# Multiplicity, Selection, and Modern Statistical Inference

Xianyang Zhang

Draft version, November 2026.

Comments, corrections, and suggestions are welcome.

# Contents


Preface vii
To the Reader vii
Prerequisites vii
Organization vii
Acknowledgments viii

Notation ix
Counts ix
Hypotheses and decisions ix
Error rates ix
Evidence and observables ix
Probability and statistics x
Regression and high-dimensional models x
Conventions x

Chapter 1. Introduction: Multiplicity, Selection, and Modern Inference 1
1. Targets Before Methods 1
2. A Common Structure 2
3. How the Chapters Fit Together 2
4. How to Read the Book 3

Chapter 2. Global Testing: Extremes, Aggregates, and Detection Boundaries 4
1. From Many Hypotheses to One Question 4
2. Tests Based on Small P-Values 5
3. Tests That Combine Evidence 6
4. The Gaussian Sequence Model 8
5. $L_2$ Tests for Dense Signals 12
6. A Lower Bound for Dense Alternatives 13
7. Comparing the Maximum and $L_2$ Tests 16
8. Assumptions in Plain Language 17
9. Bibliographic Notes 18
10. Exercises 18

Chapter 3. Simes, Goodness-of-Fit, and Higher Criticism 21
1. A Uniformity Problem 21
2. Simes as a Crossing Rule 22
3. Empirical Processes and Classical Goodness-of-Fit Tests 24
4. Higher Criticism, Berk-Jones, and Average Likelihood Ratios 25
5. Rare and Weak Gaussian Mixtures 27
6. Dependence Can Hurt, but It Can Also Help 30
7. Assumptions in Plain Language 32

8. Bibliographic Notes 33
9. Exercises 33

Chapter 4. Familywise Error Rate, Closure, and Graphical Procedures 36
1. Why Strong Control Is Needed 36
2. Bonferroni, Sidak, and Holm 38
3. Step-Up Procedures and Simes 40
4. The Closure Principle 41
5. Closed Bonferroni, Closed Simes, and Hommel 42
6. Graphical Error Spending 44
7. Weighted Bonferroni and Consonance 48
8. Simulation: Error and Power 50
9. Assumptions in Plain Language 51
10. Bibliographic Notes 52
11. Exercises 52

Chapter 5. False Discovery Rate and the BH Principle 54
1. FDP and FDR 54
2. The BH Step-Up Rule 54
3. Adjusted P-Values 56
4. What the BH Estimate Means 59
5. The Independent-Null Proof 59
6. Z-Score Thresholds 61
7. Weighted BH 62
8. Learning Weights from Covariates 62
9. A Generalized BH Template 65
10. Arbitrary Dependence: The BY Correction 66
11. Mirror Estimation 68
12. Assumptions in Plain Language 70
13. Bibliographic Notes 71
14. Exercises 72

Chapter 6. Dependence, Adaptive FDR, and Empirical Bayes 74
1. Positive Dependence and PRDS 74
2. Why PRDS Does Not Protect Mirror Thresholds 79
3. Empirical Processes and Reverse Martingales 84
4. Estimating the Null Fraction 85
5. Q-Values 88
6. The Two-Groups Model and Local FDR 89
7. Optimal Thresholding by Local FDR 91
8. Covariate-Assisted Local FDR 95
9. AdaPT: Masking Guided by Local FDR 97
10. Positive FDR 100
11. Conditional Calibration 101
12. Assumptions in Plain Language 103
13. Bibliographic Notes 103
14. Exercises 104

Chapter 7. Structured and Hierarchical Multiple Testing 106
1. Why Flat BH Is Not Enough 106

2. Group BH via Simes Aggregation 107
3. The p-Filter 110
4. Trees of Hypotheses and TreeBH 111
5. Replicability and Partial Conjunction 116
6. Connections to Regulatory Practice 119
7. Assumptions in Plain Language 119
8. Bibliographic Notes 120
9. Exercises 121

Chapter 8. E-Values, Safe Testing, and Multiple Testing 124
1. Setup and Notation 124
2. Evidence on an Expectation Scale 125
3. Likelihood Ratios, Bayes Factors, and Composite Nulls 130
4. Betting Scores and Safe Tests 134
5. Optional continuation and e-processes 140
6. E-BH: FDR Control by Aggregate Null Evidence 141
7. How Threshold Rules Create E-Values 143
8. Combining and Assembling E-Values 150
9. Assumptions in Plain Language 154
10. Bibliographic Notes 155
11. Exercises 156

Chapter 9. Conditional Randomization and Knockoff Filters 158
1. Conditional Feature Nulls 158
2. Conditional Randomization Tests 159
3. Why Marginal Resampling Fails 161
4. Fixed-X Knockoffs 162
5. Knockoff Statistics and Thresholds 164
6. Model-X Knockoffs 169
7. Gaussian Model-X Knockoffs 172
8. Sequential Conditional Independent Pairs and Structured Feature Laws 173
9. Computation and Approximation 175
10. Multilayer Knockoffs 177
11. Assumptions in Plain Language 180
12. Bibliographic Notes 181
13. Exercises 181

Chapter 10. Conformal Prediction and Conformal P-Values 184
1. The Target 184
2. The Rank Argument 185
3. Full Conformal Prediction 188
4. Split Conformal Prediction 189
5. What conditional coverage means 191
6. Randomization and Ties 192
7. Conformal P-Values 194
8. Conformalized Quantile Regression 197
9. Jackknife and Jackknife+ 200
10. Conformal e-prediction 202
11. Assumptions in Plain Language 203
12. Bibliographic Notes 204

13. Exercises 205

Chapter 11. Debiased Lasso and High-Dimensional Confidence Intervals 207
1. Model and Target 207
2. Regularization and the Lasso 208
3. What Lasso Consistency Gives 210
4. Why the Lasso Limit Is Not Enough 212
5. Debiasing by an Approximate Inverse 213
6. Nodewise Lasso and Precision Approximation 218
7. The Optimal-Projection View 220
8. Beyond Linear Regression: Decorrelated Scores 221
9. Many Coordinates and Multiple Testing 224
10. Assumptions in Plain Language 226
11. Failure Modes and Diagnostics 227
12. Limits of Adaptivity 229
13. Bibliographic Notes 230
14. Exercises 231

Chapter 12. Selective Inference and False Coverage Rate 233
1. A Selected-Interval Problem 233
2. Targets After Selection 234
3. Conditional Coverage and Its Limits 234
4. False Coverage Rate 235
5. The Benjamini–Yekutieli FCR Adjustment 236
6. POSI: Protection Against Arbitrary Selection 238
7. Polyhedral Selective Inference for the Lasso 239
8. Selective Inference for Clustering 243
9. Data Thinning and Data Fission 245
10. Selective Inference Beyond Fixed-$\lambda$ Models 249
11. Assumptions in Plain Language 250
12. Failure Modes and Diagnostics 251
13. Bibliographic Notes 251
14. Exercises 252

Chapter 13. Prediction-Powered and Active Statistical Inference 255
1. Cheap Predictions, Expensive Measurements 255
2. Correcting Predictions with Observed Errors 256
3. How Much Should Predictions Contribute? 259
4. Estimating Equations 261
5. Cross-Prediction 264
6. Recalibration 265
7. Choosing Which Outcomes to Measure 267
8. Assumptions in Plain Language 272
9. Bibliographic Notes and Further Directions 273
10. Exercises 274

Chapter 14. Applications and Research Frontiers 276
1. A Shared Workflow 276
2. Genomic Screening 277
3. Dependence, Linkage Disequilibrium, and Structured Discoveries 277

4. Benchmark Multiplicity for Language Models 279
5. LLM-as-a-Judge and Latent Entanglement 281
6. Watermarked Generation 283
7. Watermark Detection and Scanning 285
8. Contamination Auditing by Exchangeability 289
9. Reference-Relative Factuality 291
10. What the Assumptions Are Doing 294
11. Bibliographic Notes 294
12. Exercises 295

Appendix A. Advanced Safe and Sequential Inference 297
1. Numeraires and reverse information projection 297
2. Optional Continuation and E-Processes 300
3. Universality of e-BH 305
4. Asymptotic E-Values 306
5. Online FDR Control 309
6. Confidence Sequences 312

Appendix B. Advanced Conformal Theory 314
1. What Conditional Coverage Means 314
2. The conformal PRDS argument 320
3. Coverage After Selection 321
4. The jackknife+ tournament argument 325
5. Online Conformal P-Values and Testing Exchangeability 329
6. When Exchangeability Fails 335
7. Conformal Risk Control and Learn-Then-Test 340

Appendix C. Details for Decorrelated-Score Inference 344
1. From Estimation Errors to Score Errors 344
2. Normal Approximation and the One-Step Correction 346
3. General Losses and Misspecified Models 347

Appendix D. Method-Selection Matrix 349

Appendix E. Glossary 351

Bibliography 353

Index 365

# Preface

## To the Reader

I wrote this book for students who have already met probability, mathematical statistics, and linear regression, and who now want a coherent route into modern large-scale inference. The recurring question is simple: after many hypotheses, model choices, or benchmark comparisons have been examined, what claim can still be made honestly?

These chapters grew out of lectures for statistics Ph.D. students. I have tried to make the written version more patient than lectures can be: definitions before use, examples that show what the theory is buying, reproducible numerical experiments, and exercises that turn reading into active checking. Some arguments need to be read twice; that is normal. The goal is not to memorize a catalogue of methods, but to learn how to ask careful inferential questions when modern data analysis can easily outrun intuition.

## Prerequisites

To read the technical chapters comfortably, it helps to have seen the following material before opening Chapter 2.

- Graduate probability, including the measure-theoretic language of expectation, conditional expectation, modes of convergence, and the Gaussian and multinomial distributions.
- Mathematical statistics: hypothesis testing, Neyman-Pearson, asymptotic normality, likelihood ratios, contiguity in some form.
- Linear regression in matrix form, including the normal equations and the projection geometry of ordinary least squares (OLS).
- Basic real analysis sufficient to read $\epsilon$-style arguments and routine inequalities (Markov, Chebyshev, Cauchy-Schwarz, union bound).

Familiarity with the Lasso, martingales, and empirical processes is helpful but not assumed. When these tools first appear, I introduce what is needed for the chapter and point to standard references for readers who want to go deeper.

## Organization

Chapter 1 introduces the book's common language: inferential targets, evidence measures, error rates, dependence, and selection. Chapters 2 and 3 develop global testing: detecting whether any signal is present in a high-dimensional collection of hypotheses. Chapters 4 to 6 cover individual decisions under the familywise error rate (FWER) and the false discovery rate (FDR), including closure, graphical procedures, the Benjamini–Hochberg (BH) procedure, positive regression dependence on a subset (PRDS), adaptive estimation, and the two-groups empirical-Bayes model. Chapter 7 extends these tools to non-flat families: grouped hypotheses, tree-structured multiplicity, and replicability across studies. Chapter 8 introduces e-values, safe testing, and the e-value analogue of BH (e-BH); advanced sequential derivations are collected in Appendix A. Chapter 9 develops the conditional randomization test (CRT), knockoff filters, and

multilayer knockoffs for high-dimensional variable selection. Chapters 10 to 12 treat distribution-free prediction with conformal inference, confidence intervals for high-dimensional regression coefficients via debiasing and decorrelated scores, and selective inference after data-driven model selection (including data thinning and fission for unsupervised targets). Advanced conformal limits, proofs, drift bounds, and risk control are collected in Appendix B; proofs for decorrelated-score inference are in Appendix C. Chapter 13 develops prediction-powered inference (PPI): using a small labeled sample to correct predictions and estimate population quantities. It then studies how to choose which labels to acquire while retaining valid inference. Chapter 14 uses applications in genomics, watermarking, contamination auditing, and reference-relative factuality to show how the same statistical principles reappear in modern scientific and artificial intelligence (AI) applications.

The chapters can mostly be read in order, but they are not strictly linear. A reader who wants to learn FDR control quickly may go straight from Chapter 4 to Chapters 5 and 6. A reader interested in e-values may read Chapter 8 immediately after Chapter 5. A reader focused on prediction and selective inference may go directly from Chapter 6 to Chapters 10 and 12. Readers interested specifically in population estimation with scarce labels can read the core route of Chapter 13 using expectations, variances, and the central limit theorem (CLT). Readers interested in hierarchical genomic or regulatory testing can read Chapter 7 immediately after the FDR chapters.

## Acknowledgments

This book grew out of the STAT 689 lecture sequence at Texas A&M University. I am grateful to the students and colleagues whose questions and comments sharpened the exposition. I also benefited a lot from Emmanuel Candès's Stats 300C lecture materials.

# Notation

The book uses the conventions below. Chapter-specific notation is introduced where it appears.

### Counts

| | |
|---|---|
| $m$ | number of hypotheses, tests, p-values, or e-values |
| $n$ | sample size (number of observations) |
| $N$ | unlabeled sample size in prediction-powered inference (Chapter 13) |
| $p$ | number of features in a regression model |
| $m_0$ | number of true null hypotheses |
| $m_1$ | number of nonnulls, $m_1 = m - m_0$ |

### Hypotheses and decisions

| | |
|---|---|
| $H_i$, $H_{0i}$, $H_{1i}$ | $i$th hypothesis with its null and alternative |
| $H_I = \bigcap_{i \in I} H_i$ | intersection hypothesis on index set $I$ |
| $\mathcal{I}_0$ | index set of true nulls, so $m_0 = \lvert\mathcal{I}_0\rvert$ |
| $\mathcal{R}, R$ | rejection set and its size $R = \lvert\mathcal{R}\rvert$ |
| $V$ | number of false rejections, $V = \lvert\mathcal{R} \cap \mathcal{I}_0\rvert$ |
| $U$ | true nulls not rejected, $U = m_0 - V$ |
| $D$ | number of true discoveries, $D = R - V$ |
| $M$ | nonnulls not rejected (missed discoveries), $M = m_1 - D$ |
| $\widehat{\mathcal{S}}$ | data-driven selected set (Chapters 9, 12) |

### Error rates

| | |
|---|---|
| $\mathrm{FWER} = \mathbb{P}(V \geq 1)$ | familywise error rate |
| $\mathrm{FDP} = V/(R \vee 1)$ | false discovery proportion |
| $\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$ | false discovery rate |
| $\mathrm{pFDR} = \mathbb{E}[V/R \mid R > 0]$ | positive false discovery rate |
| FCR | false coverage rate (Chapter 12) |

### Evidence and observables

| | |
|---|---|
| $p_i$ | p-value for $H_i$, super-uniform under $H_{0i}$; also written $P_i$ when distinguishing a random p-value from its observed value |
| $E_i$ | e-value for $H_i$, satisfying $\mathbb{E}[E_i] \leq 1$ under $H_{0i}$ |
| $Z_i$ | z-score; in the Gaussian sequence model $Z_i \sim N(\mu_i, 1)$ |
| $\mathrm{lfdr}(z)$ | local false discovery rate $\mathbb{P}(\theta = 0 \mid Z = z)$ |
| $W_j$ | knockoff statistic (Chapter 9) |
| $A$ | generic conformal conformity or nonconformity score |
| $A_{\mathrm{safe}}$ | strict-safe score for reference-relative factuality |
| $T_{\mathrm{wm}}$ | watermark detection or maximum-scan statistic |

### Probability and statistics

| | |
|---|---|
| $\mathbb{P}$, $\mathbb{E}$, Var, Cov | probability, expectation, variance, covariance |
| $\mathcal{L}(X)$ | law (distribution) of a random variable |
| $\Phi$, $\phi$ | standard normal cumulative distribution function (CDF) and density |
| $\mathrm{Unif}(0,1)$ | uniform distribution on $[0,1]$ |
| $\chi^2_k$, $\chi_d$ | chi-square ($k$ df) and chi ($d$ df, the norm of a standard normal vector in $\mathbb{R}^d$) distributions |
| $\mathrm{TN}(\mu, \sigma^2, [a,b])$ | normal $N(\mu,\sigma^2)$ truncated to $[a,b]$ |
| $\xrightarrow{P}$, $\xrightarrow{d}$ | convergence in probability and in distribution |
| $o_p(\cdot)$, $O_p(\cdot)$ | stochastic little-o and big-O: $X_n = o_p(a_n)$ means $X_n/a_n \xrightarrow{P} 0$, and $X_n = O_p(a_n)$ means $X_n/a_n$ is bounded in probability |
| $\mathbf{1}\{A\}$ | indicator of the event $A$ |
| $a \vee b$, $a \wedge b$ | $\max(a,b)$ and $\min(a,b)$, respectively |

### Regression and high-dimensional models

| | |
|---|---|
| $X \in \mathbb{R}^{n\times p}$ | design matrix (rows are observations, columns are features) |
| $X_j \in \mathbb{R}^n$ | $j$th column of $X$ (the $j$th feature across all observations) |
| $X_{-j}$ | the design matrix with column $j$ removed |
| $x_i^\top \in \mathbb{R}^{1\times p}$ | $i$th row of $X$ (the feature vector for observation $i$) |
| $y$, $Y$ | response vector |
| $\beta \in \mathbb{R}^p$ | regression coefficients (target) |
| $\widehat{\beta}$ | estimator of $\beta$ |
| $\mathcal{S}_\beta$ | coefficient support $\{j : \beta_j \neq 0\}$ |
| $\theta$, $\Theta_0$ | generic parameter and null parameter set |
| $\Sigma$, $\Gamma$ | covariance matrix and precision matrix $\Sigma^{-1}$ |
| $\widetilde{X}$ | knockoff copy of $X$ (Chapter 9) |

### Conventions

- Vectors are not boldfaced; context distinguishes scalars from vectors.
- Uppercase/lowercase distinguish random variables/realizations when needed. Chapters 1–5 may use $p_i$ for both; Chapter 6 uses $P_i$ and $p_i$, respectively.
- $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ are the real, integer, and natural numbers.
- Unless stated otherwise, limits, including $o(\cdot)$ and $O(\cdot)$, take $m$ or $n$ to infinity.
- $q$ is reserved for an FDR target; $\alpha$ denotes the stated single-test Type I error, familywise error budget, or prediction-set or confidence-interval miscoverage level.
- Hypothesis-specific q-values use uppercase $Q$, such as $Q_i$, $Q_{(i)}$, or $Q(z)$, distinct from the prespecified scalar target $q$.
- Threshold rules use $R(t) \vee 1$ to avoid division by zero. Supremum and infimum conventions are local. Proofs using an inequality at the selected threshold ensure attainment through a finite search grid or sufficient one-sided continuity or closedness.
- Blackboard $\mathbb{P}, \mathbb{E}$ are abstract probability/expectation operators; italic $P_0, P_1, P_\theta, E_\theta$ denote specific measures/expectations, identified by subscript.
- Roman distribution names: binomial $\mathrm{Bin}(n,p)$, Bernoulli$(p)$, uniform $\mathrm{Unif}(0,1)$.
- Group counts are $G$ (one layer) or $G_\ell$ (layer $\ell$). Locally declared $m$ may count benchmark tasks or model–task cells. Cluster separation $\delta_{\mathrm{sep}}$ differs from normal density $\phi$. Terminal time may use $T$; decision-table counts use $D, M$.

CHAPTER 1

# Introduction: Multiplicity, Selection, and Modern Inference

Modern data analysis rarely asks just one question. A genomic study may test 20,000 genes, a regression may consider hundreds of variables, and a large language model (LLM) benchmark may compare many models, tasks, or subgroups. When we look in many places and report only the most promising result, ordinary chance gets many opportunities to look convincing. This is the problem of *multiplicity*.

Call a gene *null* when it has no genuine association with the outcome. A p-value is a number between zero and one that measures compatibility with this no-association model: smaller values are less compatible. If all 20,000 genes are null and their p-values are independent and uniform on $[0, 1]$, then

$$\mathbb{P}\{\min_j p_j \le 10^{-5}\} = 1 - (1 - 10^{-5})^{20000} \approx 0.18.$$

Thus, even when every gene is null, there is about an 18 percent chance—roughly one in five—of seeing a p-value at least this small somewhere. A small p-value can still be useful evidence, but its meaning depends on how many chances there were to find it.

Selection creates the same issue in less visible ways. A regression analyst might choose a variable after seeing which variables look most promising. A benchmark paper might highlight the largest gain among many model–task comparisons. A prediction method might be tuned across many subgroups and reported only where its coverage looks best. These choices are not inherently wrong, but a fixed-in-advance uncertainty statement may no longer describe the claim that was actually selected. Honest inference must account for both the opportunities to look and the rule that chose what to report.

This book develops tools for these situations. Its readers are graduate students in statistics, applied researchers who need reliable large-scale inference, and statisticians who want a coherent path from classical multiple testing to modern methods such as e-values, knockoffs, conformal prediction, debiased Lasso, selective inference, and PPI. The level is mathematical, but the organizing question is practical: what claim are we trying to make, and what calibration makes that claim honest?

## 1. Targets Before Methods

**Route: Core**

The same data set can support several different inferential targets. A global test asks whether any signal is present. A multiple-testing procedure reports a discovery list while controlling a false-positive error rate. A conformal method reports a prediction set with a coverage guarantee for a future observation. PPI uses predictions together with a smaller labeled sample to estimate a population quantity, such as a mean. A post-selection method reports an interval or p-value after the object of inference has been chosen using the data.

These targets are related, but they are not interchangeable. A valid global test can tell us that a genomic experiment contains signal, but it does not identify the non-null genes. A BH discovery list controls the expected proportion of false positives among the selected genes

(Chapter 5), but it does not promise that every selected gene is non-null. A conformal prediction interval gives a coverage statement for a future response, but it does not rank variables by importance. Selective inference can repair a confidence statement after model selection, but only relative to the selection event or randomization that is actually analyzed.

This is the book's first habit of mind: state the target before choosing the procedure. The mathematics then has a job to do. It must connect the reported object to the error rate or coverage statement that the reader will interpret.

## 2. A Common Structure

**Route: Core**

Most chapters follow the same sequence:

$$\text{define the claim} \quad \longrightarrow \quad \text{calibrate evidence} \quad \longrightarrow \quad \text{report the result.}$$

The evidence may be a p-value, an e-value, a residual score, a feature statistic, or a randomized comparison. The guarantee may control the chance of any false rejection (FWER; Chapter 4) or the expected fraction of false rejections among discoveries (FDR; Chapter 5). Other guarantees concern coverage, inference after selection, or validity when the data determine when to stop collecting observations (Chapter 8). Each requires assumptions about the null distribution or the data collection process that justify its calibration.

This matching is where many statistical mistakes enter. A method calibrated for independent p-values should not be casually used under strong dependence. An interval for a pre-specified coefficient should not be read as an interval for the coefficient selected by an algorithm. A discovery procedure designed for marginal hypotheses may not answer a conditional variable-importance question. Later chapters will state assumptions carefully because, in large-scale inference, the assumptions often define the claim.

## 3. How the Chapters Fit Together

**Route: Core**

Figure 1.1 summarizes the book's structure. The arrows are conceptual dependencies, not a rigid reading order. The early chapters develop global testing and multiple-testing error rates; the middle chapters address dependence, structure, adaptivity, and safe evidence; the later chapters handle algorithmic targets, prediction, high-dimensional estimation, and selection.

One branch uses goodness-of-fit (GoF) tests, including higher criticism (HC). These compare a null cumulative distribution function (CDF) with the empirical cumulative distribution function (ECDF) of the observed p-values.

Chapters 2 and 3 begin with the broadest question: is there signal anywhere? Chapters 4, 5, and 6 move from global evidence to discovery lists, contrasting FWER control with FDR control and then studying dependence, adaptive estimation, and empirical Bayes interpretations. Chapter 7 extends the FDR framework to non-flat families: grouped hypotheses, tree-structured genomic and benchmark testing, and replicability across studies. Chapter 8 introduces e-values and safe testing, where evidence can be accumulated under optional continuation, and develops online FDR and confidence sequences for sequential workflows.

The remaining chapters move closer to modern data-analysis workflows. Chapter 9 studies conditional feature importance through CRTs, knockoff filters, and multilayer knockoffs. Chapter 10 develops distribution-free prediction through conformal methods, including conformal risk control (CRC) and beyond-exchangeability extensions. Chapters 11 and 12 treat inference after high-dimensional modeling and data-dependent selection, including data thinning and fission

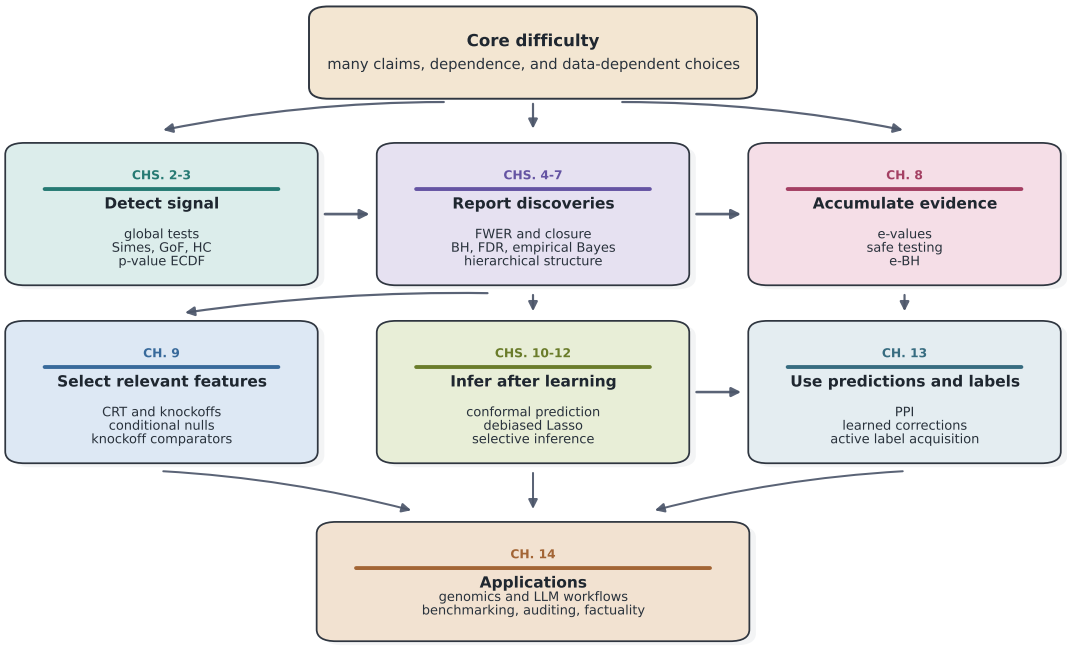



FIGURE 1.1. Conceptual map of the book. Global testing detects signal; FWER, FDR, hierarchical, and e-value methods calibrate discovery rules. Knockoffs, conformal prediction, debiased Lasso, and selective inference address algorithmic or selected targets. PPI and active inference combine predictions with acquired labels to estimate population quantities. Applications connect these ideas to genomics and language-model evaluation.

for valid inference after unsupervised discovery steps such as clustering. Chapter 13 shows how a small labeled sample can correct a large collection of predictions, and how randomized label acquisition can improve precision at a fixed labeling budget. Chapter 14 closes the book by showing how the same ideas reappear in genomics and language-model evaluation.

## 4. How to Read the Book

**Route: Core**

A graduate course can follow the chapter order. Other readers may start from their target: global signal detection in Chapters 2–3, discovery lists in Chapters 4–7, sequential evidence in Chapter 8, and algorithmic or post-selection inference in Chapters 9–12. For population estimation with scarce labels, Chapter 13 requires only expectations, variances, and the CLT on its core route.

Section labels mark *Core*, *Advanced*, *Optional proof*, or *Research frontier*. A first course can follow the Core route and assign advanced material as further reading. Appendices A, B, and C collect technical derivations for safe inference, conformal prediction, and decorrelated scores. Each subsequent chapter's "Assumptions in practice" card distinguishes what the study design supports from what remains an assumption, and suggests sensitivity checks or alternative methods.

Before applying a method, ask five questions. What is reported? What family or selection rule does the guarantee cover? Which assumptions justify calibration? How does dependence affect the argument? Does the stated error rate match the scientific claim?

CHAPTER 2

# Global Testing: Extremes, Aggregates, and Detection Boundaries

Large-scale inference often begins before selection, estimation, or ranking. The first question is global: does the data set contain any signal at all? In a genomic screen, this asks whether at least one gene behaves differently between cases and controls. In a high-dimensional regression, it asks whether at least one coefficient is nonzero. In a model-evaluation study, it asks whether a collection of discrepancies can plausibly be explained by noise.

This chapter develops the global testing material that will be used throughout the book. The main theme is that different global tests search for different shapes of evidence. Bonferroni and maximum tests are tuned to a few strong signals. Fisher, Stouffer, and chi-square tests accumulate many weak signals. The Cauchy combination test sits closer to the extreme-value side while retaining a useful robustness to dependence. The chapter ends by making these statements quantitative in the Gaussian sequence model.

## 1. From Many Hypotheses to One Question

**Route: Core**

Suppose a study produces $m$ null hypotheses

$$H_{01}, \dots, H_{0m}.$$

The global null is the intersection

$$H_0 = \bigcap_{j=1}^{m} H_{0j}.$$

Rejecting $H_0$ provides evidence that at least one individual null is false. It does not identify which one. A level-$\alpha$ global test limits the probability of rejection to at most $\alpha$ whenever all the individual nulls are true; this probability is its Type I error rate. The localization problem, where we decide which hypotheses to reject while controlling FWER or FDR, begins in later chapters.

**Example 2.1: Benchmark-comparison screen**

Suppose two prediction systems are compared on $m$ related benchmark tasks, as in modern model-evaluation studies [53, 127]. For task $j$, the systems are evaluated on the same $n_j$ test cases. Let $L_{jk}^A$ and $L_{jk}^B$ be losses on test case $k$, and define the paired improvement

$$D_{jk} = L_{jk}^B - L_{jk}^A,$$

so positive values favor system $A$. A one-sided individual null is

$$H_{0j} : \mathbb{E}D_{j1} \le 0,$$

meaning that system $A$ has no positive mean improvement over system $B$ on task $j$. For the benchmark Gaussian approximation, assume that within each task the paired differences

$D_{j1}, \ldots, D_{jn_j}$ are independent and identically distributed with

$$\delta_j = \mathbb{E}D_{j1}, \qquad 0 < \sigma_j^2 = \mathrm{Var}(D_{j1}) < \infty.$$

No independence across different tasks is needed for this marginal approximation. Define the standardized mean shift

$$\mu_j = \frac{\sqrt{n_j}\,\delta_j}{\sigma_j}.$$

With

$$\bar{D}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} D_{jk}, \qquad s_j^2 = \frac{1}{n_j - 1}\sum_{k=1}^{n_j}(D_{jk} - \bar{D}_j)^2,$$

the standardized statistic

$$Z_j = \frac{\sqrt{n_j}\,\bar{D}_j}{s_j}$$

is approximately $N(0,1)$ at the boundary $\delta_j = 0$, by the CLT and Slutsky's theorem. For power calculations, the shifted approximation $N(\mu_j, 1)$ also applies along alternatives with bounded $\mu_j$, provided the centered-data CLT and variance consistency continue to hold. These are local alternatives: the mean improvement decreases on the $n_j^{-1/2}$ scale as the sample size grows. Here $\delta_j > 0$ corresponds to a real improvement. The global question is whether any benchmark task carries signal. Throughout this chapter we work directly with the Gaussian z-score model, viewing it as an idealized summary that already absorbs whatever distributional approximation produced it.

The common input to many global tests is a collection of p-values $p_1, \ldots, p_m$. Let $\Theta_{0j}$ denote all parameter values satisfying $H_{0j}$. A p-value for this possibly composite null is valid if its probability of falling below any threshold $t$ is at most $t$ under every parameter in the null:

$$\sup_{\theta \in \Theta_{0j}} \mathbb{P}_\theta(p_j \le t) \le t, \qquad 0 \le t \le 1.$$

This property is called *super-uniformity*. Equivalently, the inequality must hold under every distribution in the null; for a simple null, the supremum contains only one distribution. Exact uniformity is convenient but not necessary for Type I error control.

For the Gaussian model $Z_j \sim N(\mu_j, 1)$, the two-sided p-value for $H_{0j} : \mu_j = 0$ is

$$p_j = 2\{1 - \Phi(|Z_j|)\},$$

whereas the one-sided upper-tail p-value for $H_{0j} : \mu_j \le 0$ is $p_j = 1 - \Phi(Z_j)$. The direction of the alternative is part of the scientific model and should be fixed before looking at the data.

## 2. Tests Based on Small P-Values

**Route: Core**

**2.1. Bonferroni and Sidak.** The Bonferroni global test rejects $H_0$ at level $\alpha$ if

$$\min_{1 \le j \le m} p_j \le \frac{\alpha}{m}.$$

**Proposition 2.2: Bonferroni validity**

If all null p-values are super-uniform, then the Bonferroni global test has Type I error at most $\alpha$, with no assumptions on the dependence among the p-values.

**Proof of Proposition 2.2**

Under the global null,

$$\mathbb{P}\left(\min_j p_j \le \frac{\alpha}{m}\right) = \mathbb{P}\left(\bigcup_{j=1}^{m}\{p_j \le \alpha/m\}\right) \le \sum_{j=1}^{m}\mathbb{P}(p_j \le \alpha/m) \le m \cdot \frac{\alpha}{m} = \alpha.$$

□

This proof is only a union bound, but that simplicity is exactly why the guarantee is so robust. Bonferroni remains valid under arbitrary dependence. When the p-values are independent and exactly $\mathrm{Unif}(0,1)$ under the global null, the Sidak threshold

$$1-(1-\alpha)^{1/m}$$

gives exact level $\alpha$. Indeed,

$$\mathbb{P}\left(\min_j p_j \le 1-(1-\alpha)^{1/m}\right) = \alpha.$$

With independent but merely super-uniform null p-values, the same threshold still gives level at most $\alpha$, but equality need not hold.

Bonferroni is often described as conservative. This is sometimes true, but it is not the right mental model in independent large-scale problems. If $p_1, \ldots, p_m$ are independent uniforms under the global null, then the *size* of the Bonferroni test — the actual Type I error probability — is

$$1-\left(1-\frac{\alpha}{m}\right)^m \longrightarrow 1-e^{-\alpha}.$$

For $\alpha = 0.05$, the limit is 0.0488, about 0.12 percentage points below the nominal level. So under independence Bonferroni gives up almost nothing; it is a near-exact level-$\alpha$ test. In the equicorrelated Gaussian model of Figure 2.1, increasing positive correlation makes the p-values move together, reduces the effective number of distinct chances to cross $\alpha/m$, and makes Bonferroni increasingly conservative. This is a model-specific illustration, not a claim that every notion of positive dependence must reduce Bonferroni's size.

## 3. Tests That Combine Evidence

**Route: Core**

Bonferroni reacts to the smallest p-value. If the alternative consists of many weak effects, no individual p-value may be spectacular, even though the combined evidence is overwhelming. Combination tests are designed for this regime.

**3.1. Fisher's Combination Test.** If $p \sim \mathrm{Unif}(0,1)$, then $-2\log p \sim \chi^2_2$. Therefore, if the p-values are independent and uniform under the global null,

$$T_F = -2\sum_{j=1}^{m}\log p_j \sim \chi^2_{2m}.$$

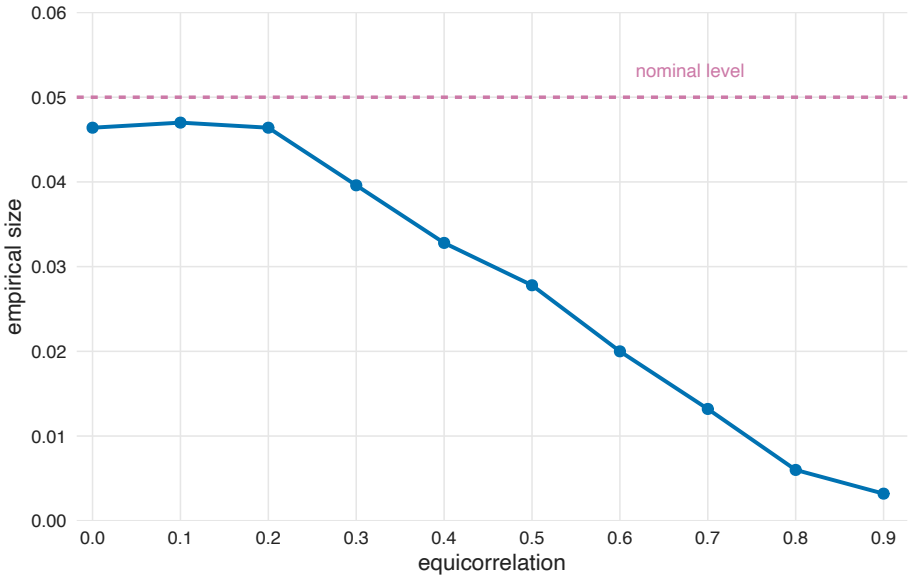



FIGURE 2.1. Empirical size of Bonferroni's two-sided global test under an equicorrelated Gaussian null with $m = 200$, $\alpha = 0.05$, based on 5,000 independent simulation runs. The test remains valid, but positive correlation makes it increasingly conservative.

Fisher's test rejects for large $T_F$. The exact chi-square calibration is a statement about both independence and uniform null distributions. Under independence, super-uniform p-values give a conservative upper-tail test with the same cutoff. Under dependence, the null calibration must be justified separately.

Fisher's statistic is not the only classical combination rule. Pearson's statistic

$$T_P = -2\sum_{j=1}^{m} \log(1 - p_j)$$

is also $\chi^2_{2m}$ under independent uniform p-values and rejects for large values. Its direction is easiest to see on the observed-score axis. For the one-sided upper-tail p-value $p_j = 1 - \Phi(Z_j)$,

$$\underset{Z_j \ll 0,\ p_j \to 1}{\text{harmful score}} \longleftarrow \underset{Z_j = 0,\ p_j = 1/2}{\text{no observed difference}} \longrightarrow \underset{Z_j \gg 0,\ p_j \to 0}{\text{beneficial score}}$$

Thus Pearson combines evidence at the harmful end of this axis, opposite to Fisher's combination of unusually small p-values. The exact chi-square statement is scoped to the point global null $\mu_1 = \cdots = \mu_m = 0$. Large upper-tail p-values are *not* evidence against the composite benefit null $H_{0j} : \mu_j \le 0$, because harmful effects lie inside that null. Reversing direction changes the scientific hypothesis; it does not turn Pearson's statistic into a test for benefit.

Stouffer's method maps p-values to z-scores. For one-sided p-values, define $X_j = \Phi^{-1}(1 - p_j)$. If the null p-values are independent and uniform, these scores are independent $N(0, 1)$, so

$$Z_S = \frac{\sum_{j=1}^{m} X_j}{\sqrt{m}} \sim N(0, 1).$$

A weighted version,

$$Z_w = \frac{\sum_{j=1}^{m} w_j X_j}{\sqrt{\sum_{j=1}^{m} w_j^2}},$$

is valid under independent standard normal null scores for fixed weights. This is the first appearance of a theme that will return later: external information can be used to tilt power toward hypotheses that are more promising or measured more precisely.

For readers coming from classical one-dimensional testing, the important distinction is not the name of the combination rule but the functional being applied to the p-value vector. Bonferroni uses the minimum p-value. Fisher uses an average of $-\log p_j$. Stouffer uses an average of normal scores. Pearson uses the opposite tail and is most natural when unusually large p-values are the prespecified evidence direction under a point-null calibration. These choices encode assumptions about the shape and direction of the alternative before any asymptotic theory is invoked.

**3.2. Cauchy Combination.** The Cauchy combination test transforms p-values by

$$W = \sum_{j=1}^{m} w_j \tan\{\pi(1/2 - p_j)\}, \qquad w_j \geq 0, \quad \sum_{j=1}^{m} w_j = 1.$$

If $p_j$ is uniform, then $\tan\{\pi(1/2 - p_j)\}$ has the standard Cauchy distribution. The heavy tail means that a single very small p-value can dominate the sum. Indeed,

$$\tan\{\pi(1/2 - p)\} = \cot(\pi p) \sim \frac{1}{\pi p}, \qquad p \downarrow 0,$$

because $\sin(\pi p) \sim \pi p$ and $\cos(\pi p) \to 1$. Under mutual independence, the transformed variables are iid standard Cauchy, so Cauchy stability makes $W$ exactly standard Cauchy at every threshold. The dependent result is different and more specific. Fix $m$ and fixed nonnegative weights summing to one. Under the point global null, suppose each pair of scores $(Z_i, Z_j)$ is bivariate normal with standard normal marginals and arbitrary correlation, and form the two-sided p-values $p_j = 2\{1 - \Phi(|Z_j|)\}$. Then the upper tail of $W$ matches that of a standard Cauchy variable $C$:

$$\mathbb{P}(W > t) \sim \mathbb{P}(C > t) \sim \frac{1}{\pi t}, \qquad t \to \infty$$

[129]. The limit holds for fixed $m$ as the tail threshold $t$ diverges; it is not an $m \to \infty$ statement, and it does not say that the body of the dependent null distribution is exactly Cauchy. Thus the Cauchy test is often useful when one wants an analytic extreme-tail calibration without mutual independence.

The same idea extends beyond the Cauchy distribution: other heavy-tailed calibrators can preserve sensitivity to extreme p-values while allowing a tractable tail approximation.

Another way to combine evidence is to compare the ordered p-values with thresholds that depend on their ranks. Chapter 3 develops the Simes test, which uses such a comparison to make one global decision. Chapter 5 uses the same crossing idea in the BH procedure to produce a discovery list with FDR control. The distinction remains the one made at the start of this chapter: detecting that signal is present and deciding which hypotheses to reject are different inferential tasks.

## 4. The Gaussian Sequence Model

**Route: Advanced**

To compare global tests sharply, we now work in the Gaussian sequence model

$$Z_j = \mu_j + \xi_j, \qquad \xi_j \overset{\text{iid}}{\sim} N(0,1), \qquad j = 1, \ldots, m.$$

The global null is $\mu = 0$. This model is deliberately simple: it removes nuisance parameters so that the geometry of the alternatives is visible.

**4.1. Bonferroni as a Maximum Test.** For two-sided p-values $p_j = 2\{1 - \Phi(|Z_j|)\}$, Bonferroni rejects when

$$\max_{1\le j\le m} |Z_j| \ge \Phi^{-1}\left(1 - \frac{\alpha}{2m}\right).$$

For the one-sided upper-tail problem, the analogous maximum test rejects when

$$\max_j Z_j \ge \Phi^{-1}(1 - \alpha/m).$$

**Proposition 2.3: Null scale of the maximum**

If $Z_1, \dots, Z_m$ are independent $N(0,1)$, then

$$\frac{\max_{1\le j\le m} Z_j}{\sqrt{2\log m}} \xrightarrow{P} 1.$$

**Proof of Proposition 2.3**

Fix $0 < \epsilon < 1$. Both directions use Mills' ratio, which states that for $x > 0$,

$$\frac{x}{x^2+1}\,\phi(x) \le 1 - \Phi(x) \le \frac{\phi(x)}{x}, \tag{1}$$

where $\phi$ is the standard normal density. For the upper tail, use the union bound:

$$\mathbb{P}\left(\max_j Z_j > (1+\epsilon)\sqrt{2\log m}\right) \le m\{1 - \Phi((1+\epsilon)\sqrt{2\log m})\}.$$

The upper bound in (1) gives $m\{1-\Phi((1+\epsilon)\sqrt{2\log m})\} \le m^{1-(1+\epsilon)^2}/\sqrt{4\pi\log m} \to 0$. Here the sharper Mills prefactor contains an additional factor $(1+\epsilon)^{-1}$, which is at most one and has been dropped. For the lower tail, independence gives

$$\mathbb{P}\left(\max_j Z_j \le (1-\epsilon)\sqrt{2\log m}\right) = \Phi((1-\epsilon)\sqrt{2\log m})^m = \left[1 - \{1 - \Phi((1-\epsilon)\sqrt{2\log m})\}\right]^m.$$

The lower bound in (1) gives

$$m\{1 - \Phi((1-\epsilon)\sqrt{2\log m})\} \ge c_\epsilon \frac{m^{1-(1-\epsilon)^2}}{\sqrt{\log m}} \to \infty$$

for a constant $c_\epsilon > 0$. If $u_m = 1 - \Phi((1-\epsilon)\sqrt{2\log m})$, then $(1-u_m)^m \le \exp(-mu_m) \to 0$, so the lower-tail probability tends to zero. □

The threshold $\Phi^{-1}(1-\alpha/m)$ is also asymptotic to $\sqrt{2\log m}$: from $1 - \Phi(t) \sim \phi(t)/t$ one solves $\phi(t)/t = \alpha/m$ to find $t = \sqrt{2\log m}\{1 + o(1)\}$. Thus the maximum test looks for at least one coordinate whose mean is on the extreme-noise scale.

**4.2. One Sparse Signal.** Suppose exactly one coordinate $j^*$ has a positive mean $\mu^* > 0$, while all others have mean zero. Set

$$c_{m,\alpha} = \Phi^{-1}(1 - \alpha/m) = \sqrt{2\log m}\{1 + o(1)\}.$$

For fixed $\epsilon > 0$, if

$$\mu^* = (1+\epsilon)\sqrt{2\log m},$$

then

$$\mathbb{P}\Big(\max_j Z_j \geq c_{m,\alpha}\Big) \geq \mathbb{P}(Z_{j^*} \geq c_{m,\alpha}) \\ = \Phi(\mu^* - c_{m,\alpha}) \longrightarrow 1,$$

because $\mu^* - c_{m,\alpha} = \{\epsilon + o(1)\}\sqrt{2\log m} \to \infty$. Thus the signal coordinate alone proves that the maximum test has power tending to one above the boundary.

If instead, for fixed $0 < \epsilon < 1$,

$$\mu^* = (1-\epsilon)\sqrt{2\log m},$$

the signal is below the maximum-noise scale. Let $j^*$ be the index of the signal coordinate and write $Z_{j^*} = \mu^* + \xi_{j^*}$ with $\xi_{j^*} \sim N(0,1)$. Using independence,

$$\begin{aligned}\mathbb{P}(\max_j Z_j \leq \Phi^{-1}(1-\alpha/m)) &= \mathbb{P}(\xi_{j^*} \leq \Phi^{-1}(1-\alpha/m) - \mu^*) \prod_{j \neq j^*} \mathbb{P}(Z_j \leq \Phi^{-1}(1-\alpha/m)) \\ &= \mathbb{P}(\xi_{j^*} \leq \epsilon\sqrt{2\log m}\{1+o(1)\}) \cdot (1-\alpha/m)^{m-1} \\ &\longrightarrow 1 \cdot e^{-\alpha} = e^{-\alpha}.\end{aligned}$$

The first factor tends to one because $\Phi^{-1}(1-\alpha/m) - \mu^* = \epsilon\sqrt{2\log m}\{1+o(1)\} \to \infty$; the second factor uses the same limit as under the null. Hence the rejection probability tends to $1 - e^{-\alpha}$, the same limit as under the null. For small $\alpha$, $1 - e^{-\alpha} = \alpha + O(\alpha^2)$, so this limiting power is only the nominal size up to the usual Bonferroni conservativeness. In this sense, the maximum test cannot reliably detect a single signal below the $\sqrt{2\log m}$ scale.

**4.3. No Test Uniformly Beats This Boundary.** The preceding statement is about one test. We now show a uniform lower bound: no level-$\alpha$ test has power tending to one uniformly over the unknown-location class

$$\mathcal{A}_m(\mu^*) = \{\mu^* e_j : 1 \leq j \leq m\}$$

where $e_j \in \mathbb{R}^m$ is the $j$th standard basis vector, with a one in coordinate $j$ and zeros elsewhere. Thus $\mu^* e_j$ represents a signal of size $\mu^*$ located only in coordinate $j$. When $0 < \mu^* < (1-\epsilon)\sqrt{2\log m}$, the standard route averages the power under a uniform prior on the signal location, producing a mixture alternative that is nearly indistinguishable from the null.

Let $P_0 = N_m(0, I_m)$, let $P_{\mu^* e_j} = N_m(\mu^* e_j, I_m)$, and define the uniform mixture

$$P_1 = \frac{1}{m}\sum_{j=1}^m P_{\mu^* e_j}.$$

Equivalently, draw $J$ uniformly from $\{1,\dots,m\}$, set $\mu_J = \mu^*$, and set all other means to zero. Write $E_0$ for expectation under $P_0$. The likelihood ratio $L_m = dP_1/dP_0$ is

$$L_m(Z) = \frac{1}{m}\sum_{j=1}^m \exp\left(\mu^* Z_j - \frac{(\mu^*)^2}{2}\right).$$

**Proposition 2.4: Sparse one-signal lower bound**

If $0 < \mu^* \leq (1-\epsilon)\sqrt{2\log m}$ for some fixed $\epsilon > 0$, then $L_m(Z) \xrightarrow{P_0} 1$. Consequently, for every sequence of level-$\alpha$ rejection regions $A_m$, the mixture-average power satisfies

$$P_1(A_m) = \frac{1}{m}\sum_{j=1}^m P_{\mu^* e_j}(A_m) \leq \alpha + o(1).$$

**Proof of Proposition 2.4**

The mean $E_0 L_m = 1$ is automatic from the likelihood-ratio form. For convergence in probability, a direct second-moment calculation gives

$$E_0 L_m^2 = 1 - \frac{1}{m} + \frac{e^{(\mu^*)^2}}{m} \le 1 + \frac{m^{2(1-\epsilon)^2} - 1}{m},$$

because the $m(m-1)$ off-diagonal terms have expectation 1, while the $m$ diagonal terms have expectation $e^{(\mu^*)^2}$. The displayed bound is $1 + o(1)$ when $2(1-\epsilon)^2 < 1$, i.e., when $\epsilon$ is sufficiently large: $\epsilon > 1 - 1/\sqrt{2}$. For smaller $\epsilon$, $\mu^*$ is close to $\sqrt{2\log m}$ and the rare extreme terms $\exp(\mu^* Z_j - (\mu^*)^2/2)$ have heavy upper tails that inflate the second moment. We now carry out the truncation.

Write

$$r_m = \frac{(\mu^*)^2}{2\log m} \le (1-\epsilon)^2 < 1, \qquad Y_{mj} = \exp\{\mu^* Z_j - (\mu^*)^2/2\},$$

so $L_m = m^{-1}\sum_j Y_{mj}$. Let $a_m = \sqrt{2\log m}$ and truncate at the null maximum scale:

$$\widetilde{Y}_{mj} = Y_{mj}\mathbf{1}\{Z_j \le a_m\}, \qquad \widetilde{L}_m = m^{-1}\sum_{j=1}^{m} \widetilde{Y}_{mj}.$$

The likelihood-ratio identity for a single shifted normal gives

$$E_0\{Y_{m1}\mathbf{1}(Z_1 > a_m)\} = \mathbb{P}\{N(\mu^*, 1) > a_m\} \le 1 - \Phi\{(1 - \sqrt{r_m})\sqrt{2\log m}\} \to 0.$$

Hence $E_0\widetilde{Y}_{m1} = 1 - o(1)$, and

$$E_0|L_m - \widetilde{L}_m| = E_0\{Y_{m1}\mathbf{1}(Z_1 > a_m)\} \to 0.$$

It remains to show that $\widetilde{L}_m$ concentrates around its mean. A second likelihood-ratio calculation gives

$$E_0\widetilde{Y}_{m1}^2 = e^{(\mu^*)^2}\mathbb{P}\{N(2\mu^*, 1) \le a_m\}.$$

If $r_m \le 1/4$, this is at most $e^{(\mu^*)^2} \le m^{1/2}$, so $m^{-1}E_0\widetilde{Y}_{m1}^2 \to 0$. If $r_m > 1/4$, the Gaussian tail bound $\Phi(-x) \le e^{-x^2/2}$, valid for $x \ge 0$, gives

$$\mathbb{P}\{N(2\mu^*, 1) \le a_m\} \le \exp\{-(2\sqrt{r_m} - 1)^2 \log m\}.$$

The tail bound follows by applying Markov's inequality to $e^{xZ}$ for $Z \sim N(0,1)$. Therefore

$$\frac{1}{m}E_0\widetilde{Y}_{m1}^2 \le m^{-1+2r_m-(2\sqrt{r_m}-1)^2} = m^{-2(1-\sqrt{r_m})^2} \to 0.$$

Since the truncated summands are independent,

$$\mathrm{Var}_0(\widetilde{L}_m) \le \frac{1}{m}E_0\widetilde{Y}_{m1}^2 \to 0.$$

Thus $\widetilde{L}_m - E_0\widetilde{L}_m \to 0$ in $P_0$-probability, while $E_0\widetilde{L}_m = 1 - o(1)$. Combining this with $E_0|L_m - \widetilde{L}_m| \to 0$ yields $L_m \xrightarrow{P_0} 1$. The same $L^1$ bound also gives $E_0|L_m - 1| \to 0$.

For any rejection region $A_m$,

$$P_1(A_m) - P_0(A_m) = E_0[(L_m - 1)\mathbf{1}_{A_m}] \longrightarrow 0,$$

which gives the claim. □

The conclusion is an average-power statement, and therefore

$$\min_{1\le j\le m} P_{\mu^* e_j}(A_m) \le \frac{1}{m}\sum_{j=1}^{m} P_{\mu^* e_j}(A_m) \le \alpha + o(1).$$

Thus no test can have uniformly high power over the unknown signal location below this boundary. It does not say that every fixed, prespecified location is individually undetectable. Together with the above-boundary result, this makes the maximum test rate-optimal in the minimax, unknown-location sense.

## 5. $L_2$ Tests for Dense Signals

**Route: Core**

Sparse alternatives are not the only scientifically important regime. Many studies produce weak shifts in many coordinates: a small but real effect on a pathway of correlated genes, a fractional shift on most subgroups of a treatment effect, or a modest improvement of a model over a baseline on most of a benchmark. A maximum test may miss such a pattern because no single coordinate is large.

The natural competing statistic is the squared Euclidean norm

$$T_m = \|Z\|_2^2 = \sum_{j=1}^{m} Z_j^2,$$

the classical chi-square statistic for testing $\mu = 0$ against $\mu \neq 0$ in the Gaussian sequence model. Whereas Bonferroni rewards a single extreme coordinate, $T_m$ accumulates total squared signal: any $\mu$ of large enough $\ell_2$ norm will inflate $T_m$ above its null mean, regardless of how that norm is distributed across coordinates.

Under the global null, $T_m \sim \chi^2_m$, and

$$\frac{T_m - m}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

Thus an asymptotic level-$\alpha$ $L_2$ test rejects when

$$\frac{T_m - m}{\sqrt{2m}} \ge z_{1-\alpha}.$$

Here $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$ is the standard normal quantile.

Under a fixed mean vector $\mu$,

$$T_m = \sum_{j=1}^{m} \xi_j^2 + \|\mu\|_2^2 + 2\sum_{j=1}^{m} \mu_j \xi_j.$$

If $\|\mu\|_2 = o(\sqrt{m})$, then

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

More generally,

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m + 4\|\mu\|_2^2}} \xrightarrow{d} N(0,1).$$

Define

$$\theta_m = \frac{\|\mu\|_2^2}{\sqrt{2m}}.$$

Using the preceding normal approximation, the power of the $L_2$ test is approximately

$$1 - \Phi\left(\frac{z_{1-\alpha} - \theta_m}{\sqrt{1 + \theta_m/\sqrt{m/8}}}\right).$$

Consequently:

$$\begin{aligned} \theta_m = o(1) \quad &\Longrightarrow \quad \text{power} \to \alpha, \\ \theta_m \to c < \infty \quad &\Longrightarrow \quad \text{power} \to 1 - \Phi(z_{1-\alpha} - c), \\ \theta_m \to \infty \quad &\Longrightarrow \quad \text{power} \to 1. \end{aligned}$$

Figure 2.2 turns these three regimes into a single curve. At $\theta = 0$ the rejection probability is the nominal level; a finite positive $\theta$ produces a shifted limiting power, and increasing $\theta$ drives the power toward one. The horizontal scale is total squared signal normalized by $\sqrt{2m}$, so it directly displays the dense-signal boundary.

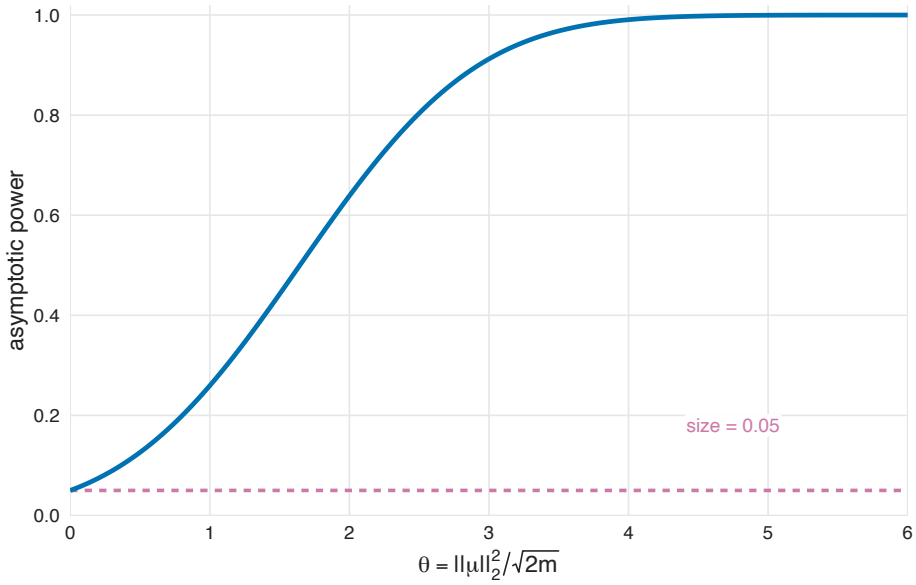



FIGURE 2.2. Asymptotic power curve of the $L_2$ test as a function of $\theta = \|\mu\|_2^2/\sqrt{2m}$, for $\alpha = 0.05$. Dense alternatives are visible once the total squared signal is on the $\sqrt{m}$ scale.

## 6. A Lower Bound for Dense Alternatives

**Route: Optional proof**

This optional route proves that the dense $L_2$ boundary is not merely a feature of the chi-square test. Readers focused on method comparison may skip the following likelihood-ratio calculation and continue at the next Advanced section.

The $L_2$ detection threshold is also optimal for dense, direction-unknown signals. Consider the testing problem

$$H_0 : \mu = 0 \qquad \text{versus} \qquad H_1 : \mu = \rho U,$$

where $U$ is drawn from the uniform distribution on the unit sphere $\mathcal{S}^{m-1}$. After integrating over $U$, the likelihood ratio is

$$L_m(Z) = \int_{\mathcal{S}^{m-1}} \exp\left(-\frac{\rho^2}{2} + \rho Z^\top u\right) d\pi_1^{(m)}(u),$$

where $\pi_1^{(m)}$ is uniform measure on the unit sphere.

If

$$\frac{\rho^2}{\sqrt{2m}} \to 0,$$

then $L_m \to 1$ in probability under the null. We prove this by a second-moment calculation, as in the sparse lower bound. Let $P_m$ denote the null law, $Q_m$ the spherical mixture alternative, and $L_m = dQ_m/dP_m$. Clearly $E_0 L_m = 1$. We now show that $E_0 L_m^2 \to 1$, which implies $L_m \to 1$ in $L^2(P_m)$, hence also in probability under $P_m$.

Under $P_m$, $Z \sim N(0, I_m)$. Using $E_0 \exp(t^\top Z) = \exp(\|t\|_2^2/2)$,

$$\begin{aligned} E_0 L_m^2 &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} E_0 \exp\left(-\rho^2 + \rho Z^\top u + \rho Z^\top v\right) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v) \\ &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} \exp\left(-\rho^2 + \frac{\rho^2}{2}\|u+v\|_2^2\right) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v) \\ &= \int_{\mathcal{S}^{m-1}} \int_{\mathcal{S}^{m-1}} \exp(\rho^2 u^\top v) d\pi_1^{(m)}(u) d\pi_1^{(m)}(v). \end{aligned}$$

For fixed $v$, choose an orthogonal matrix $B$ such that $Bv = e_1$. Since uniform measure on the sphere is rotation invariant, $Bu$ is again uniform on $\mathcal{S}^{m-1}$. Therefore

$$\int_{\mathcal{S}^{m-1}} \exp(\rho^2 u^\top v) d\pi_1^{(m)}(u) = \int_{\mathcal{S}^{m-1}} \exp(\rho^2 u_1) d\pi_1^{(m)}(u),$$

which does not depend on $v$. Hence

$$E_0 L_m^2 = E \exp(\rho^2 U_1),$$

where $U_1$ is the first coordinate of a uniform point on $\mathcal{S}^{m-1}$. This is the origin of the displayed identity.

It remains to expand the last expectation. A standard construction of uniform measure on the sphere makes the needed moments explicit [135]. Let $G = (G_1, \dots, G_m)$ have iid $N(0,1)$ coordinates and set $U = G/\|G\|_2$. Rotational invariance makes $U$ uniform on $\mathcal{S}^{m-1}$, and, for $m \geq 2$,

$$U_1^2 = \frac{G_1^2}{G_1^2 + \sum_{j=2}^m G_j^2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{m-1}{2}\right),$$

because the numerator and the remaining sum are independent chi-square variables with one and $m-1$ degrees of freedom. Hence, writing $(a)_k = a(a+1)\cdots(a+k-1)$ for the rising factorial,

$$E[U_1^{2k}] = \frac{(1/2)_k}{(m/2)_k} = \frac{(2k-1)!!}{m(m+2)\cdots(m+2k-2)}, \qquad k \geq 1.$$

By symmetry, the odd moments of $U_1$ vanish. Thus

$$\begin{aligned} E \exp(\rho^2 U_1) &= 1 + \frac{\rho^4}{2} E[U_1^2] + \sum_{k \geq 2} \frac{\rho^{4k}}{(2k)!} E[U_1^{2k}] \\ &= 1 + \frac{\rho^4}{2m} + \sum_{k \geq 2} \frac{\rho^{4k}}{(2k)!} \frac{(2k-1)!!}{m(m+2)\cdots(m+2k-2)}. \end{aligned}$$

Writing $\eta_m = \rho^4/m$, the remainder is bounded by

$$\sum_{k\geq 2} \frac{\eta_m^k}{2^k k!} = O(\eta_m^2),$$

because $(2k)! = 2^k k!(2k-1)!!$. Therefore, if $\eta_m = \rho^4/m \to 0$, equivalently $\rho^2/\sqrt{2m} \to 0$, then

$$E_0 L_m^2 = 1 + \frac{\rho^4}{2m}\{1 + o(1)\} = 1 + o(1).$$

Consequently $E_0(L_m - 1)^2 = E_0 L_m^2 - 1 \to 0$. For any rejection region $A_m$,

$$|Q_m(A_m) - P_m(A_m)| = |E_0\{(L_m - 1)\mathbf{1}_{A_m}\}| \leq \{E_0(L_m - 1)^2\}^{1/2} \to 0.$$

Hence no test can reliably separate the null from these spherical alternatives below the $L_2$ threshold.

The calculation above completes the lower bound. The next proposition relates it to more general likelihood-ratio tools. The first item is a convenient sufficient form of Le Cam's first lemma. In the usual statement, if $L_m = dQ_m/dP_m$ converges in distribution under $P_m$ to a nonnegative random variable $V$, then $Q_m$ is contiguous with respect to $P_m$ precisely when $EV = 1$. Here $L_m \to 1$ in $P_m$-probability, so the only possible limit is $V = 1$, and the condition is automatic. In the dense lower bound above we proved the stronger fact $L_m \to 1$ in $L^2(P_m)$, which even implies that the total variation distance between $P_m$ and $Q_m$ tends to zero. The second item is Le Cam's third lemma.

**Proposition 2.5: Contiguity and likelihood-ratio limits**

Let $P_m$ and $Q_m$ be sequences of probability measures, and let $L_m = dQ_m/dP_m$.

(1) If $L_m \to 1$ in $P_m$-probability, then $Q_m$ is contiguous with respect to $P_m$. Thus every event whose $P_m$-probability tends to zero also has $Q_m$-probability tending to zero.

(2) Suppose that, under $P_m$, $(T_m, \log L_m) \Rightarrow (T, Y)$ jointly, where $(T, Y)$ is bivariate normal and

$$\mathbb{E}e^Y = 1, \qquad \text{equivalently} \qquad \mathbb{E}Y = -\tfrac{1}{2}\operatorname{Var}(Y).$$

This normalization implies that $Q_m$ is contiguous with respect to $P_m$. Le Cam's third lemma then gives, under $Q_m$, the exponentially tilted limit

$$T_m \Rightarrow N(\mathbb{E}T + \operatorname{Cov}(T, Y),\ \operatorname{Var}(T)).$$

Thus the tilt shifts the Gaussian mean by $\operatorname{Cov}(T, Y)$ and leaves its variance unchanged.

In the dense lower bound above, the bound on $|Q_m(A_m) - P_m(A_m)|$ already applies to every rejection region. Thus the power of any test differs from its null rejection probability by a quantity tending to zero. This is what it means here for the null and mixture alternative to be asymptotically indistinguishable.

## 7. Comparing the Maximum and $L_2$ Tests

**Route: Advanced**

The maximum test and the $L_2$ test measure different features of $\mu$:

$$\text{maximum test:} \quad \|\mu\|_\infty, \qquad L_2 \text{ test:} \quad \frac{\|\mu\|_2^2}{\sqrt{2m}}.$$

Neither dominates the other.

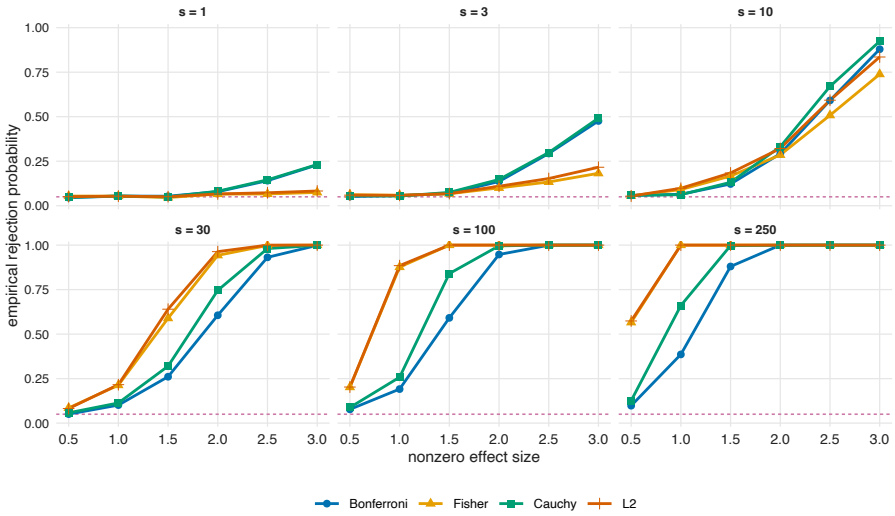



FIGURE 2.3. Empirical rejection probabilities at level $\alpha = 0.05$ in a Gaussian sequence model with $m = 500$ coordinates. Each panel fixes the number $s$ of nonzero means and varies their common effect size; each curve is based on 1,500 independent simulation runs. Bonferroni reacts fastest when $s$ is small and effects are strong. Fisher and $L_2$-type aggregation gain power as evidence is spread over many coordinates, while Cauchy remains closer to the extreme-value side.

**Example 2.6: Many strong but still sparse signals**

Suppose $m^{3/8}$ coordinates of $\mu$ equal $1.1\sqrt{2\log m}$, and all other coordinates are zero. (The exponent $3/8$ is chosen so that the signal count grows fast enough for the maximum to dominate but slowly enough that the $L_2$ total squared signal stays below the detection threshold; any exponent strictly less than $1/2$ would have the same qualitative effect.) The maximum test has power tending to one because at least one coordinate is above the extreme-noise scale. But

$$\theta_m = \frac{m^{3/8} \cdot 1.1^2(2\log m)}{\sqrt{2m}} = \frac{2.42\ \log m}{\sqrt{2}\, m^{1/8}} \to 0,$$

so the $L_2$ test is asymptotically powerless.

**Example 2.7: Many weak signals**

Suppose $\sqrt{2m}$ coordinates of $\mu$ equal a fixed constant $a > 0$, and the rest are zero. Then $\|\mu\|_\infty = a$, which is far below $\sqrt{2\log m}$. With only order $\sqrt{m}$ shifted coordinates, the probability that any of them crosses the maximum-test threshold still tends to zero, so this test has no asymptotic power beyond its null rejection probability. But $\theta_m = a^2$, and the $L_2$ test has limiting power

$$1 - \Phi(z_{1-\alpha} - a^2).$$

For example, at $\alpha = 0.05$ and $a = 2$, the limiting power is approximately 0.991.

Table 2.1 extracts three operating points from the same simulation design summarized across effect sizes in Figure 2.3. The sparse row shows the advantage of rules that react to extremes, whereas the dense row shows the gain from aggregating many modest coordinates. Reading the table and curves together also makes clear that no one global test dominates across signal geometries.

TABLE 2.1. Empirical power of four global tests. The Gaussian sequence simulation uses $m = 500$, $\alpha = 0.05$, and 1,500 independent simulation runs.

| Regime | Signals | Effect | Bonferroni | Fisher | Cauchy | $L_2$ |
|---|---|---|---|---|---|---|
| Sparse strong | 3 | 3.0 | 0.476 | 0.182 | 0.492 | 0.216 |
| Moderately sparse | 30 | 2.0 | 0.606 | 0.943 | 0.746 | 0.963 |
| Dense weak | 250 | 0.5 | 0.098 | 0.566 | 0.126 | 0.574 |

## 8. Assumptions in Plain Language

**Route: Core**

A *negative control* is a prespecified outcome, exposure, sample, or hypothesis for which the target signal should be absent for design or scientific reasons, but which passes through the same measurement, preprocessing, and analysis pipeline as the primary targets. Unexpected small p-values can expose bias or miscalibration; unremarkable controls do not prove that the primary analysis is valid. A null simulation is different: it checks calibration only under the model used to generate the simulated data.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Review how each p-value was constructed and whether different tests reuse the same observations. The study design—not the pattern of observed p-values—is the main source of information about validity and independence. |
| **What these data cannot show** | The observed p-values cannot verify their own null calibration or reveal whether nonnull effects, if present, are concentrated in a few coordinates or spread across many. |
| **Safer fallback** | If each individual p-value is valid but dependence is not justified, use Bonferroni for the global null. |

The p-value procedures in this chapter need valid null p-values. Bonferroni needs no dependence model; Fisher, Pearson, Stouffer, and the exact $L_2$ calibrations do. The Cauchy combination test uses a tail approximation that is valid under the pairwise Gaussian null condition stated above, but it is not an exact body-distribution result and needs separate justification for other null models. The Gaussian sequence lower bounds are asymptotic statements for simplified models. Their role is to explain why maximum tests are natural for rare strong signals and why $L_2$ sum-of-squares tests are natural for many weak signals.

## 9. Bibliographic Notes

**Route: Core**

Fisher's combination test is classical [61], and Stouffer's normal-score combination was developed in the social-science meta-analysis literature [171]. Sidak's exact independent-threshold calibration is due to Šidák [163]. The Cauchy combination test and its dependence-robust tail approximation are from Liu and Xie [129]. The uniform-sphere construction and coordinate moments are standard directional statistics facts; see Mardia and Jupp [135]. The contiguity arguments used in the lower bounds are Le Cam arguments; see van der Vaart [181, Chapter 6], Le Cam [111], and Le Cam and Yang [112]. Ingster's work on Gaussian sequence testing made precise how sparse alternatives can lead to non-Gaussian and infinitely divisible likelihood-ratio limits [93]; modern high-dimensional detection-boundary treatments include Donoho and Jin [52] and Arias-Castro et al. [7].

## 10. Exercises

**Route: Core**

### Basic.

**Exercise 2.8: Bonferroni and Sidak thresholds**

For $\alpha = 0.05$ and $m \in \{10, 100, 1000\}$, compute the Bonferroni threshold $b_m = \alpha/m$ and the Sidak threshold $s_m = 1 - (1-\alpha)^{1/m}$. For each $m$, report $s_m - b_m$, the ratio $s_m/b_m$, and the independent-null Bonferroni size $1 - (1-b_m)^m$; compare that size with both $\alpha$ and $1 - e^{-\alpha}$. Then, for fixed $\alpha$ and $m \to \infty$, show that

$$s_m = \frac{-\log(1-\alpha)}{m} + O(m^{-2}), \qquad \frac{s_m}{b_m} \longrightarrow \frac{-\log(1-\alpha)}{\alpha} > 1.$$

Finally use $-\log(1-\alpha) = \alpha + \alpha^2/2 + O(\alpha^3)$ to explain the different, small-$\alpha$ sense in which the two thresholds are close.

**Exercise 2.9: Fisher's test by hand**

Use Fisher's null calibration from the text: for five independent uniform null p-values, $T_F = -2\sum_{j=1}^{5} \log p_j \sim \chi^2_{10}$. Find its 0.95 quantile. For the observed p-values $0.02, 0.18, 0.07, 0.30, 0.12$, decide whether Fisher's test rejects at level 0.05.

**Exercise 2.10: Stouffer's z-score**

Convert the p-values $0.01, 0.04, 0.20, 0.32, 0.50$ to one-sided z-scores $X_j = \Phi^{-1}(1-p_j)$ and compute Stouffer's combined statistic $Z_S = \sum_j X_j/\sqrt{5}$. Compare its upper-tail p-value $1 - \Phi(Z_S)$ with Fisher's combined p-value on the same inputs.

**Exercise 2.11: Two-sided p-value calibration**

Let $Z \sim N(\mu, 1)$ and let $p = 2\{1 - \Phi(|Z|)\}$ be the two-sided p-value. Show that $p \sim \text{Unif}(0, 1)$ under $\mu = 0$. Then compute $\mathbb{P}_\mu(p \le 0.05)$ for $\mu = 1, 2, 3$ and interpret the resulting power values.

## Intermediate.

**Exercise 2.12: $L_2$ normal approximation**

In the Gaussian sequence model, prove that if $\|\mu\|_2 = o(\sqrt{m})$, then

$$\frac{T_m - \{m + \|\mu\|_2^2\}}{\sqrt{2m}} \xrightarrow{d} N(0, 1).$$

Then prove the approximation with denominator $\sqrt{2m + 4\|\mu\|_2^2}$ for an arbitrary sequence $\mu = \mu^{(m)}$. Write $Z = \mu + \xi$, where $\xi \sim N_m(0, I_m)$. Use $E\xi_j^3 = 0$ to compute the variance; uncorrelatedness alone does not establish a normal limit. For the general limit, use rotational invariance to align $\mu$ with the first coordinate and separate that coordinate from the remaining independent coordinates.

**Exercise 2.13: Cauchy combination under dependence**

Use the exact standard-Cauchy law under independence from the Cauchy-combination subsection as a benchmark. For a dependent comparison, take $m = 20$, $w_j = 1/m$, and $Z \sim N_m(0, \Sigma)$, where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = \rho = 0.5$ for $j \neq k$. Form the two-sided p-values $p_j = 2\{1 - \Phi(|Z_j|)\}$ and simulate $W = \sum_{j=1}^m w_j \tan\{\pi(1/2 - p_j)\}$. Compare the empirical and standard-Cauchy CDFs on a prespecified body grid such as $x \in \{-2, -1, 0, 1, 2\}$. Separately, for increasing $t \in \{2, 5, 10, 20\}$, report the tail ratios

$$\frac{\widehat{\mathbb{P}}(W > t)}{1/2 - \arctan(t)/\pi} \quad \text{and} \quad \frac{\widehat{\mathbb{P}}(W > t)}{1/(\pi t)}.$$

Use enough Monte Carlo repetitions to report standard errors (SEs). Assess the accuracy of the fixed-$m$, $t \to \infty$ upper-tail approximation on this finite grid. The approximation need not be accurate, or improve monotonically, at these moderate thresholds. Optionally extend the grid to 100 and 1000, increasing the simulation budget and reporting the number of observed exceedances as well as SEs. A simulation can assess finite-threshold behavior but cannot prove the tail limit.

## Computational.

**Exercise 2.14: Bonferroni under equicorrelation**

Let $m = 200$ and $Z \sim N_m(0, \Sigma_\rho)$, where $(\Sigma_\rho)_{jj} = 1$ and $(\Sigma_\rho)_{jk} = \rho$ for $j \neq k$. For $\rho \in \{0, 0.1, \ldots, 0.9\}$, estimate the empirical size of the two-sided Bonferroni global test at level 0.05 using at least 5,000 independent simulation runs. Compare your output with Figure 2.1.

**Exercise 2.15: Sparse versus dense simulation**

Reproduce Figure 2.3, using $m = 500$ independent Gaussian coordinates and level 0.05. Use positive common effects, two-sided p-values for Bonferroni, Fisher, and the equally weighted Cauchy test, and the $\chi^2_m$ cutoff for the $L_2$ test. Vary the number of nonzero coordinates and the common effect size. Compare Bonferroni, Fisher, Cauchy, and the $L_2$ test. Summarize the region in which each procedure is strongest.

**Exercise 2.16: Fisher under dependence**

Simulate $m = 200$ equicorrelated standard Gaussian z-scores under the global null, with $\rho \in \{0, 0.1, \ldots, 0.9\}$, and compute two-sided p-values. Use level 0.05 and at least 5,000 independent runs for each correlation, applying both procedures to the same simulated data. Apply Fisher's test using the independent chi-square calibration. How does the empirical size change with correlation? Compare this behavior with Bonferroni.

## Advanced.

**Exercise 2.17: Consistency of the dense-signal test**

In the Gaussian sequence model $Z \sim N_m(\mu, I_m)$, let $T_m = \|Z\|_2^2$ and suppose $\|\mu\|_2 = \rho_m$ with $\rho_m^2/\sqrt{2m} \to \infty$. Using the general normal approximation with denominator $\sqrt{2m + 4\rho_m^2}$ in Section 5, prove that the $L_2$ test rejecting when $(T_m - m)/\sqrt{2m} \geq z_{1-\alpha}$ has power tending to one. This is the above-boundary counterpart of the spherical-mixture lower bound proved in the text.

CHAPTER 3

# Simes, Goodness-of-Fit, and Higher Criticism

Chapter 2 compared global tests by asking whether evidence is concentrated in one coordinate or spread across many coordinates. This chapter gives a second view of the same question. Instead of starting from z-scores, we start from the empirical distribution of p-values. Under a clean global null, continuous p-values look uniform. A global test can therefore be read as a GoF test for the uniform distribution, with different procedures looking at different parts of the ECDF.

This viewpoint is useful because it makes the geometry of rare signals visible. If a few alternatives produce very small p-values, the ECDF rises above the diagonal near zero. If many weak alternatives are present, the deviation may be spread over a wider range. Simes, Kolmogorov–Smirnov (KS), Cramér–von Mises (CvM), Anderson–Darling (AD), HC, Berk–Jones (BJ), and average likelihood ratio (ALR) tests are different ways of measuring this upward departure.

## 1. A Uniformity Problem

**Route: Core**

Suppose a large monitoring system produces $m$ one-sided p-values

$$p_1, \dots, p_m.$$

The system might be a scientific screen, an anomaly detector, or a benchmark audit. For each coordinate $j$, small $p_j$ is evidence against the coordinate null. The global null says that all coordinates are null. Under the idealized version used in this chapter,

$$p_1, \dots, p_m \overset{\text{iid}}{\sim} \text{Unif}(0,1) \qquad \text{under } H_0.$$

Write the order statistics as

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$$

and define the ECDF

$$F_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\{p_j \le t\}, \qquad 0 \le t \le 1.$$

Under the global null, $F_m(t)$ fluctuates around $t$. Under alternatives that create too many small p-values, $F_m(t)$ exceeds $t$ for small or moderate $t$.

Figure 3.1 turns this departure into a visual diagnostic. Read the diagonal as the null benchmark, then compare the height and width of each curve's left-tail lift: a narrow early spike points to a few strong signals, whereas a sustained lift points to a denser weak alternative.

This picture also clarifies why no single global test is uniformly best. A test that averages deviations over the whole unit interval can be powerful for broad departures but inefficient for a tiny bump near zero. A test that looks only at the smallest p-value is excellent for one extreme coordinate but can miss many modest ones. Tail-adaptive procedures try to avoid choosing the scale of the departure in advance.

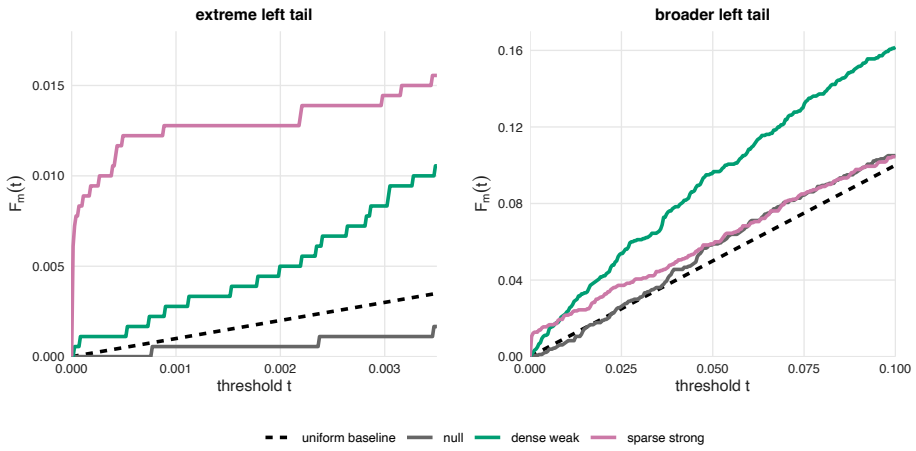



FIGURE 3.1. Left-tail ECDFs of one-sided p-values under a null model, a dense weak alternative, and a sparse stronger alternative. In this stylized draw, $m = 1800$, the dense weak case shifts 360 coordinates by 1.0, and the sparse strong case shifts 24 coordinates by 4.2. The diagonal is the uniform baseline. The extreme-left panel shows the sparse alternative creating the sharpest bump very near zero; the broader-left panel shows the dense weak alternative lifting the ECDF over a wider range.

## 2. Simes as a Crossing Rule

**Route: Core**

The Simes test rejects the global null when at least one ordered p-value crosses the line $i\alpha/m$:

$$p_{(i)} \le \frac{i\alpha}{m} \quad \text{for some } 1 \le i \le m.$$

Equivalently, define the Simes p-value

$$p_{\text{Simes}} = \min_{1 \le i \le m} \frac{m p_{(i)}}{i}.$$

The level-$\alpha$ Simes test rejects when $p_{\text{Simes}} \le \alpha$.

Figure 3.2 shows the same rule rank by rank. The first crossing encountered while scanning the ordered p-values is sufficient to reject the global null, while its rank records the tail scale at which the evidence became strong enough.

Simes is a global test in this chapter. It does not by itself say which hypotheses are false. The same boundary will reappear in Chapters 4 and 5, where it is used inside closure and FDR procedures to produce individual discoveries.

**Theorem 3.1: Simes under independent uniforms**

If $p_1, \dots, p_m$ are independent $\text{Unif}(0,1)$ random variables, then

$$\mathbb{P}\left\{ \min_{1 \le i \le m} \frac{m p_{(i)}}{i} \le \alpha \right\} = \alpha, \qquad 0 \le \alpha \le 1.$$

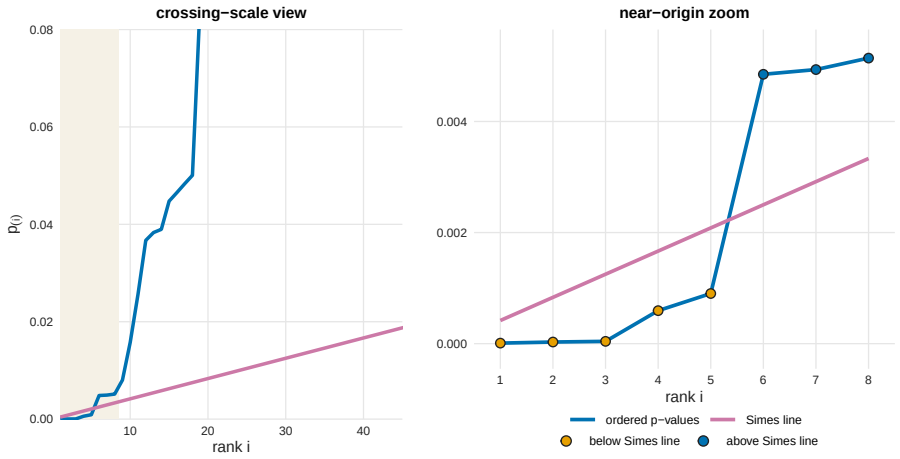



FIGURE 3.2. The Simes crossing rule. Ordered p-values are compared with the line $\alpha i/m$. A crossing gives a global rejection. Later, the same linear boundary becomes the basis of the BH multiple-testing procedure. The illustration uses $m = 120$, $\alpha = 0.05$, and one-sided p-values from eight Gaussian coordinates with mean 2.7 and 112 with mean zero. The left panel displays ranks 1–45 on the crossing-relevant vertical range from 0 to 0.08, so larger ordered p-values lie above its frame; its shaded first eight ranks are enlarged in the right panel.

**Proof of Theorem 3.1**

Let

$$T_m = \min_{1\le i\le m} \frac{m p_{(i)}}{i}.$$

We prove by induction that $T_m \sim \mathrm{Unif}(0,1)$. The case $m = 1$ is immediate.
Assume the result is true for $m-1$. Fix $0 < \alpha < 1$ and split the event $\{T_m \le \alpha\}$ according to whether the largest order statistic is at most $\alpha$:

$$\mathbb{P}(T_m \le \alpha) = \mathbb{P}\left(\min_{1\le i\le m-1} \frac{m p_{(i)}}{i} \le \alpha,\ p_{(m)} > \alpha\right) + \mathbb{P}\{p_{(m)} \le \alpha\}.$$

The second term is $\alpha^m$. The density of $p_{(m)}$ is $mt^{m-1}$, $0 < t < 1$. Conditional on $p_{(m)} = t$, the remaining $m-1$ order statistics have the same distribution as the order statistics of $m-1$ independent uniforms on $[0,t]$. Equivalently,

$$\left(\frac{p_{(1)}}{t}, \ldots, \frac{p_{(m-1)}}{t}\right)$$

has the order-statistic distribution of $m-1$ independent $\mathrm{Unif}(0,1)$ variables. Therefore, by the induction hypothesis,

$$\min_{1\le i\le m-1} \frac{(m-1)p_{(i)}}{t\, i} \sim \mathrm{Unif}(0,1) \qquad \text{conditional on } p_{(m)} = t.$$

On the event $p_{(m)} > \alpha$, the crossing among the first $m-1$ order statistics is equivalent to

$$\min_{1\le i\le m-1} \frac{(m-1)p_{(i)}}{t\, i} \le \frac{(m-1)\alpha}{mt}.$$

For $t \in (\alpha, 1)$, the right-hand side lies in $(0, 1)$. Hence

$$\mathbb{P}\left(\min_{1\le i\le m-1} \frac{mp_{(i)}}{i} \le \alpha,\ p_{(m)} > \alpha\right)$$
$$= \int_\alpha^1 mt^{m-1}\,\frac{(m-1)\alpha}{mt}\,dt = (m-1)\alpha\int_\alpha^1 t^{m-2}\,dt = \alpha(1-\alpha^{m-1}).$$

Adding the case $p_{(m)} \le \alpha$ gives

$$\mathbb{P}(T_m \le \alpha) = \alpha(1-\alpha^{m-1}) + \alpha^m = \alpha.$$

Thus $T_m$ is exactly uniform on $[0, 1]$. □

The exact statement uses independence and continuity. Under certain positive dependence conditions, Simes remains conservative rather than exact. That dependence issue is one of the main topics of Chapters 4 and 6.

## 3. Empirical Processes and Classical Goodness-of-Fit Tests

**Route: Advanced**

Under independent uniform p-values, the centered empirical process

$$\alpha_m(t) = \sqrt{m}\{F_m(t) - t\}$$

converges weakly to a Brownian bridge $B(t)$, a mean-zero Gaussian process with covariance

$$\mathrm{Cov}\{B(s), B(t)\} = \min(s, t) - st.$$

This limit is the organizing principle behind classical uniformity tests.

**Proposition 3.2: Empirical-Process Covariance**

If $p_1, \ldots, p_m$ are independent uniforms, then for fixed $s, t \in [0, 1]$,

$$\mathrm{Cov}\{\alpha_m(s), \alpha_m(t)\} = \min(s, t) - st.$$

**Proof of Proposition 3.2**

For each $j$, set $I_j(t) = \mathbf{1}\{p_j \le t\}$. Since the observations are independent,

$$\mathrm{Cov}\{F_m(s), F_m(t)\} = \frac{1}{m}\,\mathrm{Cov}\{I_1(s), I_1(t)\}.$$

For $s \le t$, $I_1(s)I_1(t) = I_1(s)$, so

$$\mathrm{Cov}\{I_1(s), I_1(t)\} = s - st.$$

Multiplying by $m$ gives the covariance of $\alpha_m$. The case $t < s$ is symmetric. □

This exact finite-$m$ identity has two roles. Its diagonal case $s = t$ gives $\mathrm{Var}\{\alpha_m(t)\} = t(1-t)$, motivating the $t(1-t)$ standardization in AD and HC. The full kernel $\min(s, t) - st$ also shows that deviations at different thresholds are correlated, so a scan or integral across thresholds needs process-level rather than pointwise calibration.

The one-sided KS statistic is

$$D_m^+ = \sup_{0\le t\le 1}\{F_m(t) - t\}.$$

It looks for the largest vertical gap above the diagonal. The two-sided version uses $\sup_t |F_m(t)-t|$, but for detecting an excess of small p-values the one-sided statistic is the relevant form.

The CvM statistic averages squared deviations:

$$W_m = m \int_0^1 \{F_m(t) - t\}^2 \, dt.$$

A one-sided version may replace the squared deviation by $\{F_m(t)-t\}_+^2$. This averaging makes the test less dependent on a single rank than KS, but it can dilute a very narrow left-tail departure.

The AD statistic adds tail weights:

$$A_m = m \int_0^1 \frac{\{F_m(t) - t\}^2}{t(1-t)} \, dt.$$

Near zero, the binomial variance of $F_m(t)$ is approximately $t/m$, so division by $t(1-t)$ gives more attention to the small-p-value region. This is the first hint of HC: if the alternatives are rare, the left tail should be standardized more aggressively than the center of the distribution. Section 4 turns this idea into an adaptive scan over left-tail thresholds.

For computation, $A_m$ has a useful ordered-p-value form. Splitting the integral over the intervals between order statistics gives

$$A_m = -m - \frac{1}{m} \sum_{i=1}^{m} (2i-1) \left\{ \log p_{(i)} + \log(1 - p_{(m+1-i)}) \right\}.$$

This formula is also good intuition: the logarithms heavily penalize unusually small $p_{(i)}$'s and unusually large $p_{(m+1-i)}$'s. In the one-sided signal-detection setting we mostly care about the small-p-value contribution, but the classical AD statistic is symmetric in the two tails.

## 4. Higher Criticism, Berk-Jones, and Average Likelihood Ratios

**Route: Advanced**

HC asks whether there is some small threshold at which the data contain more p-values than the global null would predict. Let

$$N(t) = \#\{j : p_j \le t\} = mF_m(t).$$

Under the independent global null, $N(t) \sim \mathrm{Bin}(m, t)$, so its mean is $mt$ and its standard deviation is $\sqrt{mt(1-t)}$. The standardized excess count is therefore

$$\frac{N(t) - mt}{\sqrt{mt(1-t)}} = \frac{\sqrt{m}\{F_m(t) - t\}}{\sqrt{t(1-t)}}.$$

If the useful cutoff and the number of signals were known, one could choose a single $t$. In sparse problems they are usually unknown, so HC scans a range of left-tail thresholds and reports the largest positive standardized excess. Assume $m \ge 2$. A common ordered-p-value form is

$$\mathrm{HC}_m = \max_{1 \le i \le \lfloor m/2 \rfloor} \left[ \frac{\sqrt{m}\{i/m - p_{(i)}\}}{\sqrt{p_{(i)}(1-p_{(i)})}} \right]_+ = \max_{\substack{1 \le i \le \lfloor m/2 \rfloor \\ p_{(i)} < i/m}} \frac{\sqrt{m}\{i/m - p_{(i)}\}}{\sqrt{p_{(i)}(1-p_{(i)})}}.$$

The two forms agree because only ranks with $p_{(i)} < i/m$ contribute a positive deviation; ranks where the ECDF is at or below the diagonal are skipped. Here $[x]_+ = \max(x, 0)$, and a maximum over no qualifying ranks is defined to be zero. The displayed fractions are evaluated for $0 < p_{(i)} < 1$. Endpoint values are interpreted by limits: a zero p-value contributes $+\infty$, whereas a p-value of one contributes zero to this one-sided scan. Under independent continuous uniform nulls, these endpoints occur with probability zero.

Under the global null, the maximum is not $O(1)$. Each standardized deviation $\sqrt{m}\{F_m(t) - t\}/\sqrt{t(1-t)}$ is approximately $N(0,1)$ for fixed $t$, but the supremum over the left tail picks up extreme values from a continuum of $t$'s. A law-of-the-iterated-logarithm argument of Darling–Erdős, as used in the HC analysis of Donoho and Jin [46, 52], shows that

$$\mathrm{HC}_m = \sqrt{2\log\log m}\,\{1 + o_P(1)\} \qquad \text{under } H_0.$$

The extra $\sqrt{\log\log m}$ factor over the pointwise normal scale is the price of scanning many tail thresholds. In finite samples, the statistic is usually calibrated by simulation under a specified joint-null model or by a conservative asymptotic threshold such as $\sqrt{2(1+\delta)\log\log m}$, for a fixed slack $\delta > 0$. The important point is that HC scans over many possible tail thresholds without committing to one sparsity level.

BJ replaces the normal approximation in HC by an exact binomial likelihood ratio. Let

$$h(x,t) = x\log\frac{x}{t} + (1-x)\log\frac{1-x}{1-t}, \qquad 0 \le x \le 1, \quad 0 < t < 1,$$

with $0\log 0 = 0$ for a vanishing term. Thus $h(0,t) = -\log(1-t)$ and $h(1,t) = -\log t$. This is the Kullback–Leibler (KL) divergence $D(\mathrm{Bernoulli}(x) \,\|\, \mathrm{Bernoulli}(t))$, so $m\,h(x,t)$ is the binomial log-likelihood ratio at the observed count $mx$, comparing the empirical success probability $x$ with the null success probability $t$. For the one-sided alternative of an excess count, use

$$h_+(x,t) = \begin{cases} h(x,t), & x > t, \\ 0, & x \le t. \end{cases}$$

This assigns zero evidence to a shortage of small p-values. The one-sided BJ statistic is

$$\mathrm{BJ}_m = \max_{1 \le i \le \lfloor m/2 \rfloor} m\,h_+\left(\frac{i}{m}, p_{(i)}\right).$$

It is zero when no rank shows an excess. As with HC, a zero p-value gives an infinite contribution, while a p-value of one gives zero. The natural reading of this scan: think first of a *fixed* threshold $t \in (0,1)$, under which $N(t) = \#\{j : p_j \le t\}$ is binomial with success probability $t$ under the global null. Then $N(t)/m$ is its sample proportion, and the binomial log-likelihood ratio comparing the data-driven proportion to the null $t$ is $m\,h(N(t)/m,\, t)$. The BJ statistic plugs in the data-dependent threshold $t = p_{(i)}$ (so that $N(t)/m = i/m$ when the p-values are distinct) at each rank $i$ and takes the maximum. Its null cutoff must account for this entire scan, rather than use the null distribution for one fixed threshold.

Figure 3.3 compares what the different weightings buy in finite samples. Compare methods within each alternative: the dense design rewards statistics that accumulate a broad departure, while the rare design rewards statistics that search aggressively over the extreme left tail.

An ALR statistic smooths this idea across ranks. Choose positive weights $w_i$ before examining the p-values, with $\sum_{i=1}^{\lfloor m/2 \rfloor} w_i = 1$, and define

$$\mathrm{ALR}_m = \sum_{i=1}^{\lfloor m/2 \rfloor} w_i \exp\left[m\,h_+\left(\frac{i}{m}, p_{(i)}\right)\right].$$

An inactive rank contributes $w_i \exp(0) = w_i$, so the statistic is one when no rank shows an excess. Calibrate the complete weighted statistic under the specified joint null, using the same weights throughout. The weights $w_i$ keep the average from being dominated by the many ranks available away from the extreme tail. A convenient way to think about the left tail is by dyadic blocks of ranks:

$$i \in [1,2),\ [2,4),\ [4,8), \ldots.$$

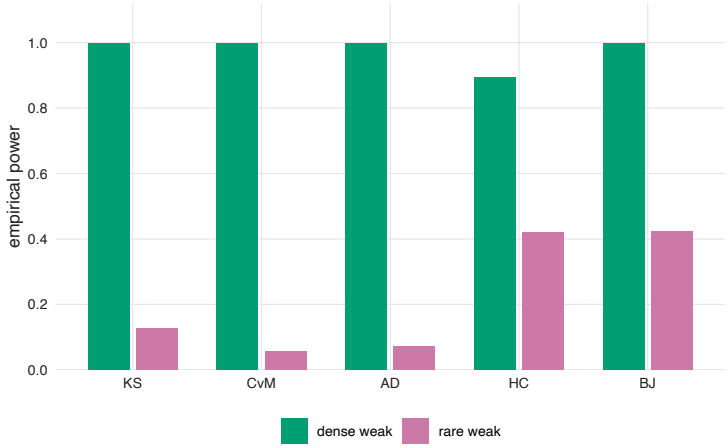



FIGURE 3.3. Empirical power for KS, CvM, AD, HC, and BJ with $m = 700$ one-sided Gaussian p-values and nominal level 0.05. Each method's cutoff is its empirical 95th percentile from 1,200 null replicates of 700 iid Unif$(0, 1)$ p-values; power uses 900 replicates per alternative. In each alternative replicate, the dense design randomly locates 126 $N(1, 1)$ signals, and the rare design randomly locates 8 signals with mean $\sqrt{0.72 \log 700}$; all remaining coordinates are $N(0, 1)$.

There are $O(\log m)$ such resolution levels up to $m/2$, and the block $[2^k, 2^{k+1})$ contains order $2^k$ ranks. Weights of order $1/i$ therefore give each dyadic block comparable total weight, while the extra $1/\log m$ normalizes across the logarithmic number of blocks. Thus a normalized choice is

$$w_i = \frac{i^{-1}}{\sum_{j=1}^{\lfloor m/2 \rfloor} j^{-1}}, \qquad 1 \le i \le \lfloor m/2 \rfloor.$$

HC takes the largest standardized tail deviation, BJ takes the largest binomial likelihood ratio, and ALR averages likelihood ratios across tail locations. These are not cosmetic differences: they behave similarly in the rare/weak asymptotic theory, but in finite samples averaging can be less sensitive to a single noisy rank or cutoff than a pure maximum.

## 5. Rare and Weak Gaussian Mixtures

**Route: Advanced**

We now connect the p-value empirical process to the Gaussian sequence model. Let

$$Z_j \sim (1 - \varepsilon_m)N(0, 1) + \varepsilon_m N(\mu_m, 1), \qquad j = 1, \dots, m,$$

independently, and use one-sided p-values $p_j = 1 - \Phi(Z_j)$. The rare/weak parameterization is

$$\varepsilon_m = m^{-\beta}, \qquad \mu_m = \sqrt{2r \log m},$$

where $1/2 < \beta < 1$ and $r > 0$. The expected number of signals is

$$m\varepsilon_m = m^{1-\beta}.$$

Thus $\beta$ controls sparsity and $r$ controls signal strength on the extreme-value scale.

The detection boundary for this model is

$$\rho^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1 - \beta})^2, & 3/4 \le \beta < 1. \end{cases}$$

For $r > \rho^*(\beta)$, there exist level-$\alpha$ tests whose power tends to one. For $r < \rho^*(\beta)$, no sequence of level-$\alpha$ tests is consistent: in the standard formulation of the lower bound, the total-variation distance between the null and the least-favorable mixture under the alternative tends to zero, so the sum of Type I and Type II errors of any sequence of tests cannot fall below one. This is the Ingster–Donoho–Jin rare/weak boundary [52, 93].

Figure 3.4 plots these two detection regimes. The solid curve is the optimal boundary; the dashed curve is the boundary reached by Bonferroni's maximum test.

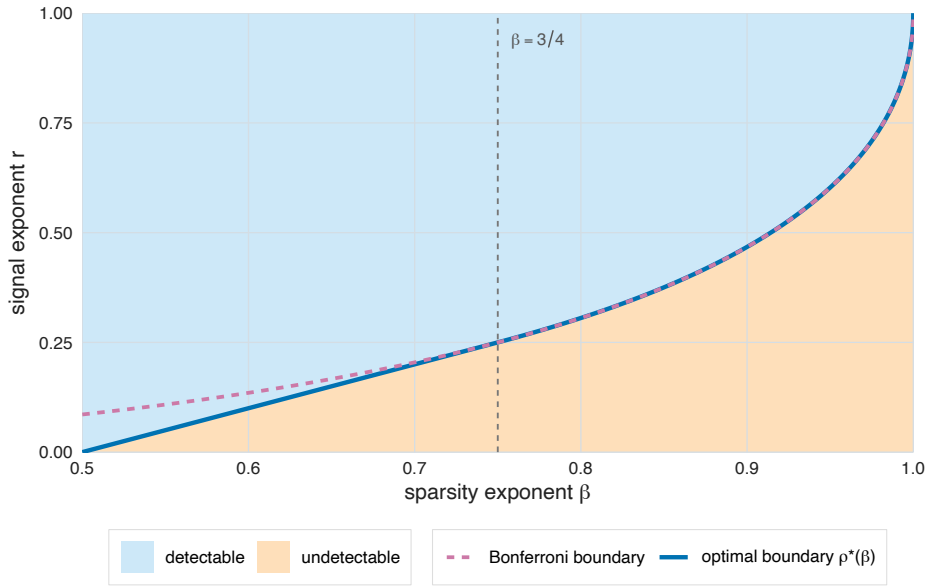



FIGURE 3.4. Rare/weak detection boundary for Gaussian mixtures. Above $\rho^*(\beta)$, detection is possible; below it, detection is impossible. The dashed curve is the boundary for Bonferroni's maximum test, which is rate-optimal in the very sparse region $3/4 \le \beta < 1$, but not throughout the whole rare/weak regime. The vertical dotted line marks $\beta = 3/4$, where the two boundary branches meet and Bonferroni's maximum test becomes rate-optimal.

The shape of the boundary explains the roles of the tests. When $\beta \ge 3/4$, the alternatives are so sparse that the best evidence is often an extreme coordinate. Bonferroni, or equivalently a maximum test, reaches the optimal boundary because the maximum among the signal coordinates can exceed the maximum-noise scale. The signal count is approximately $m\varepsilon_m = m^{1-\beta}$, each signal mean is $\sqrt{2r\log m}$, and the maximum of $m^{1-\beta}$ independent $N(0,1)$ noise terms concentrates near $\sqrt{2(1-\beta)\log m}$ by the same Mills-ratio calculation used in Chapter 2. Therefore

$$\max_{j\in\text{nonnull}} Z_j = \{\sqrt{2r\log m} + \sqrt{2(1-\beta)\log m}\}\{1 + o_P(1)\},$$

which exceeds the Bonferroni threshold $\sqrt{2\log m}\{1 + o(1)\}$ exactly when

$$\sqrt{r} + \sqrt{1-\beta} > 1, \qquad \text{equivalently} \qquad r > (1 - \sqrt{1-\beta})^2.$$

When $1/2 < \beta < 3/4$, the signals are still rare but not rare enough for the minimum p-value to be the whole story. The optimal boundary is $\beta - 1/2$, which lies below the Bonferroni boundary. HC and BJ recover this region by scanning over many tail cutoffs.

Here is the exponent calculation behind that statement. Fix a p-value cutoff $t = m^{-u}$, $0 < u < 1$. Under the null, the expected count below $t$ is

$$mt = m^{1-u}, \qquad \mathrm{sd}\{\#(p_i \le t)\} \asymp m^{(1-u)/2}.$$

For a nonnull observation $Z \sim N(\sqrt{2r\log m}, 1)$,

$$\mathbb{P}(p \le m^{-u}) = \mathbb{P}\{Z \ge \Phi^{-1}(1-m^{-u})\} \approx \mathbb{P}\{N(0,1) \ge \sqrt{2u\log m} - \sqrt{2r\log m}\}.$$

The approximation uses the normal tail quantile $\Phi^{-1}(1-m^{-u}) = \sqrt{2u\log m}\{1+o(1)\}$. At the boundary $u = r$, the sharper expansion

$$\Phi^{-1}(1-m^{-r}) = \sqrt{2r\log m} + O\left(\frac{\log\log m}{\sqrt{\log m}}\right) = \sqrt{2r\log m} + o(1)$$

justifies the limit stated below. If $u < r$, the threshold lies below the signal mean by a diverging amount, so a nonnull coordinate crosses with probability tending to one. At $u = r$, the threshold differs from the signal mean by $o(1)$, and the crossing probability tends to $1/2$. If $u > r$, Mills' ratio gives

$$\mathbb{P}\{N(0,1) \ge (\sqrt{2u} - \sqrt{2r})\sqrt{\log m}\} = m^{-(\sqrt{u}-\sqrt{r})^2 + o(1)}.$$

Thus the nonnull crossing probability is approximately

$$\begin{cases} 1 - o(1), & u < r, \\ 1/2 + o(1), & u = r, \\ m^{-(\sqrt{u}-\sqrt{r})^2+o(1)}, & u > r, \end{cases}$$

with the Mills-ratio line understood up to subpolynomial factors. The excess signal count has exponent

$$1 - \beta - (\sqrt{u} - \sqrt{r})_+^2,$$

whereas the null standard deviation has exponent $(1-u)/2$. At threshold $m^{-u}$, the numerator of the HC statistic is the observed count minus its null expectation, and the denominator is its null standard deviation. Thus the mean shift of the standardized statistic is the expected excess count divided by that standard deviation. For this signal-to-noise ratio to diverge, the excess-count exponent must be larger than the null-standard-deviation exponent; equivalently, HC is powerful when

$$1 - \beta - (\sqrt{u} - \sqrt{r})_+^2 > \frac{1-u}{2}.$$

Scanning over $u$ lets the procedure find the cutoff where this exponent gap is largest. To see the optimization, write the left minus right side as

$$G(u; r, \beta) = 1 - \beta - (\sqrt{u} - \sqrt{r})_+^2 - \frac{1-u}{2}.$$

For $u > r$,

$$G(u; r, \beta) = \frac{1}{2} - \beta - r - \frac{u}{2} + 2\sqrt{ur}.$$

The unconstrained maximizer is $u = 4r$. If $4r < 1$, the best exponent is $G(4r; r, \beta) = r - \beta + 1/2$, giving the condition $r > \beta - 1/2$. This is the relevant case when $1/2 < \beta < 3/4$. If the optimizer would lie beyond the search range, the best cutoff is at the endpoint $u = 1$, giving

$$G(1; r, \beta) = 1 - \beta - (1 - \sqrt{r})^2,$$

which is positive exactly when $r > (1 - \sqrt{1-\beta})^2$. This gives the second branch for $3/4 \le \beta < 1$. The full theorem in Donoho and Jin [52] controls the stochastic fluctuations uniformly over the scanned thresholds; the calculation here is the deterministic signal-to-noise comparison that explains the shape of the phase diagram.

**Theorem 3.3: Adaptive rare/weak detection, informal**

In the Gaussian mixture model above, HC has asymptotic power tending to one whenever $r > \rho^*(\beta)$, without knowing $\beta$ or $r$ in advance. If $r < \rho^*(\beta)$, no sequence of level-$\alpha$ tests has power tending to one uniformly over the corresponding rare/weak alternatives.

Why the boundary has this form. For a fixed threshold $t = m^{-u}$, the null count $mF_m(t)$ has mean approximately $m^{1-u}$ and standard deviation approximately $m^{(1-u)/2}$. Under the mixture, the nonnull coordinates add an excess count whose exponent depends on $\beta$, $r$, and $u$. HC standardizes the excess by the null standard deviation and then maximizes over $u$. The best cutoff yields a positive exponent exactly when $r > \rho^*(\beta)$. The impossibility statement is proved by likelihood-ratio or second-moment arguments for the mixture distribution under the null. The full proof is technical; the exponent calculation is the part that explains why a tail scan is adaptive.

## 6. Dependence Can Hurt, but It Can Also Help

**Route: Research frontier**

The preceding calibration assumes independent p-values. Dependence changes the fluctuations of $F_m(t)$, so an independence-calibrated HC threshold can be wrong. Positive correlation, for example, can make small p-values arrive in clusters even under the global null.

Figure 3.5 studies the HC test under first-order autoregressive (AR(1)) noise. The left panel compares the empirical size obtained with an iid-null cutoff and an AR(1)-null cutoff; the right panel compares the corresponding empirical power values under the same AR(1) alternative.

Dependence is not only a nuisance. In structured Gaussian problems it can make signals more visible if the statistic uses the covariance structure. The idea appears clearly in innovated HC. If $X = \mu + Z$ with $Z \sim N(0, \Sigma)$ and known positive-definite $\Sigma$, then a whitening operator $A_{\mathrm{wh}}$ satisfying $A_{\mathrm{wh}}\Sigma A_{\mathrm{wh}}^\top = I$ turns the noise white:

$$A_{\mathrm{wh}}X = A_{\mathrm{wh}}\mu + A_{\mathrm{wh}}Z, \qquad A_{\mathrm{wh}}Z \sim N(0, I).$$

Whitening makes the Gaussian noise coordinates independent, but it also changes the signal from $\mu$ to $A_{\mathrm{wh}}\mu$. The transformed signal need not remain sparse, so whitening alone does not guarantee a power gain. For suitable sparse alternatives and structured correlation models, carefully chosen transformations can improve detection relative to independent noise [78]. Ordinary HC applied to the untransformed coordinates can miss this opportunity as well as use an incorrect null calibration.

A related likelihood-ratio calculation explains why precision adjustment is a principled statistic for sparse high-dimensional mean testing. Consider $n$ independent Gaussian observations $X_1, \ldots, X_n \sim N_m(\mu, \Sigma)$, with known positive-definite covariance matrix $\Sigma$, and let $\Gamma = \Sigma^{-1}$ be the precision matrix. The sample mean $\bar{X} = n^{-1}\sum_{k=1}^n X_k$ has covariance $\Sigma/n$. Define the precision-adjusted vector $S_\mu = \Gamma\bar{X}$; then

$$S_\mu = \Gamma\bar{X} \qquad \text{so that} \qquad \mathbb{E}S_\mu = \Gamma\mu, \quad \mathrm{Cov}(S_\mu) = \Gamma/n.$$

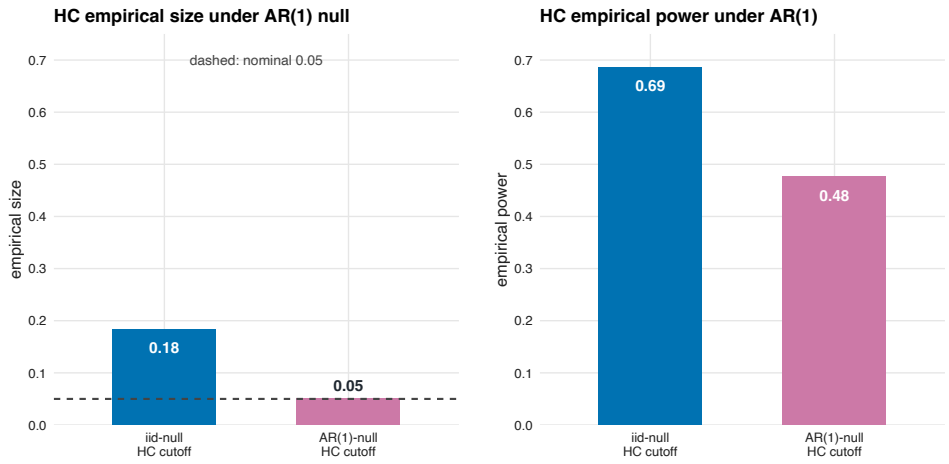



FIGURE 3.5. Empirical size and power of HC under dependence. The simulation uses $m = 400$ one-sided Gaussian p-values with AR(1) correlation $\rho = 0.85$. Both HC cutoffs are empirical 95th percentiles computed from 5,000 simulated null samples: independent uniforms for the iid-null cutoff and correlated Gaussian samples for the AR(1)-null cutoff. Under the AR(1) null, the iid-null cutoff is too low and has inflated empirical size. Under an AR(1) alternative with 8 randomly located signals of size 2.4, it likewise gives inflated apparent power; the AR(1)-null cutoff restores empirical size control and gives the calibrated empirical power. The dashed line in the left panel marks the nominal level 0.05. Each displayed empirical size and power value uses 5,000 independent simulation runs.

To see why this vector is called a score, drop terms that do not depend on $\mu$ and write the Gaussian log likelihood as

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2}\sum_{k=1}^{n}(X_k-\mu)^\top\Gamma(X_k-\mu)+C, \\ \nabla_\mu\ell(\mu) &= n\Gamma(\bar{X}-\mu), \\ \nabla_\mu\ell(\mu)\big|_{\mu=0} &= n\Gamma\bar{X} = nS_\mu. \end{aligned}$$

A score is the derivative of the log likelihood with respect to a parameter. Thus $S_{\mu,j}$, apart from the common factor $n$, measures the instantaneous change in log likelihood when $\mu_j$ moves away from zero while the other mean coordinates remain at zero. Its covariance is $\Gamma/n$, not the identity, so this is a precision adjustment rather than a whitening.

Now fix a candidate support $S$, a set of coordinates allowed to have nonzero means, and restrict the alternative to vectors satisfying $\mu_{S^c} = 0$. The likelihood equation for the active coordinates and its solution are

$$0 = n\{S_{\mu,S} - \Gamma_{S,S}\widehat{\mu}_S\}, \qquad \widehat{\mu}_S = \Gamma_{S,S}^{-1}S_{\mu,S}.$$

Substituting this restricted maximum-likelihood estimate into the log likelihood gives twice the maximized log-likelihood ratio,

$$T_S = 2\{\ell(\widehat{\mu}_S, 0_{S^c}) - \ell(0)\} = n\,S_{\mu,S}^\top\Gamma_{S,S}^{-1}S_{\mu,S}.$$

Under $\mu = 0$, $S_{\mu,S} \sim N(0, \Gamma_{S,S}/n)$. Consequently,

$$\sqrt{n}\,\Gamma_{S,S}^{-1/2}S_{\mu,S} \sim N(0, I_{|S|}), \qquad T_S = \left\|\sqrt{n}\,\Gamma_{S,S}^{-1/2}S_{\mu,S}\right\|_2^2 \sim \chi^2_{|S|}.$$

If the alternative is sparse but the support is unknown, the ideal sparse likelihood-ratio test scans over candidate sets $S$. For single-coordinate supports this reduces to the maximum of precision-adjusted standardized scores,

$$\max_{1\le j\le m} \frac{nS_{\mu,j}^2}{\gamma_{jj}},$$

where $\gamma_{jj}$ is the $j$th diagonal of $\Gamma$. This is the one-sparse case of the likelihood-ratio derivation in Zhang [200] and the basis of the two-sample precision-transformed maximum test of Cai et al. [37].

This likelihood-ratio view also gives the right interpretation of when dependence helps. For sparse or maximum-type tests, what matters after precision adjustment is not the raw mean $\mu_j$, but the leading standardized transformed coordinates

$$\frac{(\Gamma\mu)_j}{\sqrt{\gamma_{jj}}}.$$

Power improves when these leading coordinates are larger than the marginal signals, or when neighboring transformed coordinates carry additional evidence for the same sparse effect. Power can be worse for a maximum-type statistic when the precision transformation spreads the signal across many coordinates or creates cancellations that reduce the largest transformed coordinate. The choice of statistic should therefore match the alternative: max-type and sparse likelihood-ratio scans are aimed at sparse transformed evidence, while quadratic or sum-of-squares tests are better suited to broad dense shifts.

No observed p-value vector can verify the dependence assumptions in this chapter.

If dependence is structured, scientifically meaningful, and the covariance or precision structure is known or can be estimated accurately enough, build it into the statistic through precision adjustment or score whitening rather than treating it only as a calibration problem. In high-dimensional problems, estimating $\Gamma$ usually requires assumptions such as sparse precision rows or a sparse graphical model, with methods such as nodewise Lasso, constrained $\ell_1$ minimization for inverse matrix estimation (CLIME), or graphical Lasso. When those assumptions are doubtful, dependence-aware null calibration is safer than a poorly estimated precision transformation.

## 7. Assumptions in Plain Language

**Route: Core**

Uniform p-values are the baseline for this chapter. Exact Simes calibration uses independent continuous uniforms. Empirical-process approximations use large $m$, and the usual Brownian-bridge limit is an independence result. HC and BJ rely on tail standardization; their finite-sample thresholds should be calibrated by simulating from a justified joint-null model that reflects the relevant dependence. The rare/weak boundary is an asymptotic statement for an idealized Gaussian mixture. It is valuable because it explains which signal shapes different tests are designed to see.

| **Assumptions in practice** | |
|---|---|
| **What you can check** | If the joint null distribution is specified by the design or by a model chosen before seeing the results, simulate complete p-value vectors, recompute the statistic, and record the fraction that exceed the proposed cutoff. |

| | |
|---|---|
| **What these data cannot show** | One observed p-value vector cannot verify independence, the joint null model, or the accuracy of an extreme-tail asymptotic approximation. |
| **Safer fallback** | If only the validity of the individual p-values is justified, use Bonferroni for the global null; it does not require a dependence model. |

## 8. Bibliographic Notes

**Route: Core**

The Simes global test is due to Simes [164]. HC and the rare/weak detection boundary are presented in the form of Donoho and Jin [52]. The earlier minimax testing work of Ingster [93] is important background: it shows that sparse Gaussian testing can have likelihood-ratio limits outside the classical Gaussian regime, which is why tail-sensitive tests are not merely finite-sample heuristics. The $\sqrt{2\log\log m}$ null scale for HC follows from law-of-the-iterated-logarithm arguments going back to Darling and Erdős [46]. BJ statistics go back to Berk and Jones [25]; modern comparisons of HC, BJ, and related phi-divergence statistics include Jager and Wellner [96]. Innovated HC and the possibility that dependence can improve sparse-signal detection are developed by Hall and Jin [78]; related heterogeneous-mixture boundaries appear in Cai et al. [35]. Precision-transformed maximum and sparse likelihood-ratio tests for high-dimensional means are developed by Cai et al. [37] and Zhang [200]. Precision-matrix estimation methods used in such settings include nodewise Lasso [136], CLIME [36], and graphical Lasso [65].

## 9. Exercises

**Route: Core**

### Basic.

**Exercise 3.4: Simes crossing**

For $m = 5$ p-values

$$0.012,\ 0.018,\ 0.021,\ 0.31,\ 0.78,$$

compute $p_{\text{Simes}}$ and decide whether the Simes test rejects at $\alpha = 0.05$. Compare with Bonferroni.

### Intermediate.

**Exercise 3.5: Anderson-Darling ordered formula**

Derive the ordered-p-value expression for $A_m$ by integrating $\{F_m(t)-t\}^2/\{t(1-t)\}$ over the intervals $[0, p_{(1)}]$, $[p_{(i)}, p_{(i+1)}]$, and $[p_{(m)}, 1]$.

**Exercise 3.6: Berk-Jones likelihood ratio**

Fix $t \in (0,1)$, let $N(t) = mF_m(t)$, and consider the binomial model $N(t) \sim \text{Binomial}(m, \theta)$. The null is $\theta = t$. For an observed fraction $\widehat{\theta} = N(t)/m > t$, derive the binomial log likelihood ratio comparing $\theta = \widehat{\theta}$ with $\theta = t$:

$$m \left\{ \widehat{\theta} \log \frac{\widehat{\theta}}{t} + (1 - \widehat{\theta}) \log \frac{1 - \widehat{\theta}}{1 - t} \right\} = m\, h(\widehat{\theta}, t).$$

Start from the binomial probability mass function, using $0 \log 0 = 0$ at the endpoint. This is the fixed-threshold calculation behind the BJ scan in Section 4.

**Exercise 3.7: Simulation comparison**

Simulate one-sided Gaussian p-values under a sparse mixture. Choose and report the dimension, signal proportion, signal strength, and replication counts. Compare the chapter's KS statistic $D_m^+$, AD statistic $A_m$, HC, and BJ at level 0.05, using their displayed definitions. Estimate null cutoffs by Monte Carlo, then use independent simulation runs to estimate size and power. Report Monte Carlo SEs.

## Computational.

**Exercise 3.8: Average likelihood ratio**

Implement the ALR statistic with weights

$$w_i = \frac{i^{-1}}{\sum_{j=1}^{\lfloor m/2 \rfloor} j^{-1}}, \qquad 1 \le i \le \lfloor m/2 \rfloor,$$

fixed before observing the data. Calibrate its null cutoff by simulation and compare its rare/weak power with HC and BJ at the same nominal level. Use separate runs for calibration and evaluation. For numerical stability, subtract the largest exponent before exponentiating, then add it back after taking the logarithm of the weighted sum.

**Exercise 3.9: Rare/weak phase diagram**

Choose and report a fixed dimension $m$. For a grid with $1/2 < \beta < 1$ and $r > 0$, simulate independent Gaussian coordinates from the mixture with $\varepsilon_m = m^{-\beta}$ and $\mu_m = \sqrt{2r \log m}$, and form one-sided p-values. Estimate the power of Bonferroni, HC, and BJ at level 0.05, calibrating HC and BJ on independent null simulations. Report replication counts and Monte Carlo SEs. Overlay the theoretical boundary $\rho^*(\beta)$ and summarize the finite-sample power patterns relative to this asymptotic boundary.

**Exercise 3.10: Dependence calibration of HC**

Repeat the size comparison in the left panel of Figure 3.5 for $m \in \{200, 1000\}$. Generate mean-zero Gaussian vectors with unit variances and $\text{Corr}(Z_j, Z_k) = 0.5^{|j-k|}$, and form one-sided p-values $1 - \Phi(Z_j)$. Estimate level-0.05 HC cutoffs under the iid null and under this AR(1) null. Use fresh AR(1) null simulations to estimate the actual size of each cutoff, and report Monte Carlo SEs.

**Advanced.**

**Exercise 3.11: Dependence-aware sparse testing**

Let $X_1, \dots, X_n$ be independent $N_m(\mu, \Sigma)$ observations, with known $\Sigma \succ 0$ and sample mean $\bar{X}$. Compare the ordinary maximum statistic $\max_j n\bar{X}_j^2/\Sigma_{jj}$ with

$$\max_{1 \le j \le m} \frac{n(\Gamma \bar{X})_j^2}{\gamma_{jj}}, \qquad \Gamma = \Sigma^{-1}.$$

Calibrate both tests at the same nominal level, for example using Bonferroni $\chi_1^2$ cutoffs. Construct an example where precision adjustment increases power and one where it decreases power. For each example, compute $\max_j |\mu_j|/\sqrt{\Sigma_{jj}}$ and $\max_j |(\Gamma\mu)_j|/\sqrt{\gamma_{jj}}$, and use these values to interpret your power comparison.

CHAPTER 4

# Familywise Error Rate, Closure, and Graphical Procedures

Chapters 2 and 3 asked whether a collection of tests contains any signal. That is a global question. This chapter turns to settings in which the reported output is a set of individually named claims: which endpoints, genes, models, subgroups, or validation checks show convincing evidence? When the inferential goal is to guard against even one mistaken claim in that set, FWER is a natural criterion: it bounds the probability of making at least one false rejection.

We develop FWER control as an error-budgeting problem. Bonferroni divides the budget among the hypotheses. Holm uses the ordered p-values to spend it less conservatively. Closure explains why these procedures provide strong control, and graphical procedures give a practical language for prespecified testing plans in which alpha can move from one hypothesis to another after rejections.

The central issue is not merely that many p-values are present. It is that the reported claims are usually read together. A paper that declares three biomarkers significant, a trial report that claims benefit on a primary and two secondary endpoints, or a model audit that flags several failure modes is making a family of statements. Controlling FWER at level $\alpha$ means that, with probability at least $1 - \alpha$, none of the reported rejections is a false rejection. This guarantee is particularly useful when one false claim could trigger a costly follow-up experiment, regulatory conclusion, or scientific headline.

## 1. Why Strong Control Is Needed

**Route: Core**

We begin with a generic family, before introducing any priorities among its members. Let $m$ hypotheses be tested. For each hypothesis, there are two binary states: the null is true or false, and the procedure rejects or does not reject. The usual decision table is shown in Figure 4.1.

Write $\mathcal{I}_0$ for the unknown set of true-null indices, $m_0 = |\mathcal{I}_0|$ for its size, and $m_1 = m - m_0$ for the number of nonnulls. The decision table partitions the $m$ hypotheses into four cells according to whether the null is true or false and whether the procedure rejects or not:

$$U = \#\{i \in \mathcal{I}_0 : \text{not rejected}\}, \qquad V = \#\{i \in \mathcal{I}_0 : \text{rejected}\},$$
$$M = \#\{i \notin \mathcal{I}_0 : \text{not rejected}\}, \qquad D = \#\{i \notin \mathcal{I}_0 : \text{rejected}\},$$

with marginals $U + V = m_0$, $M + D = m_1$, and $R = V + D$ the total number of rejections. In FWER analysis, $V$ is the only random variable we directly bound, but the full table is the conceptual object that organizes every multiple-testing error rate. The FWER is

$$\text{FWER} = \mathbb{P}(V \geq 1),$$

and a procedure controls FWER at level $\alpha$ if $\mathbb{P}(V \geq 1) \leq \alpha$ for the relevant class of data-generating distributions.

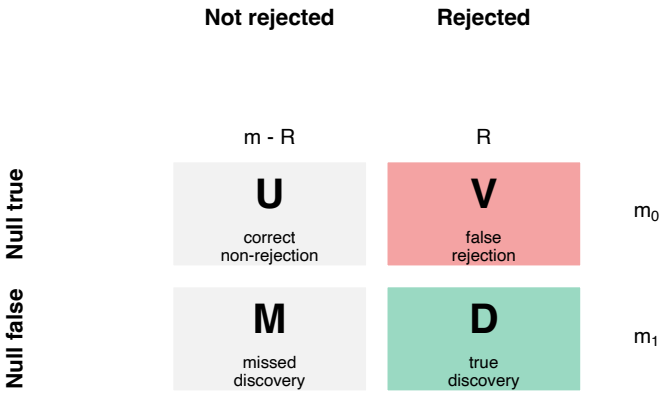



FIGURE 4.1. Multiple-testing decision table. Rows indicate whether the null hypothesis is true or false; columns indicate the procedure's decision. Among true nulls, $U$ are correctly not rejected and $V$ are falsely rejected; among false nulls, $M$ are missed discoveries and $D$ are true discoveries. Row totals are $m_0$ and $m_1 = m - m_0$; the column total of rejections is $R = V + D$.

**Definition 4.1: Weak and strong FWER control**

A procedure has weak FWER control at level $\alpha$ if

$$\mathbb{P}(V \geq 1) \leq \alpha$$

under the global null, when all individual null hypotheses are true. It has strong FWER control at level $\alpha$ if the same guarantee holds for every configuration of true and false null hypotheses.

Strong control is the useful guarantee for individual discoveries. It says that even if some alternatives are real, the probability of falsely rejecting any true null remains at most $\alpha$.

The distinction is not pedantic. Suppose a procedure first tests the global null at level $\alpha$, and, if the global null is rejected, reports every individual p-value below 0.05 without adjustment. Under the global null, this has weak control if the global test is valid. But if one nonnull hypothesis makes the global test reject almost surely, then the true nulls are effectively tested at unadjusted level 0.05. With $m_0$ true nulls and independent p-values, the chance of at least one false rejection is

$$1 - 0.95^{m_0},$$

which can be far above $\alpha$.

For example, with $m_0 = 20$ independent true nulls,

$$1 - 0.95^{20} \approx 0.642.$$

Thus the chance of at least one false rejection is about 64.2%, even though each true null is tested at the seemingly familiar 0.05 level.

One can also control generalized errors such as $k$-FWER,

$$\mathbb{P}(V \geq k),$$

but this chapter focuses on the classical $k = 1$ case.

## 2. Bonferroni, Sidak, and Holm

**Route: Core**

Let $p_1, \dots, p_m$ be valid p-values for hypotheses $H_1, \dots, H_m$, with true-null set $\mathcal{I}_0$ as above.

> **Remark 4.2: P-value validity recalled**
> Recall from Chapter 2 that validity requires $\mathbb{P}(p_i \le t) \le t$ for every $t \in [0,1]$ under every distribution satisfying $H_i$. The FWER arguments below use this marginal super-uniformity; they do not require exact uniformity unless stated otherwise.

The Bonferroni procedure rejects

$$H_i \quad \text{if} \quad p_i \le \frac{\alpha}{m}.$$

**Proposition 4.3: Bonferroni strong FWER control**

If the p-values for true nulls are super-uniform, then Bonferroni controls FWER at level $\alpha$ under arbitrary dependence.

**Proof of Proposition 4.3**

If any false rejection occurs, then $p_i \le \alpha/m$ for at least one $i \in \mathcal{I}_0$. Therefore

$$\mathbb{P}(V \ge 1) \le \sum_{i \in \mathcal{I}_0} \mathbb{P}(p_i \le \alpha/m) \le |\mathcal{I}_0| \frac{\alpha}{m} \le \alpha.$$

□

Under independence, the Sidak procedure replaces the threshold by

$$1 - (1-\alpha)^{1/m}.$$

With $m$ mutually independent, exactly $\mathrm{Unif}(0,1)$ p-values under the global null, this threshold gives FWER exactly $\alpha$. With independent super-uniform true-null p-values it is conservative and still gives strong control; neither the exactness nor the independence calculation is available under arbitrary dependence. Sidak is slightly less conservative than Bonferroni, but Holm's step-down procedure usually gives a more important gain.

Order the p-values as

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)},$$

with corresponding hypotheses $H_{(1)}, \dots, H_{(m)}$. Holm's procedure compares them with

$$\frac{\alpha}{m}, \quad \frac{\alpha}{m-1}, \quad \dots, \quad \alpha.$$

It starts at $p_{(1)}$. If

$$p_{(i)} \le \frac{\alpha}{m-i+1},$$

it continues to the next rank. It stops at the first failure and rejects all hypotheses before that failure.

Figure 4.2 should be read as a stopping-rule diagram: the critical values widen as fewer hypotheses remain, but a single ordered p-value above its current bound closes the path to every later rank.

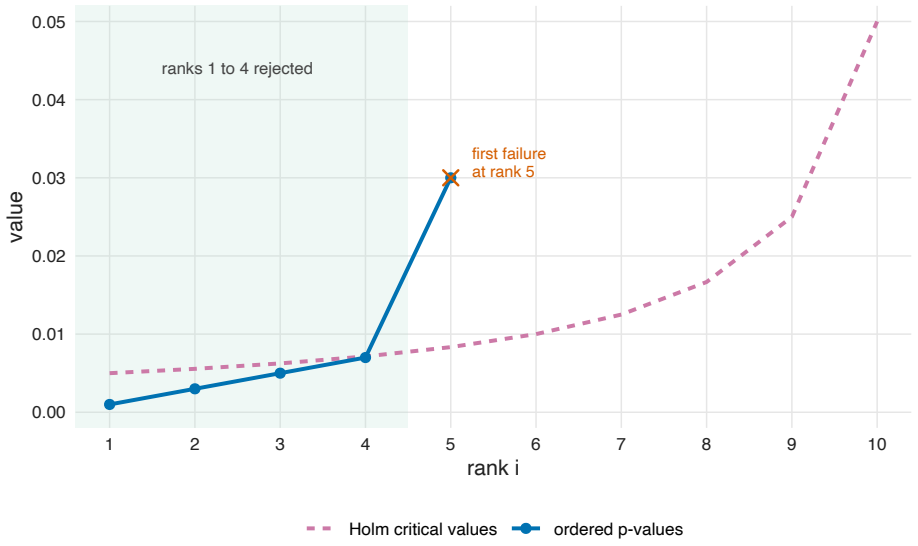



FIGURE 4.2. Holm's step-down rule for $m = 10$ and $\alpha = 0.05$. The solid blue curve shows the ordered p-values through the first failure, at rank 5 (red $\times$); the shaded ranks 1–4 are rejected, and later p-values are not inspected. The dashed purple curve shows all Holm critical values $\alpha/(m-i+1)$, ending at $\alpha$ for rank 10.

**Theorem 4.4: Holm strong FWER control**

Holm's procedure controls FWER at level $\alpha$ under arbitrary dependence whenever the true-null p-values are super-uniform.

**Proof of Theorem 4.4**

If $m_0 = 0$, then $V = 0$ and there is nothing to prove. Otherwise fix the ordering used by Holm, including any tie-breaking, and let $i^*$ be the first rank occupied by a true null:

$$i^* = \min\{i : H_{(i)} \text{ is true}\}.$$

Then $i^* \le m - m_0 + 1$, because only the $m - m_0$ false nulls can appear before the first true null. If Holm makes any false rejection, its step-down path must reach and reject this first true null: all earlier rejected hypotheses may be false nulls, but the first false rejection can only occur when the scan first reaches a true null. Hence

$$p_{(i^*)} \le \frac{\alpha}{m - i^* + 1}.$$

The rank bound gives

$$i^* \le m - m_0 + 1, \qquad \frac{\alpha}{m - i^* + 1} \le \frac{\alpha}{m_0}.$$

A false rejection therefore implies

$$p_{(i^*)} \le \frac{\alpha}{m_0}.$$

Since $p_{(i^*)}$ is one of the true-null p-values, the last event is contained in

$$\left\{\min_{j\in\mathcal{I}_0} p_j \le \frac{\alpha}{m_0}\right\}.$$

By Bonferroni applied to the $m_0$ true-null p-values, this event has probability at most $m_0 \cdot (\alpha/m_0) = \alpha$. □

## 3. Step-Up Procedures and Simes

**Route: Core**

Holm is a step-down procedure: it starts with the smallest p-value and moves toward larger ones until it fails. Hochberg's procedure uses the same critical values but scans in the opposite direction. It finds the largest $i$ such that

$$p_{(i)} \le \frac{\alpha}{m-i+1},$$

and rejects $H_{(1)}, \dots, H_{(i)}$.

Figure 4.3 isolates the consequence of scan direction. Holm advances only while each successive comparison from the left succeeds, whereas Hochberg can use one successful comparison found from the right to reject the entire prefix, which explains both its possible power gain and its stronger dependence requirements.

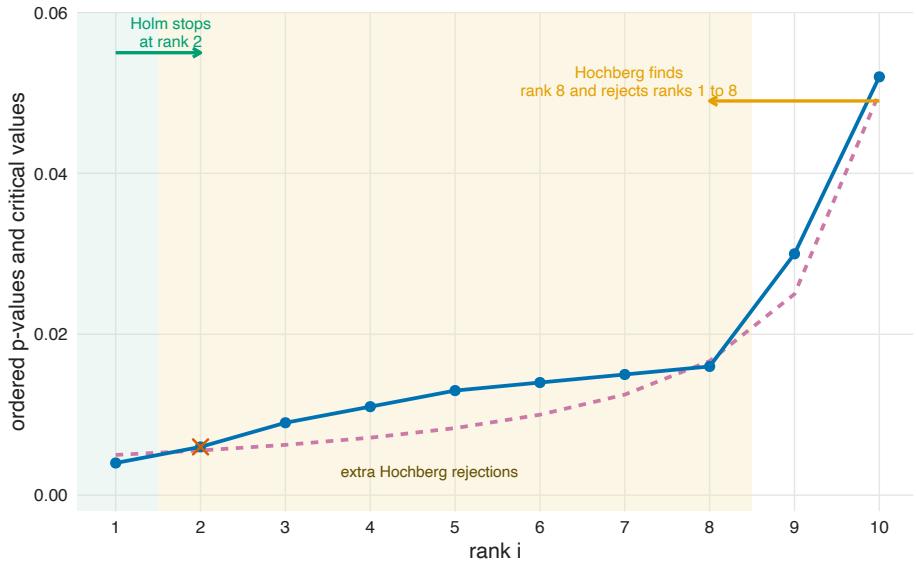



FIGURE 4.3. Step-down and step-up scans using the same critical curve. Holm starts from the smallest p-value and stops at the first failure, here at rank 2. Hochberg starts from the largest p-value and looks backward for the first success, here at rank 8, so it rejects ranks 1 through 8. The illustration uses $m = 10$ and $\alpha = 0.05$: the solid blue curve shows the ordered p-values, the dashed purple curve shows the common critical values, and the green and orange arrows show the two scan directions.

Hochberg can reject more hypotheses than Holm, but it is not an arbitrary-dependence procedure. Its validity relies on Simes tests being valid within the relevant intersections. Recall

from Chapter 3 that the Simes global test looks for a crossing of $i\alpha/m$ by an ordered p-value. It is valid for independent p-values and for several forms of positive dependence, but not under arbitrary dependence. We return to the exact closed-testing interpretation after introducing the closure principle. The important point for now is that Hochberg is a step-up rule whose validity comes from Simes-type assumptions, unlike Holm's arbitrary-dependence guarantee.

There is also a useful warning. The BH procedure is the step-up FDR procedure that, at target $q$, rejects $H_{(1)}, \ldots, H_{(k)}$, where $k$ is the largest rank satisfying

$$p_{(k)} \leq \frac{kq}{m}.$$

It is studied in detail in Chapter 5. For the weak-FWER comparison, set $q = \alpha$. Under the global null, BH then rejects at least one hypothesis exactly when the level-$\alpha$ Simes global test rejects. Under independence, or under PRDS (see Chapter 6, Section 1), the Simes inequality makes this a *weak* FWER guarantee: under the global null, the probability that BH rejects any hypothesis is at most $\alpha$. Weak control is not strong control: when some nulls are false, BH may reject true nulls with probability larger than $\alpha$. The dependence assumption is essential: under arbitrary dependence, Simes itself can fail, so this weak guarantee is not free.

## 4. The Closure Principle

**Route: Core**

Closure is the master theorem behind many FWER procedures. For every nonempty set $I \subseteq \{1, \ldots, m\}$, define the intersection hypothesis

$$H_I = \bigcap_{i \in I} H_i.$$

A *local test* is a test of one intersection hypothesis $H_I$ considered on its own. It has level $\alpha$ if its probability of rejection is at most $\alpha$ whenever all component nulls in $I$ are true. A closed testing procedure specifies such a local test for every intersection $H_I$, and it rejects an elementary hypothesis $H_i$ if and only if every intersection $H_I$ containing $i$ is rejected by its local test.

Figure 4.4 makes the proof mechanism visible for three hypotheses. To audit an elementary rejection, trace every upward path through intersections containing it; strong control follows because any false elementary rejection must force rejection of the single intersection containing all true nulls.

**Theorem 4.5: Closure principle**

If every intersection hypothesis $H_I$ is tested by a level-$\alpha$ local test, then the closed testing procedure controls FWER at level $\alpha$ strongly.

**Proof of Theorem 4.5**

Let $\mathcal{I}_0$ be the set of true null hypotheses, and note that $\mathcal{I}_0$ is among the intersection indices considered by the closed procedure (it is nonempty whenever at least one null is true; the case $\mathcal{I}_0 = \emptyset$ gives $V = 0$ deterministically). If the closed procedure makes any false rejection, then it rejects some $H_i$ with $i \in \mathcal{I}_0$. By the rule of closed testing, it must then reject every

intersection containing $i$, in particular the true intersection

$$H_{\mathcal{I}_0} = \bigcap_{i \in \mathcal{I}_0} H_i.$$

Under the true data-generating distribution, $H_{\mathcal{I}_0}$ is true, so its local test has Type I error at most $\alpha$. Therefore

$$\mathbb{P}(V \geq 1) \leq \mathbb{P}(\text{reject } H_{\mathcal{I}_0}) \leq \alpha.$$

□

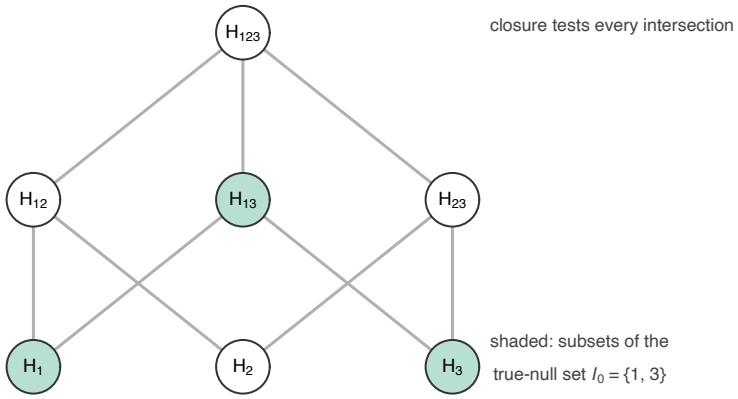



FIGURE 4.4. Closure lattice for three hypotheses. To reject $H_1$, closed testing requires rejection of every intersection containing 1: $H_1, H_{12}, H_{13}, H_{123}$. In the example shading, the true-null set is $\mathcal{I}_0 = \{1, 3\}$, so the shaded nodes are exactly the intersection hypotheses $H_I$ with $I \subseteq \mathcal{I}_0$; the rejection probability of any of them under the data-generating distribution is at most $\alpha$, and this is the lever that controls FWER.

The proof is short because closure reduces strong control to one local test: the intersection of all true nulls. The cost is computational and conceptual. There are $2^m - 1$ intersections, so useful procedures are those where the closed rule has a shortcut.

## 5. Closed Bonferroni, Closed Simes, and Hommel

**Route: Core**

Closed Bonferroni tests each intersection $H_I$ by Bonferroni within that intersection:

$$\min_{i \in I} p_i \leq \frac{\alpha}{|I|}.$$

The closed procedure is exactly Holm.

**Proposition 4.6: Closed Bonferroni equals Holm**

The closed testing procedure obtained by applying Bonferroni to every intersection hypothesis rejects exactly the same elementary hypotheses as Holm's step-down procedure.

**Proof of Proposition 4.6**

We verify both inclusions.
*Every Holm rejection is a closed-Bonferroni rejection.* Suppose Holm rejects $H_{(1)}, \dots, H_{(k)}$, and fix an index $(j)$ with $j \le k$. For any intersection $I$ containing $(j)$, let

$$h = \min\{r : H_{(r)} \in I\},$$

the smallest rank index appearing in $I$. Then $h \le j \le k$, so Holm rejected $H_{(h)}$, which means

$$p_{(h)} \le \frac{\alpha}{m-h+1}.$$

Moreover $I \subseteq \{(h), (h+1), \dots, (m)\}$, so $|I| \le m-h+1$. Combining these,

$$p_{(h)} \le \frac{\alpha}{m-h+1} \le \frac{\alpha}{|I|},$$

and the Bonferroni local test for $H_I$ rejects. Therefore closed Bonferroni rejects every $H_{(j)}$ for $j \le k$.
*No Holm nonrejection is recovered by closure.* Suppose Holm stops at rank $k$, meaning $p_{(k)} > \alpha/(m-k+1)$. Consider the intersection

$$I^* = \{(k), (k+1), \dots, (m)\}, \qquad |I^*| = m-k+1.$$

Every p-value in $I^*$ satisfies $p_{(j)} \ge p_{(k)} > \alpha/(m-k+1) = \alpha/|I^*|$, so the Bonferroni local test for $H_{I^*}$ does not reject. Since $I^*$ contains $(k)$, the closed procedure cannot reject $H_{(k)}$ either. The same nonrejected intersection $I^*$ contains every $H_{(j)}$ with $j \ge k$, so none of the hypotheses after Holm's stopping rank can be rejected by closure. Thus closure adds no extra rejections beyond Holm, but it explains why Holm is strongly valid. □

Closed Simes uses the Simes test as the local test for every intersection:

$$\min_{1 \le j \le |I|} \frac{|I| p_{(j:I)}}{j} \le \alpha,$$

where $p_{(1:I)} \le \cdots \le p_{(|I|:I)}$ are the p-values inside intersection $I$. The resulting closed procedure is Hommel's procedure. Thus Hommel, not Hochberg, is the exact shortcut for the closure of Simes local tests. Hochberg is a simpler and more conservative step-up shortcut: it rejects only hypotheses that Hommel also rejects.

**Proposition 4.7: Hochberg is contained in closed Simes**

Suppose the Simes local test is valid for every intersection hypothesis. Then every rejection made by Hochberg is also made by the closed Simes procedure, equivalently Hommel's procedure. Consequently Hochberg controls FWER at level $\alpha$ under the same Simes-validity assumptions.

**Proof of Proposition 4.7**

Let $p_{(1)} \le \cdots \le p_{(m)}$. Suppose Hochberg rejects $H_{(j)}$. Then there exists $r \ge j$ such that

$$p_{(r)} \le \frac{\alpha}{m-r+1}.$$

To show that closed Simes rejects $H_{(j)}$, fix any intersection $I$ containing $(j)$, and write $h = |I|$. We must show that the Simes local test rejects $H_I$.
Among all intersections of size $h$ that contain $(j)$, the hardest one to reject is the one that includes $H_{(j)}$ and the largest possible p-values. Call this set $K_h$. If $j \le m-h+1$, take

$$K_h = \{(j), (m), (m-1), \ldots, (m-h+2)\};$$

otherwise take $K_h = \{(m), (m-1), \ldots, (m-h+1)\}$. The ordered p-values inside any other $I$ are coordinatewise no larger than those inside $K_h$, so rejection of $K_h$ implies rejection of $I$.
If $r \le m-h+1$, then $p_{(j)} \le p_{(r)} \le \alpha/(m-r+1) \le \alpha/h$, so the first Simes comparison rejects $K_h$. If $r > m-h+1$, then $K_h$ contains $p_{(r)}$ and all $m-r$ larger ordered p-values. Thus

$$p_{(h-m+r:K_h)} \le p_{(r)}.$$

Let $a = m-r+1$ and $b = h-m+r$. Then $a+b-1 = h$, so

$$ab - h = (a-1)(b-1) \ge 0.$$

Consequently

$$p_{(h-m+r:K_h)} \le \frac{\alpha}{m-r+1} \le \frac{(h-m+r)\alpha}{h},$$

which is exactly a Simes rejection for $K_h$. Therefore every intersection containing $(j)$ is rejected by Simes, so closed Simes rejects $H_{(j)}$. Since closed Simes controls FWER by Theorem 4.5, any procedure whose rejections are a subset of closed Simes rejections also controls FWER. □

Because Simes is never less powerful than Bonferroni on the same intersection, closed Simes rejects at least everything closed Bonferroni rejects. Therefore, under the dependence conditions where Simes local tests are valid, Hommel is at least as powerful as Hochberg, and Hochberg is at least as powerful as Holm. The price is that Hommel's exact shortcut is less transparent than Holm's, and Hochberg trades away some of Hommel's power for a simpler step-up rule.

The hierarchy is therefore:

$$\text{Holm} \quad \subseteq \quad \text{Hochberg} \quad \subseteq \quad \text{Hommel},$$

with the important caveat that the latter two require Simes-type validity, not just arbitrary super-uniform p-values.

## 6. Graphical Error Spending

**Route: Core**

Closure is general, but it is not always the most convenient design language. In structured confirmatory studies, analysts often know the priority order: primary endpoint first, then secondary endpoints, then subgroups or supportive analyses. Graphical procedures encode this structure directly.

For example, a platform-trial protocol may specify that secondary endpoints are tested only after the primary endpoint is rejected, and that a related subgroup is tested only after a

designated secondary endpoint is rejected. This is *prespecified gatekeeping*: the priorities, local levels, and transfer rules are fixed before the outcomes are examined. The structured procedures of Chapter 7 also start from a scientific hierarchy, but use the observed p-values to select multiple nodes or layers while targeting structured FDRs. Here the narrower task is to move a fixed FWER budget through a prespecified graph.

Each hypothesis $H_i$ receives an initial local alpha budget $\alpha_i$, with

$$\sum_{i=1}^{m} \alpha_i \leq \alpha.$$

Directed edges carry transition weights $g_{ij} \geq 0$, where $g_{ii} = 0$ and

$$\sum_{j=1}^{m} g_{ij} \leq 1.$$

If $H_i$ is rejected, its current alpha budget is transferred to other hypotheses according to the outgoing weights. If $H_i$ is not rejected, its budget is not transferred.

Figure 4.5 is read in two stages: node labels give the local levels available now, and directed edges give the fractions transferred only after a rejection. The numerical update below shows why a transition weight is not itself a testing level.

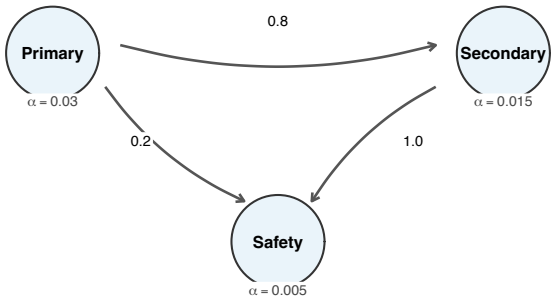



FIGURE 4.5. A graphical testing procedure. Initial alpha is assigned to nodes. After a rejection, the rejected node's local alpha is redistributed along outgoing edges. Only positive-weight edges are drawn; every omitted direction, including safety-to-secondary, has transition weight zero.

The general update after rejecting $H_i$ is:

$$\alpha_j \leftarrow \alpha_j + \alpha_i g_{ij}, \qquad j \neq i.$$

The rejected node is removed. The remaining transition weights are updated by

$$g_{jk} \leftarrow \begin{cases} \dfrac{g_{jk} + g_{ji} g_{ik}}{1 - g_{ji} g_{ij}}, & g_{ji} g_{ij} < 1, \\ 0, & g_{ji} g_{ij} = 1, \end{cases} \qquad j \neq k, \quad j, k \neq i,$$

and $g_{jj} = 0$. All quantities on the right refer to the graph before this removal. The boundary case $g_{ji} g_{ij} = 1$ means both weights are one: there is no route out of the pair to another node, so

the updated outgoing weights from $j$ are zero. This convention avoids an undefined division [32]. The random-walk explanation below derives the fraction.

The graph in Figure 4.5 gives a concrete three-endpoint calculation. Initially the primary, secondary, and safety hypotheses have local levels

$$(0.030,\ 0.015,\ 0.005).$$

If the primary endpoint rejects, its 0.030 budget is split according to the outgoing weights: 0.8 to secondary and 0.2 to safety. The remaining local levels become

$$\alpha_{\text{secondary}} = 0.015 + 0.8(0.030) = 0.039, \qquad \alpha_{\text{safety}} = 0.005 + 0.2(0.030) = 0.011.$$

The edge weights are updated separately from the local alpha levels. Write $P, S, A$ for primary, secondary, and safety. The nonzero initial transition weights are

$$g_{PS} = 0.8, \qquad g_{PA} = 0.2, \qquad g_{SA} = 1,$$

and these are the only positive-weight edges: every direction not drawn in the figure, including safety-to-secondary, has weight zero. After rejecting $P$, the remaining secondary-to-safety edge is

$$g_{SA}^{\text{new}} = \frac{g_{SA} + g_{SP}g_{PA}}{1 - g_{SP}g_{PS}} = \frac{1 + 0 \cdot 0.2}{1 - 0 \cdot 0.8} = 1.$$

The reverse safety-to-secondary edge remains zero:

$$g_{AS}^{\text{new}} = \frac{g_{AS} + g_{AP}g_{PS}}{1 - g_{AP}g_{PA}} = 0.$$

Thus the updated graph has a single edge from secondary to safety with weight one. If secondary then rejects, safety receives the secondary budget, giving

$$\alpha_{\text{safety}} = 0.011 + 0.039 = 0.050.$$

Thus, for p-values $(0.020, 0.030, 0.018)$, the procedure rejects all three: primary at 0.030, secondary at 0.039, and safety after recycling at 0.050. If the primary endpoint had not been rejected, its alpha would not have been transferred; secondary and safety would have remained at their initial 0.015 and 0.005 levels.

Two common special cases are fixed-sequence and fallback testing.

**Example 4.8: Fixed sequence**

Assume the testing plan lists the hypotheses in a priority order

$$H_1,\ H_2,\ \ldots,\ H_m,$$

before the data are examined. Fixed sequence puts the entire error budget on the first unrejected item in this list. Thus $H_1$ is tested at level $\alpha$; only if $H_1$ is rejected does $H_2$ get tested, again at level $\alpha$; and the same gatekeeping logic continues down the list. A single nonrejection closes the gate for all hypotheses that follow. As a graph, this is the chain with $\alpha_1 = \alpha$, $\alpha_i = 0$ for $i > 1$, $g_{i,i+1} = 1$ for $i < m$, and no other transitions. For instance, with $\alpha = 0.025$ and p-values in this hypothesis order

$$(0.004,\ 0.031,\ 0.006,\ 0.001),$$

only $H_1$ is rejected. The second p-value exceeds 0.025, so the procedure stops there; the smaller p-values later in the sequence are not eligible for confirmatory claims. The advantage is that every reached test uses the full level $\alpha$; the disadvantage is the complete dependence on the preassigned order.

**Example 4.9: Fallback**

Fallback testing also uses an ordered list, but it does not put all of the initial budget on the first hypothesis. Choose nonnegative local levels $\alpha_1, \dots, \alpha_m$ with $\sum_i \alpha_i \le \alpha$. Hypothesis $H_i$ starts with its own reserved amount $\alpha_i$, and if $H_i$ is rejected, its current level is transferred to $H_{i+1}$. The graph is again a chain, $g_{i,i+1} = 1$, with all other transition weights zero.
For example, take $\alpha = 0.05$, initial levels

$$(0.030,\ 0.015,\ 0.005),$$

and p-values

$$(0.040,\ 0.012,\ 0.018).$$

The first hypothesis is not rejected at level 0.030, so no alpha is passed forward from $H_1$. The second hypothesis is still tested at its reserved level 0.015, and it rejects. Its 0.015 is then added to the third hypothesis, so $H_3$ is tested at $0.005 + 0.015 = 0.020$ and also rejects. This is the essential distinction from fixed sequence: a nonrejection prevents that hypothesis's alpha from being transferred, but hypotheses with their own reserved alpha can still be tested.

**Route: Advanced**

The random-walk view explains both the update denominator and its connection to closed testing. The same graph does two jobs. First, its initial local levels define a weighted Bonferroni test: each component is tested against its own allocated fraction of $\alpha$, and the intersection rejects if any component passes. For the global intersection,

$$H_{\{1,\dots,m\}} = \bigcap_{i=1}^{m} H_i \quad \text{is rejected if} \quad p_i \le \alpha_i \text{ for some } i.$$

Under this global intersection, every component null is true, so super-uniformity and the union bound give

$$\mathbb{P}\left(\bigcup_{i=1}^{m} \{p_i \le \alpha_i\}\right) \le \sum_{i=1}^{m} \mathbb{P}(p_i \le \alpha_i) \le \sum_{i=1}^{m} \alpha_i \le \alpha.$$

This calculation does not require independence. Second, the directed edges specify the shortcut after an elementary rejection: remove the rejected node, recycle its current local level, and update the remaining graph so that it represents the same closed weighted-Bonferroni logic on the smaller family. Section 7 formalizes the local weights and the property that makes this sequential shortcut agree with closure.

Normalize the current local levels by the total level, $w_i = \alpha_i/\alpha$, and think of one unit of error budget as a walker that starts at node $i$ with probability $w_i$; any initially unused alpha starts outside the testing family. At a given stage, the hypotheses still under consideration are stopping states: alpha that reaches one of them stays there and becomes part of its current local level. Rejected hypotheses are transient states: alpha that reaches a rejected node follows the outgoing transition weights until it is absorbed by a remaining hypothesis. If an outgoing row sums to less than one, the missing probability can be represented by adding a lost-alpha state $H_{m+1}$ with

$$g_{j,m+1} = 1 - \sum_{k=1}^{m} g_{jk}, \qquad g_{m+1,k} = 0 \quad (k \le m), \qquad g_{m+1,m+1} = 1.$$

On this expanded graph, define a Markov chain by

$$\mathbb{P}(Z_0 = H_i) = \frac{\alpha_i}{\alpha}, \qquad i = 1, \dots, m,$$
$$\mathbb{P}(Z_0 = H_{m+1}) = 1 - \sum_{i=1}^{m} \frac{\alpha_i}{\alpha},$$
$$\mathbb{P}(Z_{t+1} = H_k \mid Z_t = H_j) = g_{jk}, \qquad j, k = 1, \dots, m+1.$$

For a remaining intersection $I \subseteq \{1, \dots, m\}$, let

$$\tau_I = \inf\{t \geq 0 : Z_t \in \{H_i : i \in I\} \cup \{H_{m+1}\}\}.$$

The weight assigned to $H_i$ in the local test for $H_I$ is its hitting probability

$$w_i(I) = \mathbb{P}(\tau_I < \infty,\ Z_{\tau_I} = H_i), \qquad i \in I,$$

so its current local level is

$$\alpha w_i(I) = \alpha\, \mathbb{P}(\tau_I < \infty,\ Z_{\tau_I} = H_i).$$

The lost-alpha state is absorbing and is not tested, so $\sum_{i \in I} w_i(I) \leq 1$. Strict inequality means that some original alpha can be absorbed by $H_{m+1}$ before reaching any hypothesis in $I$, or can remain forever in a closed class disjoint from those states. Thus any non-hitting probability is conservatively treated as lost alpha.

The denominator in the update formula is the two-node special case of this hitting-probability calculation. When the rejected node is $i$, a future unit of alpha leaving $j$ can reach $k$ directly through $j \to k$, through $j \to i \to k$, or after one or more round trips $j \to i \to j \to i$. The total rerouted weight is

$$(g_{jk} + g_{ji}g_{ik})\{1 + g_{ji}g_{ij} + (g_{ji}g_{ij})^2 + \cdots\} = \frac{g_{jk} + g_{ji}g_{ik}}{1 - g_{ji}g_{ij}}.$$

The geometric series requires $g_{ji}g_{ij} < 1$. In the boundary case $g_{ji} = g_{ij} = 1$, there is no escape from this pair to another node; the zero-outgoing-weight convention above records this fact. The update and its validity are derived by Bretz et al. [32].

The strong-FWER connection is now explicit. For every remaining intersection $I$, the hitting probabilities form weighted Bonferroni weights with total at most one, so the local test that rejects when $p_i \leq \alpha w_i(I)$ has level at most $\alpha$. The graph update preserves these intersection weights after each removal; applying the closure principle, Theorem 4.5, therefore yields strong FWER control, while the graph is the sequential shortcut to those closed tests.

## 7. Weighted Bonferroni and Consonance

**Route: Core**

We now state the two properties used in the graphical argument. Valid local tests give FWER control through closure. Monotone weights ensure that the sequential implementation reaches the same elementary decisions as those closed tests.

**Definition 4.10: Weighted Bonferroni local test**

For an intersection $H_I$, choose weights $w_i(I) \geq 0$ for $i \in I$ with

$$\sum_{i \in I} w_i(I) \leq 1.$$

The weighted Bonferroni local test rejects $H_I$ if

$$p_i \le \alpha w_i(I) \quad \text{for at least one } i \in I.$$

By the union bound, this is a level-$\alpha$ test for $H_I$. Closing these local tests gives strong FWER control by Theorem 4.5.

**Definition 4.11: Monotone weights**

The intersection weights are monotone if, whenever $J \subseteq I$ and $i \in J$,

$$w_i(J) \ge w_i(I).$$

Intuitively, removing rejected hypotheses should not reduce the alpha available to the hypotheses that remain.

**Definition 4.12: Consonance**

A family of local tests is consonant if, whenever it rejects an intersection $H_I$, there is some $i \in I$ such that every smaller intersection $H_J$, $i \in J \subseteq I$, is also rejected.

For weighted Bonferroni tests, monotone weights imply this property directly. If $p_i \le \alpha w_i(I)$ rejects $H_I$, then for every $i \in J \subseteq I$,

$$p_i \le \alpha w_i(I) \le \alpha w_i(J).$$

The same component therefore rejects each of those smaller intersections.

In the resulting closed procedure, rejection of an intersection therefore leads to the rejection of at least one elementary hypothesis. This is why a consonant closed procedure can be implemented as a sequential rejection rule, rather than ending with only a rejected intersection and no individual claim.

With monotone weights, the closed weighted Bonferroni procedure can be carried out by a sequential shortcut. The input of the shortcut is the whole collection of weight vectors $w(I) = (w_i(I))_{i \in I}$, one for each nonempty $I \subseteq \{1, \dots, m\}$, fixed before the data are examined. Start with $I = \{1, \dots, m\}$. At any stage, $I$ is the set of hypotheses not yet rejected. Test each remaining $H_i$, $i \in I$, at local level $\alpha w_i(I)$. If $p_i > \alpha w_i(I)$ for every $i \in I$, stop. Otherwise reject one $H_j$ with $p_j \le \alpha w_j(I)$; when several hypotheses qualify, any one of them may be chosen. If $j$ was the only index in $I$, every hypothesis has been rejected; stop. Otherwise replace $I$ by $I \setminus \{j\}$, the set $I$ with $j$ removed, and continue with the weights $w(I \setminus \{j\})$. Updating the weights after a rejection therefore means passing to the weight vector already assigned to the smaller intersection. No weight is computed from the p-values. The agreement with the closed procedure is shown below.

In practice the collection is generated by a rule rather than listed. Equal weights $w_i(I) = 1/|I|$ give closed Bonferroni, which is Holm's procedure by Proposition 4.6. The shortcut runs through Holm's thresholds: after each rejection $\alpha/|I|$ becomes $\alpha/(|I| - 1)$. Fixed positive base weights $v_1, \dots, v_m$, renormalized over the remaining hypotheses as $w_i(I) = v_i / \sum_{k \in I} v_k$, give a weighted version of Holm's procedure. In a graphical procedure, $\alpha w_i(I)$ is the local level of node $i$ after the nodes outside $I$ have been removed by the node and edge updates of Section 6. The result does not depend on the order of removal [32], so the graph fixes the whole collection before the data are examined. When $H_j$ is rejected, one more update, applied with $j$ as the rejected node,

turns the current levels $\alpha w_i(I)$ into $\alpha w_i(I \setminus \{j\})$. All three rules produce monotone weights. Under the first two, removing hypotheses shrinks the denominator and leaves the numerator unchanged. In the graph, each node update adds a nonnegative amount to the local level of every remaining node, and the levels for any $J \subseteq I$ are reached from those for $I$ by removing the indices in $I \setminus J$ one at a time.

The shortcut visits at most $m$ of the $2^m - 1$ intersections. Monotonicity is the property that makes it agree with the full closed-testing rule rather than merely resemble it. Suppose the shortcut rejects $H_i$, and let $J \subseteq \{1, \ldots, m\}$ be any index set containing $i$. Since $i \in J$, the shortcut rejects at least one hypothesis with index in $J$; let $H_r$ be the first, and let $I_r$ be the set $I$ at that stage, before $r$ is removed. Then $r \in I_r$, and the rejection step gives $p_r \leq \alpha w_r(I_r)$. No index in $J$ was removed earlier, so $J \subseteq I_r$, and monotonicity gives $w_r(I_r) \leq w_r(J)$. Hence

$$p_r \leq \alpha w_r(I_r) \leq \alpha w_r(J).$$

The local test therefore rejects $H_J$. Since $J$ was arbitrary, the closed procedure rejects $H_i$. Conversely, if the shortcut stops with a nonempty remaining set $I$, then the local test does not reject $H_I$, so the closed procedure rejects no $H_k$ with $k \in I$. If no hypothesis remains, there is nothing further to show. The two rejection sets therefore coincide. In particular, when several hypotheses cross their local thresholds at the same stage, the final rejection set does not depend on which of them is rejected.

Nonconsonance is easy to see with a combining test. Suppose $p_1$ and $p_2$ are independent valid p-values under their two null hypotheses. For the intersection $H_{12} = H_1 \cap H_2$, use Fisher's local test,

$$T_{12} = -2\{\log(p_1) + \log(p_2)\}, \qquad \text{reject } H_{12} \text{ if } T_{12} > \chi^2_{4,0.95}.$$

For the singleton intersections, use the ordinary level-0.05 tests $p_i \leq 0.05$. If $p_1 = p_2 = 0.07$, then

$$T_{12} = -4\log(0.07) \approx 10.64,$$

which exceeds $\chi^2_{4,0.95} \approx 9.49$. The intersection therefore rejects at level 0.05. However, both singleton tests fail because $0.07 > 0.05$. The rejected intersection does not point to any elementary hypothesis that can be rejected, which is exactly the nonconsonant behavior.

## 8. Simulation: Error and Power

**Route: Core**

Figure 4.6 separates a practical sparse setting from a dense stress test. In both settings, $m = 100$ one-sided Gaussian statistics have unit marginal variance, nonnull mean 3, and equicorrelation $\rho \in \{0, 0.3, 0.6, 0.9\}$; the target is $\alpha = 0.05$, and each point is based on 25,000 independent simulation runs. The sparse setting has 10 nonnulls and 90 nulls. Its empirical FWER panel should be read first: the dashed line marks the target that each procedure must respect. In the sparse panel of empirical power, Holm and Hochberg are nearly indistinguishable, a useful reminder that a theoretical ordering need not produce a practically important gain. The final panel uses a deliberately dense stress test with 95 nonnulls and five nulls. Hochberg is never below Holm there and separates clearly at the strongest correlation, especially $\rho = 0.9$; both methods exceed Bonferroni throughout the displayed correlation range.

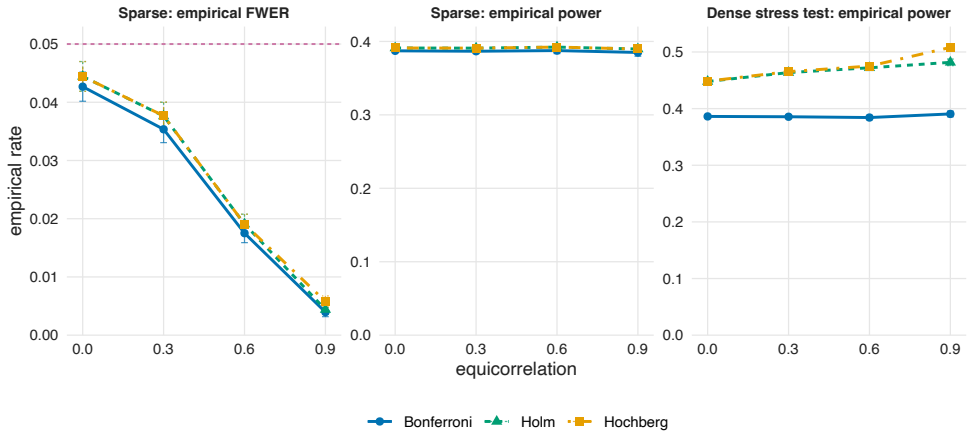



FIGURE 4.6. Empirical FWER and empirical power for Bonferroni, Holm, and Hochberg with $m = 100$, $\alpha = 0.05$, nonnull mean 3, and equicorrelations $\rho = 0, 0.3, 0.6, 0.9$. The first two panels use the sparse setting with 10 nonnulls and 90 nulls; the third is a dense stress test with 95 nonnulls and five nulls. Each point uses 25,000 independent simulation runs. Thin vertical bars are 95% simulation intervals, and the dashed line in the empirical-FWER panel marks 0.05. The common bottom legend applies to all three panels.

## 9. Assumptions in Plain Language

**Route: Core**

Bonferroni and Holm need only valid p-values for the true null hypotheses; they do not require independence. Sidak's exact threshold requires independence. Hochberg and Hommel get their extra power from Simes-type local tests, so their validity depends on conditions under which Simes is valid. Graphical procedures have strong FWER control only when the graph implements a valid weighted-Bonferroni test for every intersection hypothesis. The closure principle gives the resulting familywise guarantee.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Before seeing the p-values, specify the family, testing order, gatekeeping rules, weights, and graph. Verify every graph update and that each node's outgoing weights sum to at most one. |
| **What these data cannot show** | Observed p-values cannot show whether those choices were made after the results. They also cannot establish the independence and exact uniformity needed for Sidak's equality or the dependence conditions for Hochberg and Hommel. |
| **Safer fallback** | With valid p-values but uncertain dependence, use Holm. If prespecified priorities are essential, use the weighted-Bonferroni closure described above. |

## 10. Bibliographic Notes

**Route: Core**

Bonferroni and Sidak are classical. Holm's step-down procedure is due to Holm [84]. The closure principle is due to Marcus et al. [134]. Hochberg's step-up procedure was introduced by Hochberg [82], and the closed Simes shortcut leading to Hommel's procedure is due to Hommel [85]. Simes' test, used inside closed Simes procedures, is from Simes [164]. Modern graphical testing procedures and gatekeeping formulations are described by Bretz et al. [32] and Dmitrienko et al. [50].

Closure has also become a general tool for post hoc error statements beyond FWER. The exploratory-research formulation of Goeman and Solari [208] uses closed testing to give simultaneous true-discovery guarantees for sets chosen after seeing the data. Goeman et al. [209] show that closed testing is essentially unavoidable for admissible control of false discovery proportion (FDP) tail probabilities. More recent e-value work extends the same logic to online testing [207] and to broad classes of expected-loss error rates, including FDR [210].

The regulatory side of multiple testing in clinical trials is shaped by a small set of official guidance documents. The International Council for Harmonisation (ICH) E9 statistical-principles guideline [94] provides the statistical framework for primary and secondary endpoints, while its E9(R1) addendum [95] explains how estimands specify the corresponding targets of inference. The European Medicines Agency (EMA) guideline on multiplicity issues [59] and the U.S. Food and Drug Administration (FDA) guidance on multiple endpoints in clinical trials [179] specify how graphical and gatekeeping procedures should be planned and reported. The practical face of closure and graphical procedures is heavily shaped by these documents.

## 11. Exercises

**Route: Core**

### Basic.

**Exercise 4.13: Decision table**

For a testing problem with $m = 20$, $m_0 = 15$, $R = 6$, and $V = 2$, compute the entries $U$, $M$, and $D$ of the decision table in Figure 4.1, verify the identity $R = V + D$, compute the realized FDP, $\mathrm{FDP} = V/(R \vee 1)$, and determine whether a familywise error occurred.

**Exercise 4.14: A combined p-value for the nonconsonance example**

In the nonconsonance example of Section 7, $p_1 = p_2 = 0.07$ gives $T_{12} = -4\log(0.07)$. Compute the Fisher combined p-value $\mathbb{P}\{\chi^2_4 \geq T_{12}\}$.

### Intermediate.

**Exercise 4.15: Closure lattice**

For $m = 4$ hypotheses, draw the full closure lattice (15 nonempty intersections). For p-values

$$(0.004,\ 0.012,\ 0.030,\ 0.200),$$

mark which intersections are rejected by their local Bonferroni tests $\min_{i \in I} p_i \leq 0.05/|I|$. Deduce the elementary rejection set of closed Bonferroni, and identify the rejection set of the shortcut Holm procedure. Check that the two elementary rejection sets agree.

### Computational.

**Exercise 4.16: Simulation**

Reproduce Figure 4.6. Vary the effect size, correlation, and number of nonnull hypotheses, and summarize when the extra power of Hochberg is most visible. Report empirical FWER and empirical average power. For $m_1 > 0$, average power is estimated by the mean over runs of $D/m_1$, the fraction of nonnull hypotheses rejected; this is the power measure plotted in the figure. Assess FWER alongside power, since a method whose FWER exceeds $\alpha$ is not a valid comparison even if its power is larger.

**Exercise 4.17: Gatekeeping for a multi-endpoint trial**

Design a graphical procedure for a clinical trial with one primary endpoint, two co-secondary endpoints, and one prespecified subgroup endpoint. Specify the initial alpha budget and the transition weights. State the joint distribution used to simulate the four test statistics and define the power measure you report. Simulate FWER under the global null and under at least two partial-null configurations, and simulate power under one realistic alternative.

### Advanced.

**Exercise 4.18: Hommel and closed Simes**

For a small family, say $m \leq 8$, implement the closed Simes procedure directly by testing all intersection hypotheses. Compare its rejection set with Hommel's procedure from a standard software implementation, and verify that they agree on several p-value vectors. For validity under independent uniform null p-values, see Theorem 3.1; the PRDS conditions are discussed in Chapter 6.

CHAPTER 5

# False Discovery Rate and the BH Principle

FWER control is designed for settings where even one false rejection is costly. Many modern studies have a different objective. In a genome-wide screen, a model-comparison benchmark, or a feature-discovery pipeline, the analyst may be willing to tolerate some false discoveries if the overall list remains reliable. FDR is the standard error rate for that regime.

The central procedure in this chapter is the BH procedure. The main idea is simple: BH searches for a cutoff and rejects every p-value at or below it. It chooses the cutoff so that the estimated fraction of false discoveries among the rejected hypotheses is no larger than the target level.

## 1. FDP and FDR

**Route: Core**

Use the same decision table as in Chapter 4. Let $V$ be the number of false rejections and $R$ the total number of rejections. The FDP is

$$\mathrm{FDP} = \frac{V}{R \vee 1}.$$

The FDR is its expectation:

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

The convention $R \vee 1 = \max(R, 1)$ makes $\mathrm{FDP} = 0$ when no hypotheses are rejected.

FDR is weaker than FWER. Since $\mathrm{FDP} \leq \mathbf{1}\{V \geq 1\}$,

$$\mathrm{FDR} \leq \mathrm{FWER}\,.$$

Thus any FWER procedure controls FDR, but it may be unnecessarily conservative for exploratory discovery. Conversely, FDR control does not guarantee that the realized FDP in a particular study is below $q$. It controls the average of that random fraction across repeated studies under the assumed model. Nor does FDR control bound the probability that the realized FDP exceeds $q$. Procedures that target realized or post-hoc FDP statements are different objects: for example, simultaneous FDP confidence bounds give guarantees for many data-chosen rejection sets at once [72]. BH should therefore be read as an expected-error-rate guarantee, not as a claim that every selected list has small FDP.

## 2. The BH Step-Up Rule

**Route: Core**

Let $p_1, \ldots, p_m$ be p-values, with order statistics

$$p_{(1)} \leq \cdots \leq p_{(m)}.$$

Throughout this chapter, if a step-up search has no passing rank or admissible positive threshold, we set the selected rank or threshold to zero; this means that no hypotheses are rejected. The

BH procedure at level $q$ sets

$$\widehat{k} = \max\left\{ i \in \{1, \dots, m\} : \; p_{(i)} \le \frac{qi}{m} \right\}.$$

The procedure rejects all hypotheses with

$$p_i \le p_{(\widehat{k})}$$

when $\widehat{k} > 0$, and rejects nothing when $\widehat{k} = 0$.

Figure 5.1 turns the step-up definition into a visual search: compare each ordered p-value with the rising line $qi/m$, keep the largest rank that lies below the line, and reject the entire ordered prefix through that rank. The word "largest" matters because an early rank can miss its cutoff even when a later rank crosses it. In the displayed sample, $p_{(11)} = 0.01065 > 0.00917 = 11q/120$, whereas $p_{(17)} = 0.01224 < 0.01417 = 17q/120$. Thus rank 11 misses its own comparison, but rank 17 passes. Because no later rank passes, BH sets $\widehat{k} = 17$ and rejects the first 17 hypotheses, including the hypothesis at rank 11.

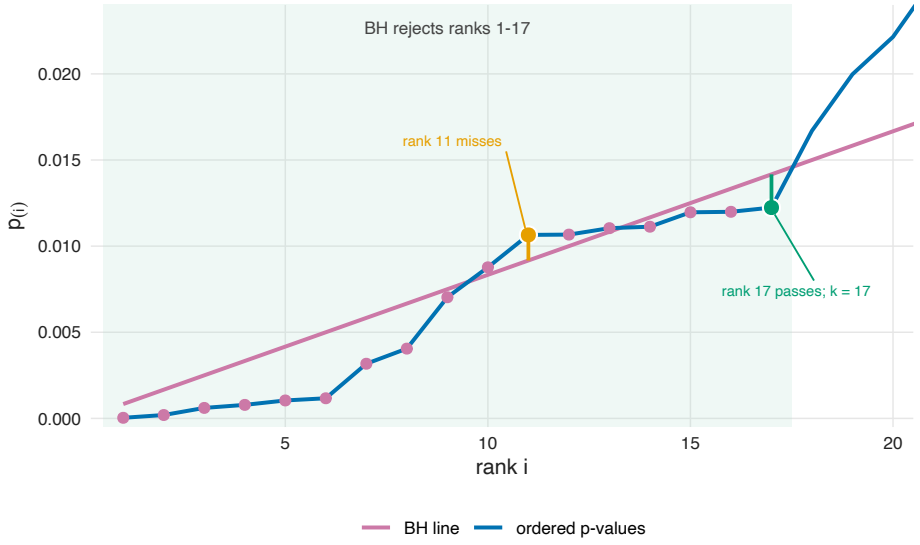



FIGURE 5.1. BH in ordered-p-value form. The procedure finds the largest rank at which $p_{(i)} \le qi/m$ and rejects all p-values up to that rank. The simulated example has $m = 120$, $q = 0.10$, and 16 one-sided Gaussian signals of mean 2.6. The vertical segments show the gaps between two ordered p-values and their own BH critical values: rank 11 misses because $0.01065 > 0.00917$, whereas rank 17 passes because $0.01224 < 0.01417$. No later rank passes, so $\widehat{k} = 17$ and BH rejects the entire prefix through rank 17, including rank 11. The display is a small-p-value zoom of the first 20 ranks.

For very small $m$, the difference between FWER and FDR procedures can be seen geometrically. Figure 5.2 colors the number of rejections made at each p-value configuration. Bonferroni and Holm are FWER procedures from Chapter 4; BH is the FDR step-up rule.

For $m = 2$, Bonferroni rejects each hypothesis with $p_i \le q/2$. Thus, in the Bonferroni panel, the blue arms are one-rejection regions: exactly one of $p_1, p_2$ is below $q/2$. Holm rejects one hypothesis if $p_{(1)} \le q/2$ and $p_{(2)} > q$, and rejects both if $p_{(1)} \le q/2$ and $p_{(2)} \le q$. BH rejects one hypothesis under the same one-rejection condition, but rejects both whenever $p_{(2)} \le q$. Thus

BH includes the square $q/2 < p_1, p_2 \le q$, where both p-values are moderately small but neither passes the Bonferroni threshold.

For $m = 3$, the figure shows the ordered slice with $p_{(3)} = 0.20 > q$, so at most two hypotheses can be rejected. Bonferroni rejects $\#\{i : p_i \le q/3\}$ hypotheses. Holm rejects none if $p_{(1)} > q/3$, one if $p_{(1)} \le q/3$ and $p_{(2)} > q/2$, and two if $p_{(1)} \le q/3$ and $p_{(2)} \le q/2$. BH rejects two whenever $p_{(2)} \le 2q/3$, and otherwise rejects one only if $p_{(1)} \le q/3$. This is the geometric face of the tradeoff: FWER procedures protect against any false rejection, while BH protects the expected FDP.

The same rule can be written as a threshold rule. For $0 \le t \le 1$, define

$$\mathcal{R}(t) = \{i : p_i \le t\}, \qquad R(t) = |\mathcal{R}(t)|.$$

If all null p-values were exactly uniform, then at threshold $t$ one would expect about $m_0 t$ false discoveries, where $m_0$ is the number of true nulls. Since $m_0 \le m$, $mt$ is a conservative estimate. BH chooses the canonical cutoff

$$\widehat{T}_{\mathrm{BH}} = \frac{q\widehat{k}}{m},$$

where $\widehat{k}$ is the step-up rank defined above. The procedure rejects $p_i \le \widehat{T}_{\mathrm{BH}}$ when $\widehat{k} > 0$, and rejects nothing when $\widehat{k} = 0$. Equivalently, when at least one p-value is rejected, $\widehat{T}_{\mathrm{BH}}$ is the largest positive threshold satisfying

$$R(t) > 0, \qquad \frac{mt}{R(t)} \le q.$$

The condition $R(t) > 0$ restricts the search to thresholds that actually make a rejection and keeps the ratio well defined. The canonical cutoff $\widehat{T}_{\mathrm{BH}}$ is not generally an observed p-value: when $\widehat{k} < m$, it sits in the open interval $(p_{(\widehat{k})}, p_{(\widehat{k}+1)})$ whenever $p_{(\widehat{k})} < q\widehat{k}/m < p_{(\widehat{k}+1)}$. What matters for the rejection decision is that, for $\widehat{k} > 0$, the set $\{i : p_i \le \widehat{T}_{\mathrm{BH}}\}$ coincides with $\{i : p_i \le p_{(\widehat{k})}\}$, so the rejection set is the ordered-p-value rejection set above.

There is a third view, useful for visualizing dependence corrections. The ECDF of the p-values is $F_m(t) = R(t)/m$, so the BH inequality $mt/R(t) \le q$ at thresholds with $R(t) > 0$ is the same as

$$F_m(t) \ge \frac{t}{q},$$

that is, the ECDF must lie above a line of slope $1/q$ through the origin. The rejection decision is still determined by the largest ordered rank $\widehat{k}$ satisfying $p_{(\widehat{k})} \le q\widehat{k}/m$. The horizontal ECDF segment at height $\widehat{k}/m$ meets the BH boundary $t/q$ at the canonical cutoff $\widehat{T}_{\mathrm{BH}} = q\widehat{k}/m$. This cutoff can lie to the right of the last rejected observed p-value $p_{(\widehat{k})}$, but no observed p-value lies between the two values, so they produce the same rejection set. For the Benjamini–Yekutieli (BY) procedure, defined fully in Section 10, the analogous canonical cutoff is $\widehat{T}_{\mathrm{BY}} = q\widehat{k}/(mH_m)$, where $H_m = \sum_{j=1}^{m} j^{-1}$ is the $m$th harmonic number. Replacing the line by a steeper boundary corresponds to a stricter procedure: this is how BY's harmonic correction enters in Figure 5.3 below.

## 3. Adjusted P-Values

**Route: Core**

BH-adjusted p-values report the smallest nominal BH level at which each hypothesis would be rejected by BH. To distinguish a hypothesis-specific adjusted value from the book-wide target $q$, write the former with an uppercase $Q$. In ordered form,

$$Q_{(i)} = \min_{j \ge i} \left\{ \frac{mp_{(j)}}{j} \right\} \wedge 1.$$

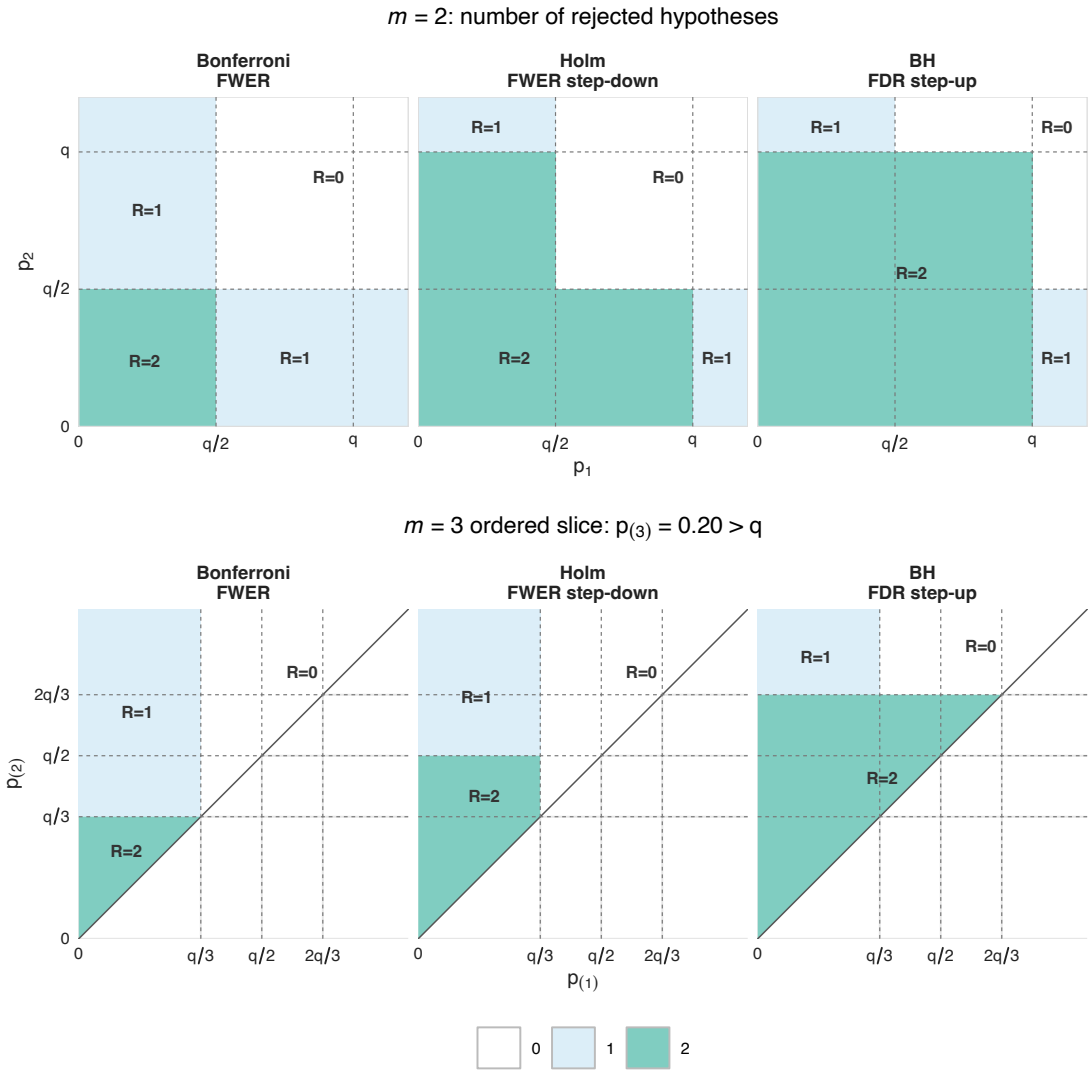



FIGURE 5.2. Rejection regions at $q = 0.10$ for $m = 2$ and $m = 3$. Color gives the number $R$ of rejected hypotheses: white is $R = 0$, blue is $R = 1$, and green is $R = 2$. Each panel recomputes those regions for its own method. The $m = 2$ row shows the full $(p_1, p_2)$ square near the origin. The $m = 3$ row shows the ordered slice $(p_{(1)}, p_{(2)})$ with $p_{(3)} = 0.20 > q$. Bonferroni rejects an individual hypothesis only when its p-value is at most the common cutoff $q/m$. Holm checks the ordered p-values sequentially against $q/m, q/(m-1), \ldots, q$, whereas BH uses $qi/m$. Thus BH's two-rejection region is visibly larger in the ordered $m = 3$ slice.

The running minimum over $j \geq i$ is essential: adjusted p-values must be monotone in the ordered ranks. Return the adjusted values to the original hypothesis order, assigning $Q_{(i)}$ to the hypothesis that produced $p_{(i)}$.

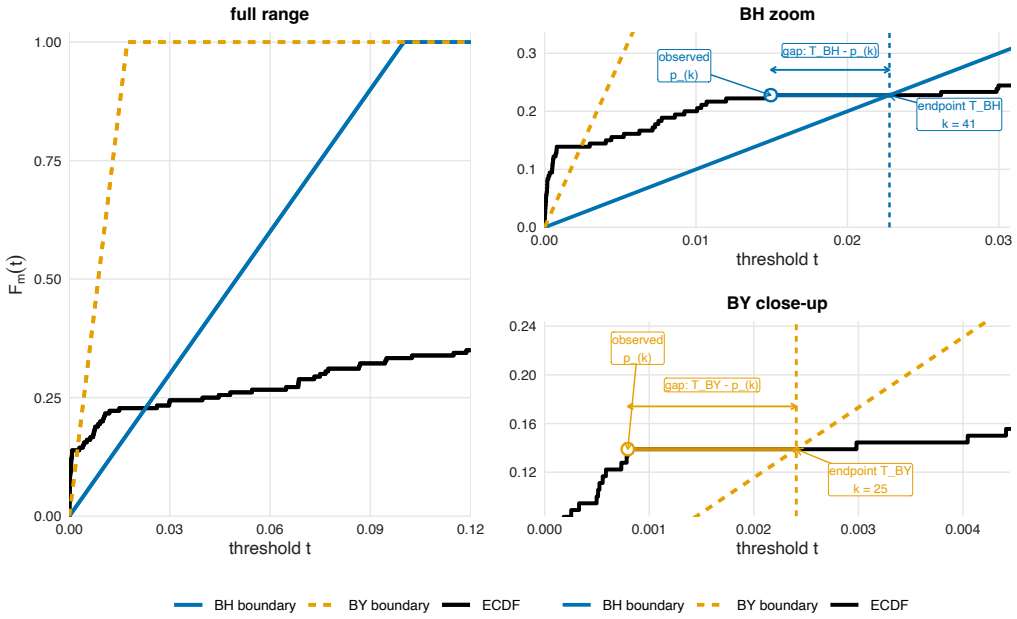



FIGURE 5.3. ECDF view of BH and BY. The ECDF is a step function. Open circles mark the observed last rejected p-values $p_{(\widehat{k})}$, while dashed vertical lines mark the canonical cutoffs $q\widehat{k}/m$ and $q\widehat{k}/(mH_m)$. The zoom panels explicitly mark the slack intervals $\widehat{T}_{\mathrm{BH}} - p_{(\widehat{k})}$ and $\widehat{T}_{\mathrm{BY}} - p_{(\widehat{k})}$: there are no observed p-values in such an interval, so the canonical cutoff and the last rejected p-value produce the same rejection set even though they are different numbers.

**Example 5.1**

Suppose $m = 5$ and the ordered p-values are

$$0.003, \quad 0.014, \quad 0.041, \quad 0.20, \quad 0.62.$$

The raw BH ratios $mp_{(i)}/i$ are

$$0.015, \quad 0.035, \quad 0.0683, \quad 0.25, \quad 0.62.$$

These are already increasing, so they are the BH-adjusted p-values in ordered rank. At level 0.05, BH rejects the first two hypotheses.

These BH-adjusted p-values describe the rejection rule; its FDR guarantee still requires the assumptions of the BH theorem below. The term *q-value* is also used for related quantities. To keep them distinct, we call the values here BH-adjusted p-values. Chapter 6 explains how Storey's estimated q-values replace $m$ in the adjustment by $\widehat{\pi}_0 m$, an estimate of the number of true nulls; Section 10 then introduces population q-values.

## 4. What the BH Estimate Means

**Route: Core**

BH uses

$$\widehat{\mathrm{FDP}}_{\mathrm{BH}}(t) = \frac{mt}{R(t) \vee 1}$$

as a conservative estimate of the FDP. It is not a guaranteed upper bound on the realized FDP for every threshold or every data set. Its role is more subtle: under independence, the random threshold chosen by BH interacts with null p-values in a way that controls the expectation of FDP.

Figure 5.4 illustrates why this distinction is necessary: along one realized path, the estimator $mt/R(t)$ can lie below the realized FDP at some thresholds. BH's guarantee concerns the expected FDP at its data-selected canonical cutoff $\widehat{T}_{\mathrm{BH}}$, not pointwise domination of the realized curve.

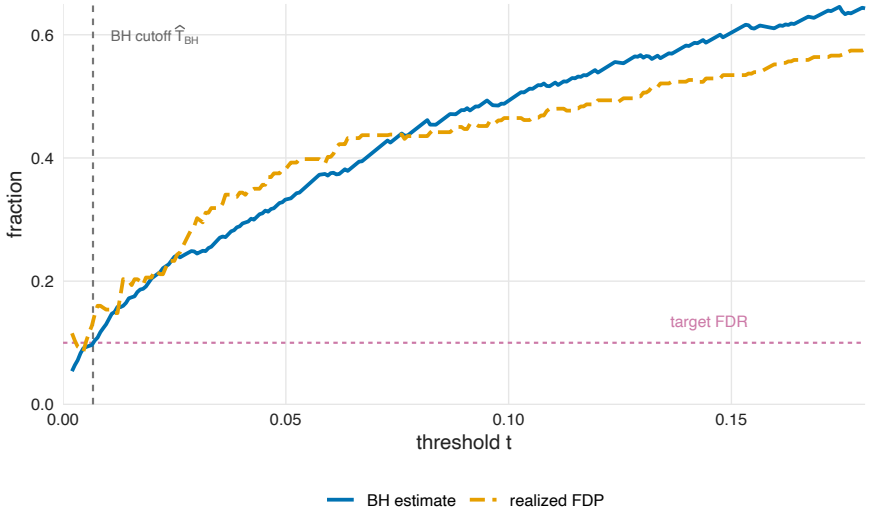



FIGURE 5.4. A single simulated path of the realized FDP and the BH estimator $mt/R(t)$. The estimator need not dominate the realized FDP pointwise; the BH theorem is an expectation statement at the selected cutoff. The gray vertical line marks the canonical BH cutoff $\widehat{T}_{\mathrm{BH}} = q\widehat{k}/m$. Here $m = 700$, $q = 0.10$, and 90 one-sided Gaussian signals have mean 2.45.

## 5. The Independent-Null Proof

**Route: Core**

The proof below is a template for much of modern FDR theory. Its two important features are self-consistency—every rejection satisfies the selected cutoff—and a leave-one-out argument that separates one true-null p-value from the random threshold.

**Theorem 5.2: BH under independent nulls**

Suppose the true-null p-values are independent uniforms and are independent of the false-null p-values. Then BH at level $q$ satisfies

$$\mathrm{FDR} \le \frac{m_0}{m} q \le q,$$

where $m_0$ is the number of true nulls.

**Proof of Theorem 5.2**

Let $\mathcal{I}_0$ be the set of true nulls. Write $R$ for the number of BH rejections. Then

$$\mathrm{FDR} = \sum_{i \in \mathcal{I}_0} \mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\right].$$

For a true null $i$, define $R_i$ as the number of rejections that BH would make if $p_i$ were set to zero while all other p-values were left fixed. Two observations make the leave-one-out argument go through. First, setting $p_i = 0$ forces the smallest p-value in the modified vector to be 0, which satisfies the smallest BH threshold $0 \le q/m$; the modified BH procedure therefore rejects at least one hypothesis, giving $R_i \ge 1$ deterministically. Second, on the event that $i$ is rejected by the original procedure, self-consistency of BH gives

$$p_i \le \frac{qR_i}{m} \qquad \text{and} \qquad R = R_i.$$

The identity $R = R_i$ is the key leave-one-out fact and deserves a line. If $i$ is rejected, then by BH self-consistency, $p_i \le qR/m$; lowering $p_i$ further to 0 only moves this already rejected p-value earlier among the first $R$ ordered positions. The $R$ originally rejected p-values still satisfy the rank-$R$ BH inequality. For any rank $s > R$, the $s$th ordered p-value after lowering $p_i$ is the same as the original $s$th ordered p-value, because only the already rejected $p_i$ moved within the first $R$ positions. Since $R$ was the largest crossing, no new rank $s > R$ can cross. Hence the rejection count is unchanged and $R_i = R$. Conversely, if $p_i \le qR_i/m$, replacing the zero in the modified vector by $p_i$ leaves at least $R_i$ p-values at or below $qR_i/m$. The original procedure therefore rejects at least $R_i$ hypotheses. Increasing a p-value cannot increase the BH rejection count, so it rejects exactly $R_i$, including $i$. Thus the two events are equal, and $R \vee 1 = R_i$ whenever either occurs. Consequently,

$$\frac{\mathbf{1}\{i \text{ rejected}\}}{R \vee 1} = \frac{\mathbf{1}\{p_i \le qR_i/m\}}{R_i}.$$

Conditional on $\{p_j : j \ne i\}$, the quantity $R_i$ is fixed and $p_i \sim \mathrm{Unif}(0,1)$. Therefore

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le qR_i/m\}}{R_i} \,\middle|\, \{p_j : j \ne i\}\right] = \frac{1}{R_i} \cdot \frac{qR_i}{m} = \frac{q}{m}.$$

Summing over the $m_0$ true nulls and taking expectations yields $m_0 q/m \le q$. □

**Corollary 5.3: BH under independent super-uniform nulls**

Suppose the true-null p-values are independent, super-uniform, and independent of the false-null p-values. Then BH at level $q$ satisfies

$$\mathrm{FDR} \le \frac{m_0}{m} q \le q.$$

**Proof of Corollary 5.3**

Repeat the leave-one-out proof of Theorem 5.2. Conditional on $\{p_j : j \neq i\}$, the leave-one-out count $R_i$ is fixed and $qR_i/m \in [0,1]$. Super-uniformity gives

$$\mathbb{P}\{p_i \le qR_i/m \mid \{p_j : j \neq i\}\} \le \frac{qR_i}{m},$$

so the contribution of each true null is at most $q/m$ instead of being exactly $q/m$. □

## 6. Z-Score Thresholds

**Route: Core**

Many large-scale testing problems begin with z-scores rather than p-values. For two-sided tests, let

$$R(t) = \sum_{i=1}^{m} \mathbf{1}\{|Z_i| \ge t\}.$$

Under a standard normal null, the expected number of null z-scores beyond $\pm t$ is at most $2m\{1-\Phi(t)\}$, so this is a conservative estimate of the false-discovery count. The clean way to state the procedure is to form the two-sided p-values

$$p_i = 2\{1-\Phi(|Z_i|)\}$$

and apply ordinary BH. With $\widehat{k}$ denoting the resulting step-up rank, its canonical p-value cutoff is

$$\widehat{T}_{\mathrm{BH}} = q\widehat{k}/m.$$

If $\widehat{k} > 0$, map this canonical p-value cutoff back to the z-score scale by setting

$$\widehat{T}_z = \Phi^{-1}\left(1 - \frac{\widehat{T}_{\mathrm{BH}}}{2}\right).$$

For every hypothesis,

$$|Z_i| \ge \widehat{T}_z \quad \Longleftrightarrow \quad 2\{1-\Phi(|Z_i|)\} \le \widehat{T}_{\mathrm{BH}} \quad \Longleftrightarrow \quad p_i \le \widehat{T}_{\mathrm{BH}}.$$

Under a true null, $Z_i \sim N(0,1)$, so $p_i = 2\{1-\Phi(|Z_i|)\}$ is uniform on $[0,1]$. Thus the z-score rule and p-value BH have exactly the same rejection set. If $\widehat{k} = 0$, no finite z-score cutoff is reported and the rejection set is empty.

Equivalently, when at least one BH rejection occurs, $\widehat{T}_z$ is the smallest z-score cutoff that makes at least one rejection while satisfying the z-scale FDP inequality:

$$\widehat{T}_z = \inf\left\{t \ge 0 :\ R(t) > 0,\ \frac{2m\{1-\Phi(t)\}}{R(t)} \le q\right\},$$

where the condition $R(t) > 0$ excludes cutoffs above all observed $|Z_i|$'s. The infimum is appropriate because larger $|Z_i|$ represents stronger evidence, so the rule selects the first cutoff at which the

inequality holds. At $t = \widehat{T}_z$, the numerator equals $q\widehat{k}$ and $R(t) = \widehat{k}$. If a smaller cutoff satisfied the inequality, it would correspond to a passing BH rank larger than $\widehat{k}$, contradicting the definition of $\widehat{k}$. Because this is only a monotone rewriting of the same BH decisions, the assumptions and bound transfer directly: if the true-null z-scores are independent $N(0,1)$ variables and are independent of the false-null z-scores, then Theorem 5.2 gives

$$\mathrm{FDR} \le \frac{m_0}{m} q \le q.$$

No separate FDR proof on the z-score scale is needed.

## 7. Weighted BH

**Route: Core**

External information can improve power. Some hypotheses may be more promising, more scientifically important, or measured with higher precision. A fixed weighted BH procedure assigns weights $w_i \ge 0$ satisfying

$$\frac{1}{m}\sum_{i=1}^{m} w_i = 1.$$

It applies BH to the transformed p-values

$$\frac{p_i}{w_i},$$

with the convention that hypotheses with $w_i = 0$ cannot be rejected. Larger weights make a hypothesis easier to reject; smaller weights make it harder. When $w_i > 1$, the transformed quantity $p_i/w_i$ is not itself a valid p-value under the null; the validity comes from the normalized weight budget. Section 9 proves this by the same leave-one-out argument as ordinary BH: a true null with weight $w_i$ contributes at most $qw_i/m$, and these charges sum to at most $q$ because $\sum_i w_i = m$.

Figure 5.5 shows the geometric effect of the weights. Dividing by a large $w_i$ moves a high-priority hypothesis downward in the ordered plot, so it may cross the BH boundary sooner; a small weight moves it upward. The normalization redistributes this advantage rather than enlarging the total weight budget.

Weights must be fixed independently of the null p-values they are used to test, or learned in a way justified by additional theory such as sample splitting, external covariates, or structural modeling. Reusing the same noise twice can destroy FDR control.

## 8. Learning Weights from Covariates

**Route: Advanced**

Modern multiple-testing problems often come with auxiliary information for each hypothesis: gene length, technical replicates, baseline expression, prior significance, distance to a known feature, or a side-data summary derived from an entirely different experiment. Let $X_i \in \mathcal{X}$ denote such side information for hypothesis $i$, and write $\mathbf{X} = (X_1, \dots, X_m)$. A useful covariate may predict power or the chance that $H_i$ is false, but it must not make a null p-value more likely to be small. A convenient sufficient formulation is conditional validity: for every true null $i$,

$$\mathbb{P}\{p_i \le t \mid \mathbf{X}\} \le t, \qquad 0 \le t \le 1.$$

For example, gene length can be suitable when it affects precision but does not change the null distribution of the p-value. A quantity computed from the same test statistic generally is not suitable unless its null independence is proved.

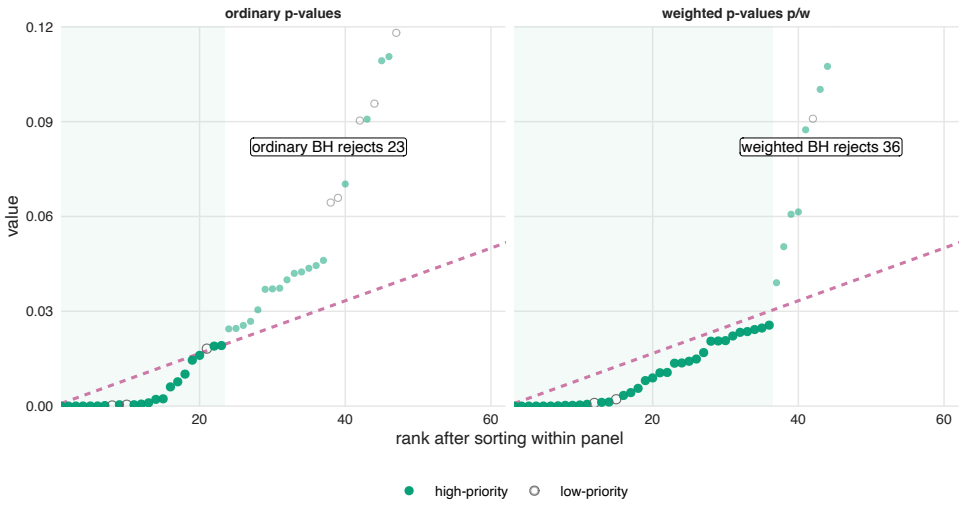



FIGURE 5.5. Weighted BH as BH on transformed p-values $p_i/w_i$. High-priority hypotheses receive larger weights, so their transformed p-values move downward relative to low-priority hypotheses; in this simulated example the weighted rule rejects visibly more hypotheses than ordinary BH. Here $m = 120$, $q = 0.10$, and the two equally sized priority groups receive weights 1.8 and 0.2, so the weights average to one.

Weighted BH first has a particularly clean guarantee when the weights come from prespecified information or an independent pilot sample. The following condition states exactly what the leave-one-out proof needs.

**Proposition 5.4: Independent-weight validity**

Let the possibly random, nonnegative weight vector $W = (w_1, \dots, w_m)$ be obtained from external covariates, prespecified information, or an independent pilot sample, and suppose $\sum_i w_i = m$ almost surely. Require that, for every true null $i$ and every $t \in [0, 1]$,

$$\mathbb{P}\{p_i \le t \,|\, W, \{p_j : j \ne i\}\} \le t \quad \text{almost surely.}$$

Then weighted BH applied to $p_i/w_i$ controls FDR at level $q$.

**Proof of Proposition 5.4**

For a true null $i$, condition on $\mathcal{G}_i = \sigma(W, \{p_j : j \ne i\})$. If $w_i = 0$, hypothesis $i$ cannot be rejected and its FDR contribution is zero. If $w_i > 0$, let $R_i$ be the rejection count after setting $p_i = 0$, with the same weights. Then $R_i \ge 1$. On the event that $i$ is rejected, the weighted-BH leave-one-out identity gives $R = R_i$ and

$$p_i \le \frac{q w_i R_i}{m}.$$

Therefore

$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\,\middle|\, \mathcal{G}_i\right] \le \mathbb{E}\left[\frac{\mathbf{1}\{p_i \le q w_i R_i/m\}}{R_i}\,\middle|\, \mathcal{G}_i\right] \le \frac{q w_i}{m}.$$

For the last inequality, conditional super-uniformity applies when $qw_iR_i/m \le 1$; when this threshold exceeds one, the bound follows from $1/R_i \le qw_i/m$. Summing over true nulls and then averaging gives

$$\mathrm{FDR} \le \mathbb{E}\left[\sum_{i\in\mathcal{I}_0} \frac{qw_i}{m}\right] \le q,$$

because $\sum_i w_i = m$ almost surely. □

Proposition 5.4 is practically useful when weights are fixed before the current p-values are seen—for example, protocol priorities, prespecified external annotations, or weights learned from an independent pilot and then frozen. Its displayed condition must still hold. One simple sufficient case is that, given $W$, every true-null $p_i$ is super-uniform and independent of the other p-values. The proposition therefore gives an elementary guarantee for external or frozen weighting and marks its limit: independence of $p_i$ from its own covariate alone is insufficient under dependence.

Independent hypothesis weighting (IHW) asks whether the current covariates and p-values can help learn the weights. The cross-weighting framework of Ignatiadis et al. [90] and Ignatiadis and Huber [89] combines learning from other folds (subsets of hypotheses), a separate weight budget in each fold, and—for a finite-sample guarantee—censoring. This construction is not, in general, an application of Proposition 5.4: cross-weighting stops $p_i$ from affecting its own-fold weights, but it may affect weights learned for other folds. Thus the full weight vector can reveal information about $p_i$ and violate conditional validity. Its finite-sample guarantee instead comes from the separate censored-IHW theorem below.

Fix a censoring level $\tau \in (0, 1]$. A schematic version of the formal construction is:

(1) Partition the hypotheses into $K$ folds $F_1, \dots, F_K$ without using their p-values or covariates, and require the collections of hypotheses in different folds to be independent.

(2) For each fold $k$, use hypotheses outside $F_k$ to learn a nonnegative weight function $\widehat{w}_{-k}(x)$. The learner sees the training p-values only through $p_j\mathbf{1}\{p_j > \tau\}$, so all values at or below $\tau$ look the same. IHW commonly learns regularized, piecewise-constant weights over covariate bins.

(3) For $i \in F_k$, set $\widetilde{w}_i = \widehat{w}_{-k}(X_i)$ and normalize within the held-out fold:

$$w_i = \begin{cases} \dfrac{|F_k|\widetilde{w}_i}{\sum_{j\in F_k}\widetilde{w}_j}, & \sum_{j\in F_k}\widetilde{w}_j > 0, \\ 1, & \sum_{j\in F_k}\widetilde{w}_j = 0. \end{cases}$$

Thus $\sum_{i\in F_k} w_i = |F_k|$ for every fold and consequently $\sum_i w_i = m$.

(4) Define

$$\widehat{k} = \max\left\{k \in \{1,\dots,m\} : \sum_{i=1}^{m} \mathbf{1}\left\{w_i > 0,\ p_i \le \left(\frac{qw_ik}{m}\right) \wedge \tau\right\} \ge k\right\}.$$

Set $\widehat{k} = 0$ if the set is empty and make no rejections. When $\widehat{k} > 0$, reject every $i$ satisfying

$$w_i > 0, \qquad p_i \le \left(\frac{qw_i\widehat{k}}{m}\right) \wedge \tau.$$

The within-fold normalization is concrete. Suppose $F_1 = \{1, 2\}$ receives raw weights $(2, 1)$ from the rule trained outside that fold, while $F_2 = \{3, 4\}$ receives raw weights $(1, 1)$. The

normalized weights are

$$(w_1, w_2, w_3, w_4) = \left(\frac{4}{3}, \frac{2}{3}, 1, 1\right).$$

Each fold has weight sum two and the full vector has weight sum four. Changing $p_1$ or $p_2$ leaves both weights in $F_1$ unchanged because neither p-value was used to train or normalize that fold's weights.

Keeping a p-value out of the learner for its own fold is called *honest weighting*. Foldwise normalization prevents a global denominator from carrying cross-fold dependence back into that fold. Censoring and capping handle the remaining issue: p-values at or below $\tau$ look identical to the learner, whereas p-values above $\tau$ cannot be rejected by the capped rule.

Under mutually independent null pairs $(p_i, X_i)$, independence of those pairs from the nonnull pairs, super-uniform null p-values, and null independence between $p_i$ and $X_i$, the finite-sample IHW theorem of Ignatiadis and Huber [89] shows that honest, $\tau$-censored weights used with the capped rule above control FDR at level $q$. Without censoring, cross-weighted BH has an asymptotic guarantee under additional regularity conditions.

More adaptive covariate-threshold methods require additional ideas. In particular, procedures such as adaptive p-value thresholding (AdaPT) use models for the probability that a hypothesis is null to propose thresholds depending on covariates. They hide part of each p-value during the search so that adaptation does not use all the evidence later used for testing. Those ingredients are introduced after local false discovery rate (local FDR) in Chapter 6.

## 9. A Generalized BH Template

**Route: Advanced**

A broad class of BH-type rules can be put in a single framework. Rather than invert a transformed p-value, specify its rejection cutoff directly. Let $\mathcal{T} \subset (0, 1]$ be a nonempty deterministic finite grid. For each hypothesis, let $a_i(t) \in [0, 1]$ be a deterministic, nondecreasing cutoff on this grid. A zero cutoff means that the hypothesis is ineligible for rejection, even if its observed p-value is zero. Define

$$\mathcal{R}(t) = \{i : a_i(t) > 0,\ p_i \le a_i(t)\}, \qquad R(t) = |\mathcal{R}(t)|.$$

Let $g(t) > 0$ be a deterministic boundary function on $\mathcal{T}$. The generalized BH threshold is

$$\widehat{T} = \max\left\{t \in \mathcal{T} :\ R(t) > 0, \quad \frac{mg(t)}{R(t)} \le q\right\}.$$

If the set is empty, set $\widehat{T} = 0$ and make no rejections. Otherwise reject the hypotheses in $\mathcal{R}(\widehat{T})$. The finite grid ensures that every nonzero selected threshold is attained; no inverse outside its domain or limiting endpoint is needed.

**Proposition 5.5: Generalized BH bound**

Assume the true-null p-values are independent, super-uniform, and independent of the false-null p-values. For the cutoffs and boundary just specified, define

$$C = \frac{1}{m} \sum_{i \in \mathcal{I}_0} \max_{t \in \mathcal{T}} \frac{a_i(t)}{g(t)}.$$

Then the generalized BH procedure satisfies

$$\mathrm{FDR} \le Cq.$$

**Proof of Proposition 5.5**

For a true null $i$, set $p_i = 0$ and let $R_i$ and $\widehat{T}_i$ be the resulting rejection count and threshold. First suppose the original procedure rejects $i$ at $\widehat{T}$. Because its cutoffs are nondecreasing, $i$ is already counted in $R(t)$ at every grid point $t \geq \widehat{T}$. Setting $p_i = 0$ therefore changes none of those counts. The largest passing grid point stays $\widehat{T}_i = \widehat{T}$, and $R_i = R$.
Condition on all other p-values, so $R_i$ and $\widehat{T}_i$ are fixed. If $R_i = 0$, the preceding argument shows that $i$ cannot be rejected, and its FDR contribution is zero. If $R_i > 0$, the selected grid point satisfies

$$\frac{mg(\widehat{T}_i)}{R_i} \leq q.$$

The original rejection event is contained in $\{p_i \leq a_i(\widehat{T}_i)\}$, with denominator $R_i$ on that event. Super-uniformity and independence therefore give

$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1} \,\middle|\, \{p_j : j \neq i\}\right] \leq \frac{a_i(\widehat{T}_i)}{R_i} \leq \frac{q}{m} \cdot \frac{a_i(\widehat{T}_i)}{g(\widehat{T}_i)} \leq \frac{q}{m} \max_{t \in \mathcal{T}} \frac{a_i(t)}{g(t)}.$$

Summing over $i \in \mathcal{I}_0$ and taking expectations gives the result. □

For ordinary BH, take $\mathcal{T} = \{qk/m : k = 1, \ldots, m\}$, $a_i(t) = t$, and $g(t) = t$. At $t = qk/m$, the grid condition is $R(t) \geq k$, exactly the rank-$k$ BH crossing condition. Hence this choice gives ordinary BH and

$$C = \frac{m_0}{m}.$$

For fixed weighted BH, use the same grid and boundary, with $a_i(t) = \min(w_i t, 1)$. This cap keeps every cutoff in $[0, 1]$, and the zero-cutoff convention excludes zero-weight hypotheses. The rule is BH on $p_i/w_i$, with that ratio set to infinity when $w_i = 0$. Since $a_i(t) \leq w_i t$, normalized weights satisfy

$$C \leq \frac{1}{m} \sum_{i \in \mathcal{I}_0} w_i \leq \frac{1}{m} \sum_{i=1}^{m} w_i = 1.$$

Thus fixed weighted BH controls FDR at level $q$. Informative external weights can improve power by assigning more of the weight budget to nonnull hypotheses, although poorly chosen weights can reduce power.

The value of this template is conceptual. It shows that BH belongs to a larger family of procedures that choose a cutoff by comparing how many hypotheses would be rejected with a chosen boundary $mg(t)$. The constant $C$ measures how large the true-null rejection probabilities are relative to that boundary; $C \leq 1$ is what yields control at the target $q$. Later chapters use the same logic for dependence, adaptive estimation, and e-value procedures.

## 10. Arbitrary Dependence: The BY Correction

**Route: Core**

BH is valid under independence and under PRDS, defined in Section 1 of Chapter 6. Under arbitrary dependence the BH procedure can fail. Where does the failure come from? The leave-one-out proof of Theorem 5.2 used the unconditional uniform distribution of a true-null $p_i$ to compute the expectation $\mathbb{E}[\mathbf{1}\{p_i \leq qR_i/m\}/R_i] = q/m$; under arbitrary dependence, the random threshold $R_i$ is correlated with $p_i$, and the same expectation can be larger. A careful

arbitrary dependence calculation by Benjamini and Yekutieli [21] shows that FDR $\le qH_m$ under arbitrary dependence, where

$$H_m = \sum_{j=1}^{m} \frac{1}{j}$$

is the harmonic number. Dividing the nominal level by $H_m$ restores the $q$ bound. The BY procedure rejects according to

$$p_{(i)} \le \frac{qi}{mH_m},$$

i.e., BY is BH run at level $q/H_m$. Since $H_m = \log m + O(1)$, the power cost can be substantial, but the benefit is robustness: no dependence structure among the p-values is required.

**Theorem 5.6: BY correction**

Assume every true-null p-value is marginally super-uniform, $\mathbb{P}(p_i \le t) \le t$ for $0 \le t \le 1$, with arbitrary dependence allowed among all p-values. The BY procedure at level $q$, equivalently BH run at level $q/H_m$, satisfies

$$\text{FDR} \le \frac{m_0}{m} q \le q.$$

For the proof, first consider a generic step-up rule with nondecreasing critical values

$$0 = a_0 \le a_1 \le \cdots \le a_m \le 1.$$

We reserve $\eta$ for the nominal level supplied to an ordinary BH rule, so its generic cutoffs become $a_s = \eta s/m$; the symbol $q$ remains the book-wide FDR target to be imposed at the end.

**Proof of Theorem 5.6**

Let $R$ be the number of rejections and, for a true null $i$, write $F_i(t) = \mathbb{P}(p_i \le t)$, so that $F_i(t) \le t$. Step-up self-consistency says that if hypothesis $i$ is rejected and $R = r$, then $p_i \le a_r$. Hence

$$\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1} \le \frac{\mathbf{1}\{p_i \le a_R\}}{R \vee 1}.$$

On the event $p_i \in (a_{s-1}, a_s]$, the inequality $p_i \le a_R$ can hold only if $R \ge s$. Therefore, path by path,

$$\frac{\mathbf{1}\{p_i \le a_R\}}{R \vee 1} \le \sum_{s=1}^{m} \frac{1}{s} \mathbf{1}\{a_{s-1} < p_i \le a_s\}.$$

Super-uniformity gives $F_i(a_0) = F_i(0) = 0$. Taking expectations and summing by parts gives

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\right] &\le \sum_{s=1}^{m} \frac{F_i(a_s) - F_i(a_{s-1})}{s} \\
&= \frac{F_i(a_m)}{m} + \sum_{s=1}^{m-1} F_i(a_s)\left(\frac{1}{s} - \frac{1}{s+1}\right) \\
&\le \frac{a_m}{m} + \sum_{s=1}^{m-1} a_s \left(\frac{1}{s} - \frac{1}{s+1}\right).
\end{aligned}$$

For BH run at nominal level $\eta$, $a_s = \eta s/m$, so the last line equals

$$\frac{\eta}{m} + \frac{\eta}{m}\sum_{s=1}^{m-1}\frac{1}{s+1} = \frac{\eta}{m}\sum_{s=1}^{m}\frac{1}{s} = \frac{\eta H_m}{m}.$$

Summing over the $m_0$ true nulls shows that BH at nominal level $\eta$ has

$$\mathrm{FDR} \le \frac{m_0}{m}\eta H_m.$$

Taking $\eta = q/H_m$ proves the BY bound without redefining the target $q$. □

**Remark 5.7: Reshaping the step-up boundary**

The proof shows more than the harmonic correction. For any step-up rule with critical values $a_s = q\beta(s)/m$, where $\beta(0) = 0 \le \beta(1) \le \cdots \le \beta(m)$, the same argument gives

$$\mathrm{FDR} \le \frac{m_0 q}{m}\sum_{s=1}^{m}\frac{\beta(s) - \beta(s-1)}{s}.$$

Thus arbitrary-dependence control follows whenever the displayed sum is at most one. Ordinary BH uses $\beta(s) = s$, giving the factor $H_m$; BY uses $\beta(s) = s/H_m$. This is the simplest instance of the reshaping viewpoint for self-consistent FDR procedures: the rejection boundary is bent downward just enough that the worst-case dependence calculation still sums to one. More general reshaped and self-consistent procedures are developed in Blanchard and Roquain [28].

## 11. Mirror Estimation

**Route: Research frontier**

Another way to estimate false discoveries uses symmetry. Suppose null p-values are uniform, while alternatives mostly produce small p-values. Then the upper tail near one can act as a mirror for the lower null tail. For each p-value, write

$$M_i = \min(p_i, 1 - p_i), \qquad S_i = \mathbf{1}\{p_i \le 1/2\} - \mathbf{1}\{p_i > 1/2\}.$$

The reflected magnitude $M_i$ is the distance from $p_i$ to the nearer endpoint of $[0, 1]$, so a lower-tail value and its upper-tail reflection have the same magnitude. The sign $S_i$ records the side: $+1$ for the lower half and $-1$ for the upper half. The convention at $p_i = 1/2$ is irrelevant under a continuous null distribution. For $0 < t \le 1/2$, define

$$\widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) = \frac{1 + \#\{i : p_i \ge 1 - t\}}{R(t) \vee 1}.$$

The plus-one correction is essential for the finite-sample guarantee below: under the conditional sign-symmetry and stopping-rule conditions explained below, it gives the expected-count bound needed to prove $\mathrm{FDR} \le q$. Without the plus one, the analogous uncorrected mirror rule is less conservative but need not control ordinary FDR in finite samples. A mirror procedure chooses from the finite candidate set of positive reflected magnitudes

$$\mathcal{U}_{\mathrm{mir}} = \{M_i : 1 \le i \le m,\ M_i > 0\}.$$

Because every p-value lies in $[0, 1]$, each $M_i$ is automatically at most $1/2$; an explicit upper-bound intersection is therefore redundant. Strict positivity merely removes the degenerate zero-threshold

candidate arising at $p_i \in \{0, 1\}$. The procedure then sets

$$\widehat{T}_{\mathrm{mir}} = \max\left\{t \in \mathcal{U}_{\mathrm{mir}} : \widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) \le q\right\}.$$

If no candidate passes, set $\widehat{T}_{\mathrm{mir}} = 0$ and make no rejections. Otherwise reject $p_i \le \widehat{T}_{\mathrm{mir}}$.

Figure 5.6 uses nominal level $q = 0.10$. The canonical BH cutoff is $\widehat{T}_{\mathrm{BH}} = q\widehat{k}/m$. At this cutoff, $R(\widehat{T}_{\mathrm{BH}}) = \widehat{k}$ and the conservative estimate is $m\widehat{T}_{\mathrm{BH}}/\widehat{k} = q$; the example has $\widehat{k} = 53$. The mirror rule instead uses the observed upper tail. Because that count is small at the selected threshold, $(1 + \#\{p_i \ge 1 - t\})/(R(t) \vee 1)$ remains at most 0.10 for a larger candidate and the rule rejects 84 hypotheses.

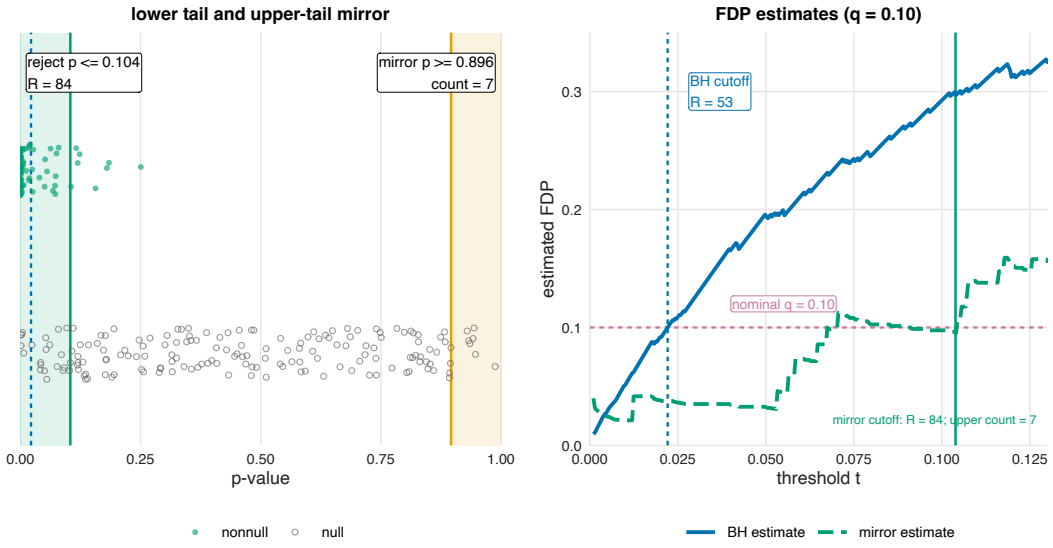



FIGURE 5.6. Mirror estimation in one simulated example at nominal level $q = 0.10$. The lower tail supplies candidate discoveries, while the upper tail $p_i \ge 1 - t$ estimates the null contribution in the lower tail. Compared with BH's conservative estimate $mt/R(t)$, the mirror estimate can allow a larger threshold when the upper tail is sparse; here BH rejects 53 hypotheses and the mirror rule rejects 84. The simulation has $m = 240$ hypotheses, including 70 one-sided Gaussian signals of mean 2.7.

This rule is attractive because it estimates the number of false discoveries from the data rather than using the conservative upper bound $mt$. One transparent sufficient condition is that, conditional on all reflected magnitudes $M_1, \ldots, M_m$ and the nonnull information, the true-null signs $S_i$ are mutually independent, with each sign equally likely to be $+1$ or $-1$. The rule can be implemented by scanning the candidate magnitudes from largest to smallest and stopping at the first passing candidate. At a current threshold $t$, the available information includes all magnitudes, the individual p-values already removed from the two tails, and the current and past aggregate lower- and upper-tail counts. These counts allow the FDP estimate to be computed without identifying which of the remaining tail p-values are small. This growing record of available information is called a *filtration*; a *stopping rule* decides whether to stop using only the information available at that stage. The aggregate counts can make the remaining signs dependent after conditioning, so the proof must account for them rather than assume that the hidden signs remain independent.

Under the stated initial conditional sign symmetry, the resulting stopped-count bound gives

$$\mathrm{FDR} \leq q.$$

The argument uses a supermartingale, a process whose conditional mean at the next step is at most its current value, to control the sign counts at the selected step. We defer the details to Barber and Candès [9]; the same sign-symmetry logic appears in the knockoff proofs of Chapter 9. Three caveats matter in practice: the orientation of the mirror count is essential (counting $p_i \geq 1 - t$ rather than $p_i \leq 1 - t$), the plus-one correction is needed for finite-sample control, and dependence among null p-values is not free — the sign-flip argument needs a suitable conditional joint-null symmetry, not merely marginal symmetry. Exact uniformity and even PRDS do not by themselves validate this adaptive two-tail comparison; Section 2 of Chapter 6 gives an exact counterexample.

## 12. Assumptions in Plain Language

**Route: Core**

The equicorrelated simulation in Figure 5.7 previews a systematic failure rather than a high-correlation curiosity. Section 2 of Chapter 6 shows that the mirror threshold can fail for exactly uniform PRDS p-values and for Gaussian scores with any fixed positive equicorrelation, even though PRDS continues to protect BH.

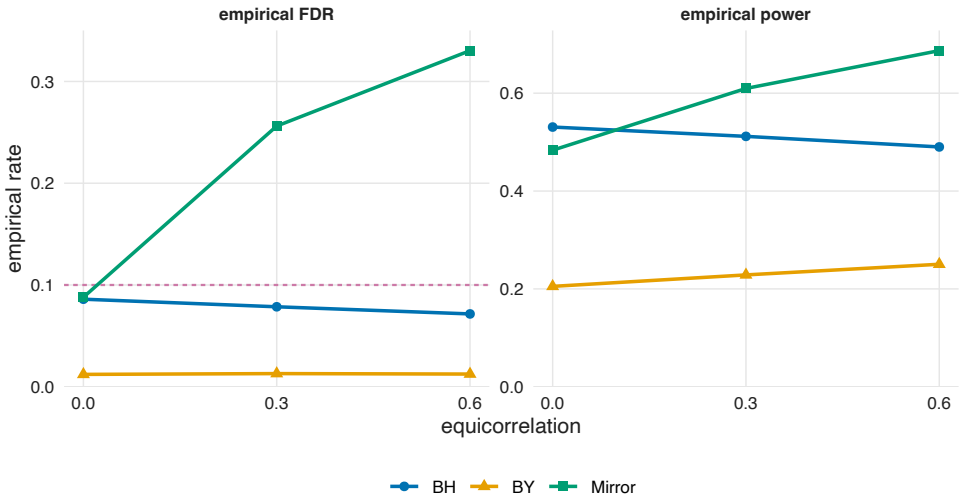



FIGURE 5.7. Simulation comparing BH, BY, and a mirror estimator. The panels show empirical FDR and empirical power. BY is robust but conservative. The mirror rule can gain power, but its validity is more sensitive to dependence and symmetry assumptions. The design uses $m = 300$ one-sided Gaussian tests, 45 signals of mean 2.45, $q = 0.10$, 1,000 replications, and equicorrelations 0, 0.3, and 0.6.

BH needs valid null p-values and a dependence condition that lets the random threshold be analyzed. Independence is the cleanest case; PRDS is treated in Chapter 6. BY sacrifices power to avoid dependence assumptions. Weighted BH is valid when the weights are external to the tested null p-values. Mirror estimators use symmetry and require more care than the formula alone suggests.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Review how each p-value was constructed and confirm that the tested family was fixed before results were examined. For fixed weighted BH, verify that the weights were chosen independently of the testing p-values and average to one. For IHW, also verify the independent fold construction, within-fold normalization, and stated censoring rule. |
| **What these data cannot show** | A single p-value vector cannot establish independence or PRDS, identify which rejections are false, or guarantee that the realized FDP is below $q$. |
| **Sensitivity checks** | For weighted BH, also report ordinary BH. When BH is used under a debatable dependence model, compare it with BY at the same target $q$. These comparisons show how the reported rejection set depends on the method; they do not verify its assumptions. |
| **Safer fallback** | Use BY when no defensible independence or PRDS argument is available. If the weighting construction lacks the conditions required by its theorem, omit the learned weights and use an unweighted procedure. |

## 13. Bibliographic Notes

**Route: Core**

The BH procedure was introduced by Benjamini and Hochberg [20]. The BY arbitrary-dependence correction is due to Benjamini and Yekutieli [21]. Generalized and reshaped FDR procedures are discussed by Blanchard and Roquain [28]. The mirror estimator and its sign-flip finite-sample argument are due to Barber and Candès [9], where the related knockoff filter is introduced. The distinction between a bare mirror or knockoff+ threshold and a valid knockoff filter under dependence is studied by Zhang [201]. The empirical-process, adaptive, and empirical-Bayes extensions are deferred to Chapter 6.

IHW was introduced by Ignatiadis et al. [90]; the cross-weighting framework and its finite-sample theory, including the role of $\tau$-censoring, were developed by Ignatiadis and Huber [89]. The procedure is widely used in genomics, where the side information is typically a measure of statistical precision such as gene length or baseline mean expression. Masking-based and local-FDR-guided covariate thresholding, including AdaPT, is deferred to Chapter 6, after local FDR has been introduced. Group and multilayer extensions to grouped or tree-structured families are from Barber and Ramdas [10] and Yekutieli [197]; these are developed in detail in Chapter 7. A frontier topic not pursued in this chapter is *differentially private* BH, which uses Laplace-noised log p-values to obtain FDR control under $(\varepsilon, \delta)$-differential privacy with a multiplicative inflation factor; see Dwork et al. [56] for a treatment in the privacy-utility literature.

## 14. Exercises

**Route: Core**

### Basic.

**Exercise 5.8: FDP versus FWER**

For a testing problem with $m = 100$, $m_0 = 70$, $V = 3$, and $R = 8$, compute the FDP and decide whether a familywise error occurred. Then, keeping $m = 100$ and $m_0 = 70$ fixed, give the smallest value of $R$ for which a procedure rejecting $V = 3$ true nulls would still achieve FDP $\leq 0.10$, and verify that this $R$ is feasible given how many nonnull hypotheses are available.

**Exercise 5.9: BH by hand**

For p-values

$$0.004,\ 0.011,\ 0.032,\ 0.036,\ 0.067,\ 0.21,\ 0.33,\ 0.46,\ 0.71,\ 0.88,$$

compute the BH rejection set at $q = 0.10$.

**Exercise 5.10: Adjusted p-values**

For the same p-values, compute the BH-adjusted p-values. Be careful to apply the running minimum from the largest rank back to the smallest rank.

### Intermediate.

**Exercise 5.11: BY correction**

Compute $H_m$ for $m = 10, 100, 1000$. Compare the BH and BY critical values and discuss the power cost of arbitrary-dependence robustness.

### Computational.

**Exercise 5.12: Pathwise FDP**

Reproduce Figure 5.4, plotting the realized FDP and the BH estimator across thresholds in one simulated data set.

**Exercise 5.13: Simulation study**

Simulate BH, BY, and the mirror procedure at $q = 0.10$, using one-sided p-values $p_i = 1 - \Phi(Z_i)$, where

$$Z_i = \mu_i + \sqrt{\rho}\, Z_0 + \sqrt{1-\rho}\, \varepsilon_i,$$

and $Z_0, \varepsilon_1, \ldots, \varepsilon_m$ are independent standard normal variables. Start with the design of Figure 5.7. Report empirical FDR, the distribution of realized FDP, and, when nonnulls are present, empirical average power as defined in Exercise 4.16. Use at least 2,000 independent runs per setting.

As a sanity check, under the all-null independent setting the empirical FDR should be close to the nominal level for BH, while BY should be visibly more conservative. Then set every hypothesis to null, fix $\rho \in \{0.05, 0.10\}$, and increase $m$ through $200, 1000, 5000$. Estimate $\mathbb{P}(R > 0)$, which equals FDR under the global null, and compare BH with the mirror rule. Use the exact mirror candidate set $\{\min(p_i, 1-p_i) : 1 \leq i \leq m, \min(p_i, 1-p_i) > 0\}$, and interpret the trend using Theorem 6.8. The theorem bounds a limit inferior; at these finite values of $m$, the mirror rejection probability need not be close to $1/2$.

**Advanced.**

**Exercise 5.14: Why the mirror correction is needed**

Consider the mirror rule of Section 11 at $q = 0.10$, with the same candidate set and rejection convention, but omit the plus-one term from its numerator. Using two independent uniform null p-values, show that this uncorrected rule fails to control FDR at level $q$.

**Exercise 5.15: BY within the generalized BH template**

Using the framework of Proposition 5.5, specify cutoffs $a_i(t)$, a deterministic finite grid $\mathcal{T}$, and a boundary $g(t)$ that recover the BY rejection rule at target $q$. The arbitrary-dependence guarantee is established separately in Theorem 5.6; the generalized template assumes independence.

CHAPTER 6

# Dependence, Adaptive FDR, and Empirical Bayes

Chapter 5 presented BH under independent null p-values and then introduced BY as an arbitrary-dependence correction. In large-scale inference, dependence often has a scientific explanation. Gene-expression measurements are correlated because genes participate in common pathways. Imaging statistics are dependent because nearby locations share features. Benchmark scores can share training data, prompts, and evaluation pipelines. Understanding this dependence can help us choose an appropriate procedure.

This chapter has two goals. First, it explains a positive-dependence condition under which BH still controls FDR. Second, it explains how estimating the fraction of true nulls or the probability that an individual case is null can improve a multiple-testing analysis. Storey's estimator, q-values, and local FDR address different parts of this problem. We will distinguish the questions they answer and the assumptions needed for their guarantees.

## 1. Positive Dependence and PRDS

**Route: Core**

The BH proof in Theorem 5.2 used a leave-one-out argument: a true-null p-value was separated from the random threshold. Under dependence, that separation is no longer automatic. PRDS supplies a different way to bound the same error probability.

Throughout this chapter, $P = (P_1, \dots, P_m)$ denotes the random vector of p-values, whereas $p = (p_1, \dots, p_m)$ denotes a possible value of that vector, such as the values observed in a data set. To describe how a procedure changes as the p-values change, write $p' \geq p$ when $p'_j \geq p_j$ for every coordinate $j$. A Borel-measurable set $A \subseteq [0,1]^m$ is *increasing* if

$$p \in A, \quad p' \geq p \quad \Longrightarrow \quad p' \in A.$$

Thus, once a vector belongs to $A$, increasing any of its coordinates keeps it in $A$. For example, $A = \{p : p_1 \geq a,\ p_2 \geq b\}$ is increasing. The set $A$ contains possible p-value vectors; $\{P \in A\}$ is the event that the random vector falls in that set. Measurability ensures that this event has a probability.

**Definition 6.1: PRDS for p-values**

A p-value vector $P = (P_1, \dots, P_m)$ satisfies *positive regression dependence on the subset* of true nulls $\mathcal{I}_0$ if, for every Borel-measurable increasing set $A \subseteq [0,1]^m$ and every $i \in \mathcal{I}_0$, a version of the conditional probability

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t)$$

can be chosen to be nondecreasing in $t$. Conditional probabilities given $P_i = t$ are interpreted through a regular conditional distribution.

The last sentence lets us use this definition with continuous p-values, for which $\mathbb{P}(P_i = t) = 0$ at each individual $t$. The condition compares the conditional distributions associated with different values of a null p-value. It says that a larger value of that coordinate cannot make any increasing-set event less likely. Although we condition on a true-null coordinate, the event can involve all coordinates, including nonnulls.

Figure 6.1 illustrates the set condition in two coordinates. Moving right or upward from a point in the shaded set keeps the point inside it. PRDS adds a probabilistic requirement: conditioning a null coordinate at a larger p-value makes membership in each such set at least as likely.

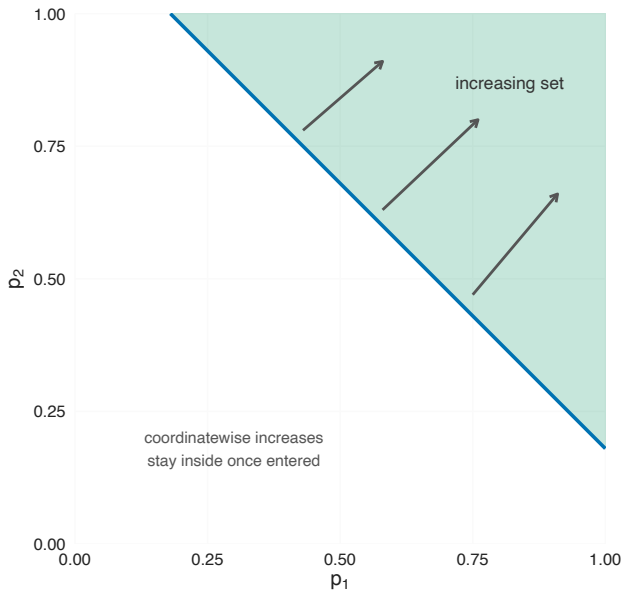



FIGURE 6.1. An increasing set in p-value coordinates. Once a point is inside the shaded set, increasing either coordinate keeps it inside. PRDS requires the conditional probability of every increasing-set event to be nondecreasing as the conditioning null p-value increases.

**Theorem 6.2: Gaussian sufficient condition for PRDS**

Let $Z = (Z_1, \ldots, Z_m)$ be multivariate normal with mean vector $\mu$ and covariance matrix $\Sigma$, with $\Sigma_{ii} = 1$ for every $i$. For testing $H_i : \mu_i = 0$ against $\mu_i > 0$, define

$$P_i = 1 - \Phi(Z_i),$$

where $\Phi$ is the standard normal distribution function. If

$$\mathrm{Cov}(Z_i, Z_j) \geq 0 \qquad \text{for every } i \in \mathcal{I}_0,\ j \in \{1, \ldots, m\},$$

then $P$ is PRDS on $\mathcal{I}_0$, and every true-null p-value is uniform on $(0, 1)$.

**Proof of Theorem 6.2**

For a true null, $Z_i \sim N(0,1)$, so $1-\Phi(Z_i)$ is uniform. To prove PRDS, fix $i \in \mathcal{I}_0$ and an increasing measurable set $A$ in p-value coordinates. The corresponding set in statistic coordinates is

$$B = \{z \in \mathbb{R}^m : (1-\Phi(z_1), \ldots, 1-\Phi(z_m)) \in A\}.$$

This set is *decreasing*. Indeed, if $z' \le z$, then $1-\Phi(z'_j) \ge 1-\Phi(z_j)$ for every $j$. Therefore $z \in B$ implies $z' \in B$, by the increasing-set property of $A$.

Write $Z_{-i}$ for the vector with its $i$th coordinate removed. The Gaussian conditional distribution of $Z_{-i}$ given $Z_i = z$ has mean

$$b(z) = \mu_{-i} + \Sigma_{-i,i}(z - \mu_i)$$

and covariance

$$C = \Sigma_{-i,-i} - \Sigma_{-i,i}\Sigma_{i,-i}.$$

Here we have used $\Sigma_{ii} = 1$. In particular, $C$ does not depend on $z$. Because every entry of $\Sigma_{-i,i}$ is nonnegative, $b(z_1) \le b(z_2)$ whenever $z_1 < z_2$.

We can compare the two conditional distributions using the same random vector $\varepsilon \sim N(0, C)$. For $r = 1, 2$, let $Y^{(r)}$ be the full vector whose $i$th coordinate is $z_r$ and whose other coordinates are $b(z_r) + \varepsilon$. Each $Y^{(r)}$ has the conditional law of $Z$ given $Z_i = z_r$. Moreover, $Y^{(1)} \le Y^{(2)}$ for every realization of $\varepsilon$. Since $B$ is decreasing,

$$\mathbf{1}\{Y^{(2)} \in B\} \le \mathbf{1}\{Y^{(1)} \in B\}.$$

Taking expectations gives

$$\begin{aligned}\mathbb{P}(P \in A \mid Z_i = z_2) &= \mathbb{E}[\mathbf{1}\{Y^{(2)} \in B\}] \\ &\le \mathbb{E}[\mathbf{1}\{Y^{(1)} \in B\}] \\ &= \mathbb{P}(P \in A \mid Z_i = z_1).\end{aligned}$$

Thus this conditional probability is nonincreasing as the conditioned statistic $z$ increases. Finally, conditioning on $P_i = t$, for $0 < t < 1$, corresponds to conditioning on $Z_i = \Phi^{-1}(1-t)$. This value of $Z_i$ decreases as $t$ increases. Substituting it into the preceding nonincreasing function therefore gives a nondecreasing function:

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t).$$

This is precisely PRDS in p-value coordinates. □

There are two order reversals in this argument. Increasing statistics make right-tail p-values smaller, so an increasing p-value set becomes a decreasing statistic set. Then increasing the conditioning p-value decreases the conditioning statistic. Keeping track of both steps explains why nonnegative Gaussian covariances give the required PRDS direction. For statistics with known positive variances other than one, first divide each statistic by its standard deviation before applying the theorem.

**Remark 6.3: MTP2, PRDS, and Gaussian signs**

Multivariate total positivity of order two (MTP2) imposes a stronger positive-dependence condition. A joint density $f$ is MTP2 if

$$f(x \vee y) f(x \wedge y) \ge f(x) f(y) \qquad \text{for all } x, y,$$

where $x \vee y$ and $x \wedge y$ take, respectively, the larger and smaller entry in each coordinate. MTP2 implies PRDS [102, 158].

For a nonsingular Gaussian distribution, MTP2 holds exactly when the *precision matrix*, the inverse covariance matrix $\Sigma^{-1}$, has nonpositive off-diagonal entries. This differs from the condition on $\Sigma$ in Theorem 6.2. For example,

$$\Sigma = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

is positive definite and has nonnegative covariances. The theorem therefore applies to its one-sided Gaussian p-values. But $(\Sigma^{-1})_{23} = 1/2 > 0$, so this Gaussian distribution is not MTP2. Nonnegative Gaussian covariances can thus establish PRDS without establishing MTP2. Also, the theorem only restricts pairs involving a true null; it does not impose a covariance sign on a pair of nonnull coordinates.

**Lemma 6.4: PRDS conditioning inequality**

Suppose $P$ is PRDS on $\mathcal{I}_0$. For any increasing measurable set $A \subseteq [0,1]^m$ and any $i \in \mathcal{I}_0$, the function

$$t \mapsto \mathbb{P}(P \in A \mid P_i \le t)$$

is nondecreasing over the thresholds for which $F_i(t) = \mathbb{P}(P_i \le t) > 0$.

**Proof of Lemma 6.4**

Choose the nondecreasing version $h(s) = \mathbb{P}(P \in A \mid P_i = s)$ from PRDS. For $F_i(t) > 0$, averaging these conditional probabilities over the lower tail gives

$$\mathbb{P}(P \in A \mid P_i \le t) = \frac{1}{F_i(t)} \int_{[0,t]} h(s)\, dF_i(s).$$

Take $t' > t$. If $F_i(t') = F_i(t)$, the two lower-tail events differ by a probability-zero event and the conditional probabilities are equal. Otherwise $F_i(t') > F_i(t) > 0$, and

$$\begin{aligned} \mathbb{P}(P \in A \mid P_i \le t') = {} & \frac{F_i(t)}{F_i(t')} \mathbb{P}(P \in A \mid P_i \le t) \\ & + \frac{F_i(t') - F_i(t)}{F_i(t')} \mathbb{P}(P \in A \mid t < P_i \le t'). \end{aligned}$$

Every argument of $h$ in $(t, t']$ is larger than every argument in $[0, t]$. Its average over the added interval is therefore at least its average over the original interval. The displayed weighted average cannot be smaller than $\mathbb{P}(P \in A \mid P_i \le t)$, proving the result. □

To connect this lemma to BH, let $R(p)$ be the number of rejections for a fixed p-value vector $p$, and let $R = R(P)$ be the random rejection count. If $p' \ge p$, then $p'_{(j)} \ge p_{(j)}$ at every ordered rank. Consequently, any rank satisfying $p'_{(j)} \le qj/m$ also satisfies $p_{(j)} \le qj/m$. The largest qualifying rank for $p'$ cannot exceed the largest qualifying rank for $p$, so $R(p') \le R(p)$. Thus

$$A_k = \{p : R(p) \le k\}$$

is an increasing set: making p-values larger cannot turn at most $k$ rejections into more than $k$ rejections. The event $\{R \le k\}$ is therefore an increasing-set event to which the lemma applies.

**Theorem 6.5: BH under PRDS**

If the p-value vector is PRDS on the set of true nulls and true-null p-values are super-uniform, then BH at level $q \in (0,1)$ satisfies

$$\mathrm{FDR} \le \frac{m_0}{m} q \le q.$$

**Proof of Theorem 6.5**

Let $R$ be the number of BH rejections. For each true null $i$, we will bound its expected contribution to the FDP by $q/m$:

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le qR/m\}}{R \vee 1}\right] \le \frac{q}{m}.$$

On the event $R > 0$, the numerator indicates that BH rejects hypothesis $i$. If $R = 0$, it is zero almost surely because super-uniformity implies $\mathbb{P}(P_i = 0) = 0$.
Write $t_k = qk/m$ and $F_i(t) = \mathbb{P}(P_i \le t)$. If $F_i(q) = 0$, hypothesis $i$ cannot be rejected, and the desired bound is immediate. Otherwise let $k_0$ be the smallest index with $F_i(t_{k_0}) > 0$. For $k < k_0$, the event $P_i \le t_k$ has probability zero. For every $k \ge k_0$, conditioning on that event is well defined, and super-uniformity gives $F_i(t_k)/k \le q/m$. Decomposing by $R$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le qR/m\}}{R \vee 1}\right] &= \sum_{k=k_0}^{m} \frac{1}{k} \mathbb{P}(P_i \le t_k, R = k) \\
&= \sum_{k=k_0}^{m} \frac{F_i(t_k)}{k} \mathbb{P}(R = k \mid P_i \le t_k) \\
&\le \frac{q}{m} \sum_{k=k_0}^{m} \mathbb{P}(R = k \mid P_i \le t_k).
\end{aligned}$$

It remains to show that the last sum is at most one. The conditioning event changes with $k$, so these terms are not probabilities from a single conditional distribution. This is the point at which we need PRDS.
Define $C_k(t) = \mathbb{P}(R \le k \mid P_i \le t)$ whenever $F_i(t) > 0$. The increasing-set argument above and Lemma 6.4 show that $C_k(t)$ is nondecreasing in $t$. Write $\mathbb{P}(R = k \mid P_i \le t_k) = C_k(t_k) - C_{k-1}(t_k)$ and regroup the sum:

$$\begin{aligned}
\sum_{k=k_0}^{m} \mathbb{P}(R = k \mid P_i \le t_k) &= C_m(t_m) - C_{k_0-1}(t_{k_0}) \\
&\quad + \sum_{k=k_0}^{m-1} \{C_k(t_k) - C_k(t_{k+1})\}.
\end{aligned}$$

The first term is one, the subtracted term is nonnegative, and each bracket is nonpositive because $t_k < t_{k+1}$. The sum is therefore at most one. When $k_0 = 1$, the subtracted term is in fact zero: $P_i \le q/m$ guarantees a BH rejection. Its nonnegativity suffices for the general case. We have bounded the contribution of each true null by $q/m$; summing over the $m_0$ true nulls proves the theorem. □

PRDS is a condition on the joint distribution, not merely on a list of pairwise correlations. The Gaussian theorem gives one useful way to verify it, but other dependence structures may

not satisfy it. When valid p-values have an otherwise unrestricted dependence structure, the BY correction from the previous chapter supplies an FDR guarantee.

**Remark 6.6: Two-sided Gaussian p-values**

For a Gaussian vector with unit marginal variances, the two-sided p-values are

$$P_i = 2\{1 - \Phi(|Z_i|)\}.$$

The map $Z_i \mapsto |Z_i|$ decreases on the negative half-line and increases on the positive half-line. Consequently, the coordinatewise comparison in Theorem 6.2 does not apply directly. This failure of the proof alone says nothing about whether BH is valid.

One question is whether BH controls FDR for *every configuration of null and nonnull means and every Gaussian correlation matrix*, when these two-sided p-values are used. This is the general two-sided Gaussian FDR conjecture. Dobriban [51] disproves it by constructing a sequence with $m = 100N$ tests, $96N$ true nulls, and $4N$ nonnulls. At target $q = 0.01$, its FDR exceeds 0.0104 for all sufficiently large $N$. A shared random factor creates realizations in which null statistics have unusually large absolute values. In those same realizations, the nonnull coordinates help increase the BH rejection count and hence the BH thresholds. The interaction leads to excess false discoveries.

A separate question restricts attention to the *global null*: $Z \sim N(0, \Sigma)$, with $\Sigma$ an arbitrary correlation matrix. Here $V = R$, so $\mathrm{FDR} = \mathbb{P}(R > 0)$. BH control is therefore equivalent to the two-sided Gaussian Simes inequality

$$\mathbb{P}\left(\exists k \in \{1, \dots, m\} : \; P_{(k)} \le \frac{kq}{m}\right) \le q.$$

This statement asks whether at least one ordered p-value can cross its Simes threshold with probability exceeding $q$ when every hypothesis is null. The cited counterexample uses nonnull means and therefore does not settle this global-null question. The two assertions must be kept separate.

## 2. Why PRDS Does Not Protect Mirror Thresholds

**Route: Research frontier**

The preceding theorem establishes a guarantee for BH. It does not establish the same guarantee for every method that chooses a threshold from the data. To see the distinction, recall the mirror procedure from Section 11 of Chapter 5. At a cutoff $0 < u \le 1/2$, define the two counts

$$R(u) = \#\{i : P_i \le u\}, \qquad L(u) = \#\{i : P_i \ge 1 - u\}.$$

The first count records candidate discoveries in the lower tail. The second records p-values in an upper interval of the same width. An exactly uniform null p-value falls in either interval with probability $u$, which motivates using the upper count to estimate the number of false discoveries in the lower count.

Let

$$D_i = \min(P_i, 1 - P_i), \qquad \mathcal{U} = \{D_i : 1 \le i \le m\} \cap (0, 1/2].$$

The distance $D_i$ tells us how close a p-value is to the nearer endpoint of $[0, 1]$, without telling us which endpoint that is. At cutoff $u$, the hypotheses contributing to either count are precisely those with $D_i \le u$. The mirror procedure chooses

$$\widehat{u} = \max\left\{u \in \mathcal{U} : \frac{1 + L(u)}{R(u) \vee 1} \le q\right\}$$

and rejects $P_i \le \widehat{u}$. If no candidate passes, it makes no rejections. The plus one is the finite-sample correction used in the mirror rule of the preceding chapter.

The equality of the two marginal null probabilities does not say how the p-values move together. If a shared source of randomness pushes many null p-values toward zero at once, $R(u)$ can be large while $L(u)$ is small. In particular, suppose every hypothesis is null, $m > 1/q$, and all p-values lie in $(0, 1/2)$. At the candidate $u = \max_i P_i$, all $m$ p-values enter the lower count and none enters the upper count. The ratio is $1/m < q$, so the procedure rejects. Under the global null this event gives FDP $= 1$, and the FDR is the probability of any rejection. The next example shows that such collective movement can occur with uniform PRDS p-values [201].

**Example 6.7: Uniform PRDS p-values with full support**

Suppose all $m$ hypotheses are null. Let $H$ be an unobserved random variable taking the values zero and one with equal probability. Given $H = h$, generate the p-values independently with density $f_h$, where $0 < \epsilon < 1/2$ and

$$f_0(u) = \begin{cases} 2(1-\epsilon), & 0 < u < 1/2, \\ 2\epsilon, & 1/2 < u < 1, \end{cases} \qquad f_1(u) = f_0(1-u).$$

When $H = 0$, a p-value lies in the lower half with probability $1 - \epsilon$; when $H = 1$, it does so with probability $\epsilon$. Thus $H$ changes the preferred half for all coordinates together. But averaging over $H$ gives $(f_0 + f_1)/2 = 1$, so each individual p-value is exactly uniform on $(0, 1)$. The joint density is

$$g(p_1, \dots, p_m) = \frac{1}{2} \prod_{j=1}^{m} f_0(p_j) + \frac{1}{2} \prod_{j=1}^{m} f_1(p_j).$$

Both terms are positive away from the irrelevant density boundary points. We may assign positive density values at those points as well. Hence every open region of the unit cube has positive probability. This is the meaning of *full support*; the example does not rely on forcing p-values to be equal. Both p-value ties and distance ties have probability zero.

We next verify PRDS. Fix an increasing measurable set $A$ and a coordinate $i$. Define

$$a_h(u) = \mathbb{P}\{(u, P_{-i}) \in A \mid H = h\}, \qquad w(u) = \mathbb{P}(H = 1 \mid P_i = u),$$

where $(u, P_{-i})$ means that $u$ is inserted in the $i$th position. For either $h$, increasing that fixed coordinate can only increase set membership, so $a_h(u)$ is nondecreasing.

The distribution with density $f_1$ is stochastically larger than the one with density $f_0$: it places more probability in the upper half. More explicitly, their distribution functions satisfy $F_1(v) \le F_0(v)$ for every $v$. On $[0, 1/2]$ their difference is $F_0(v) - F_1(v) = 2(1 - 2\epsilon)v$, and on $[1/2, 1]$ it is $2(1 - 2\epsilon)(1 - v)$, both nonnegative. Generating each coordinate under the two laws with the same uniform random number therefore makes every coordinate under $H = 1$ at least as large as its counterpart under $H = 0$. Since the coordinates are independent conditional on $H$, this comparison holds for the whole vector $P_{-i}$. It follows that $a_1(u) \ge a_0(u)$.

Bayes' rule gives the nondecreasing conditional probability

$$w(u) = \begin{cases} \epsilon, & u < 1/2, \\ 1/2, & u = 1/2, \\ 1 - \epsilon, & u > 1/2. \end{cases}$$

The value at $1/2$ is a permissible choice for conditioning at a point of probability zero. Conditioning first on $H$, we obtain

$$h_A(u) := \mathbb{P}(P \in A \mid P_i = u) = \{1 - w(u)\}a_0(u) + w(u)a_1(u).$$

For $u < v$, subtract the expressions and regroup:

$$\begin{aligned} h_A(v) - h_A(u) = {} & \{1 - w(v)\}\{a_0(v) - a_0(u)\} \\ & + w(v)\{a_1(v) - a_1(u)\} \\ & + \{w(v) - w(u)\}\{a_1(u) - a_0(u)\} \geq 0. \end{aligned}$$

Each factor in each product is nonnegative. This proves PRDS.
For the exact FDR calculation, take $q = 0.1$, $m = 11$, and $\epsilon = 0.1$. Let

$$S_i = \mathbf{1}\{P_i < 1/2\}, \qquad D_i = \min(P_i, 1 - P_i),$$

so $S_i = 1$ identifies a lower-half p-value. Given $H = h$, each $D_i$ is uniform on $(0, 1/2)$, independently of $S_i$, and

$$\mathbb{P}(S_i = 1 \mid H = 0) = 1 - \epsilon, \qquad \mathbb{P}(S_i = 1 \mid H = 1) = \epsilon.$$

For instance, on the lower half the joint density of $D_i = d, S_i = 1$ given $H = 0$ is $2(1 - \epsilon)$; this factors into the distance density 2 and the lower-half probability $1 - \epsilon$. The other cases factor in the same way.
Order the distances from smallest to largest as $D_{\pi(1)} < \cdots < D_{\pi(11)}$. Because this ordering uses only the distances, the indicators in that order remain independent given $H$, with the probabilities just displayed. At the candidate $u = D_{\pi(k)}$, exactly the first $k$ hypotheses enter the two tails, and

$$R(u) = \sum_{j=1}^{k} S_{\pi(j)}, \qquad L(u) = k - R(u).$$

The smallest possible mirror ratio for these $k$ hypotheses is $1/k$, achieved when all $k$ are in the lower tail. Consequently, no candidate with $k \leq 9$ passes. At $k = 10$, a candidate passes exactly when all ten are in the lower tail, since then the ratio is $1/10 = q$. At $k = 11$, even one upper-tail value gives a ratio of at least $2/10 > q$; thus all eleven must be in the lower tail. That event already implies that the first ten are in the lower tail. A rejection occurs exactly when the ten smallest-distance p-values all lie in the lower half.
Conditional on $H = 0$, this event has probability $(1 - \epsilon)^{10}$; conditional on $H = 1$, it has probability $\epsilon^{10}$. Hence

$$\text{FDR} = \frac{1}{2}\{(1 - \epsilon)^{10} + \epsilon^{10}\} = 0.1743392201 > 0.1.$$

If the threshold convention were $< q$ instead of $\leq q$, the ten-value candidate would fail at equality. All eleven would then have to lie in the lower half, giving

$$\text{FDR} = \frac{1}{2}\{(1 - \epsilon)^{11} + \epsilon^{11}\} = 0.15690529805.$$

Thus either convention violates the target level. Although each $S_i$ is individually a fair coin, for $i \neq j$ conditional independence gives

$$\text{Cov}(S_i, S_j) = \text{Var}\{\mathbb{E}[S_i \mid H]\} = (1/2 - \epsilon)^2 > 0.$$

The common variable $H$ makes many lower-half outcomes occur together. Uniform marginal p-values and PRDS do not prevent this dependence [201].

A similar failure occurs in a familiar Gaussian model. Here *equicorrelation* means that every pair of distinct Gaussian statistics has the same correlation. The correlation can be any fixed positive number; it need not approach one as more hypotheses are tested.

**Theorem 6.8: Fixed positive equicorrelation can defeat the mirror threshold**

Let $q \in (0,1)$, $\rho \in (0,1)$, and let $Z_0, Z_1, \dots$ be independent standard normal variables. For each $m$, apply the mirror procedure to the all-null p-values

$$P_i^{(m)} = 1 - \Phi\Big(\sqrt{\rho}\, Z_0 + \sqrt{1-\rho}\, Z_i\Big), \qquad i = 1, \dots, m.$$

These p-values are exactly uniform and PRDS, but their mirror procedure satisfies

$$\liminf_{m\to\infty} \mathrm{FDR}_m \geq \frac{1}{2}.$$

In particular, for $q < 1/2$, its FDR exceeds $q$ for all sufficiently large $m$ [201].

**Proof of Theorem 6.8**

Each statistic inside $\Phi$ has mean zero and variance one, and two distinct statistics have covariance $\rho$. Their p-values are therefore uniform and PRDS by Theorem 6.2.

To analyze the mirror counts, condition on the shared variable $Z_0 = z > 0$. Given this value, the statistics are independent with distribution $N(\mu_z, \sigma^2)$, where

$$\mu_z = \sqrt{\rho}\, z > 0, \qquad \sigma^2 = 1 - \rho.$$

Choose a positive statistic cutoff $a$ and its corresponding p-value cutoff $u = 1 - \Phi(a) < 1/2$. If $Y \sim N(\mu_z, \sigma^2)$, then the conditional lower- and upper-p-value tail probabilities are, respectively,

$$r_z(u) = \mathbb{P}(Y \geq a), \qquad l_z(u) = \mathbb{P}(Y \leq -a).$$

Let $f_z$ denote the density of $Y$. For $y > 0$, direct division of the two normal densities gives

$$\frac{f_z(-y)}{f_z(y)} = \exp\left(-\frac{2\mu_z y}{\sigma^2}\right).$$

Integrating over $y \geq a$ therefore yields

$$\begin{aligned} l_z(u) &= \int_a^\infty f_z(-y)\, dy \\ &\leq \exp\left(-\frac{2\mu_z a}{\sigma^2}\right) \int_a^\infty f_z(y)\, dy = \exp\left(-\frac{2\mu_z a}{\sigma^2}\right) r_z(u). \end{aligned}$$

Since $\mu_z > 0$, we can choose $a$ large enough that $l_z(u)/r_z(u) < q$. The probability $r_z(u)$ may be small, but is strictly positive.

For this fixed cutoff, the conditional law of large numbers gives

$$\frac{R_m(u)}{m} \longrightarrow r_z(u), \qquad \frac{L_m(u)}{m} \longrightarrow l_z(u)$$

almost surely as $m$ grows. Thus

$$\frac{1 + L_m(u)}{R_m(u) \vee 1} \longrightarrow \frac{l_z(u)}{r_z(u)} < q.$$

Eventually $R_m(u) > 0$ and the cutoff passes. Although this fixed $u$ need not be an observed distance, choose the largest observed $D_i \leq u$. It has the same two counts as $u$, so it is

a candidate that also passes. The actual procedure, which searches all candidates, must therefore reject. For every $z > 0$, its conditional probability of at least one rejection tends to one.

Under the global null, FDR is the probability of at least one rejection. Keeping only the contribution from $Z_0 > 0$, and using bounded convergence for these conditional probabilities, gives

$$\liminf_{m\to\infty} \mathrm{FDR}_m \geq \int_0^\infty 1 \cdot \varphi(z)\, dz = \frac{1}{2}.$$

Here $\varphi$ is the standard normal density. For $q < 1/2$, this lower bound on the limit inferior implies eventual FDR inflation. □

The result gives an eventual lower bound; it does not claim that FDR converges to exactly $1/2$. Its mechanism is the persistent common component $\sqrt{\rho}\, Z_0$. Even when $\rho$ is small, increasing $m$ does not average away this shared randomness. For example, the largest eigenvalue of the statistics' correlation matrix is $1 + (m-1)\rho$, which grows with $m$. Thus a small pairwise correlation need not mean that the collective effect of dependence is small.

Pairwise correlations can also miss other kinds of joint dependence. Zhang [201] constructs random scores with the same continuous distribution symmetric about zero, invariant joint law under permuting the coordinates, and zero correlation for each pair of distinct scores. Nevertheless, their mirror threshold has FDR arbitrarily close to one. Converting each score to a right-tail p-value using its common null distribution produces exactly uniform p-values with the same failure. Zero correlation of the original scores does not imply their independence, and it does not imply zero correlation of the transformed p-values. This result therefore warns against using pairwise score correlations alone to justify a procedure that depends on many coordinates together.

The same paper also limits a natural proposed repair. Suppose the mirror ratio is replaced by a deterministic function $g(r, l)$ of the two counts at the current cutoff, and the cutoff is accepted when $g(r, l) \leq q$. Assume $g$ cannot increase when the lower count $r$ increases, and cannot decrease when the upper count $l$ increases. These directions express the usual interpretation that more candidate discoveries help a cutoff pass, whereas more upper-tail controls make passing harder. For $q < 1/2$, within this class a rule that searches the current counts either fails FDR control for some full-support PRDS distribution or never rejects [201]. To see why, the most favorable possible counts are $r = m, l = 0$. If $g(m, 0) > q$, monotonicity implies $g(r, l) \geq g(m, 0) > q$ for every possible pair, so no candidate passes. If instead $g(m, 0) \leq q$, consider the preceding mixture example and let $\epsilon$ decrease toward zero. Whenever all $m$ p-values lie in the lower half, the candidate $u = \max_i P_i$ has counts $(m, 0)$ and passes. The probability of this event is

$$\frac{1}{2}\{(1-\epsilon)^m + \epsilon^m\} \longrightarrow \frac{1}{2}.$$

Because $q < 1/2$, sufficiently small positive $\epsilon$ makes the all-null rejection probability exceed $q$. Enlarging the plus one is one such change of $g$; it cannot provide a useful guarantee over that entire dependence class. The statement concerns this particular class of rules. It does not cover randomized rules, rules that violate the stated monotonicity, or rules that use information beyond the current pair of counts.

There is a useful connection to knockoffs, developed in Chapter 9. Writing a signed score as $1/2 - P_i$ turns the lower and upper p-value counts into positive and negative score counts, and the mirror formula becomes the knockoff+ threshold. A valid knockoff construction supplies additional conditional symmetry: given the score magnitudes and the nonnull scores, the null signs are independent fair coins. The counterexamples above do not have that property and

therefore do not contradict the knockoff guarantee. They show why the validity assumptions must accompany the threshold formula.

## 3. Empirical Processes and Reverse Martingales

**Route: Optional proof**

A threshold procedure chooses where to stop a search through the p-values. To study this choice, we keep track of how many p-values fall below each threshold. These counts, viewed together as the threshold varies, form an *empirical process*. For a p-value threshold $t$, write

$$V(t) = \#\{i \in \mathcal{I}_0 : P_i \le t\}, \qquad R(t) = \#\{i : P_i \le t\}.$$

The BH rejection set can be described by the largest $t \in [0,1]$ for which

$$\frac{mt}{R(t) \vee 1} \le q.$$

Assume for this section that the true-null p-values are independent and uniform on $[0,1]$. For each fixed $t$, each true null has probability $t$ of being counted, so $V(t) \sim \mathrm{Bin}(m_0, t)$ and $\mathbb{E}[V(t)] = m_0 t$. Independence alone would not give this identity without the uniform null distribution.

Imagine lowering $t$ from one toward zero. At the current threshold, we know the exact values of the null p-values above $t$, and we know which null p-values are still at or below $t$. We have not seen the exact values of those remaining p-values. A *filtration* is a mathematical description of the information available at each stage of this search. Here we use

$$\mathcal{F}_t = \sigma(\{\mathbf{1}(P_i \le u) : i \in \mathcal{I}_0,\ u \ge t\}),$$

where $\sigma(\cdot)$ means the information generated by the listed indicators. In particular, the indicators at $u = t$ already tell us the value of $V(t)$; this count is not extra information added to $\mathcal{F}_t$. The indicators at $u > t$ reveal the exact values above $t$. If $s < t$, then $\mathcal{F}_t \subseteq \mathcal{F}_s$: lowering the threshold reveals more information. This information is used in the proof, even though the analyst does not know which hypotheses are true nulls.

Given $\mathcal{F}_t$, the remaining null p-values are independent uniforms on $[0,t]$. Each has probability $s/t$ of remaining below a smaller threshold $s$. Consequently,

$$\mathbb{E}\left[\frac{V(s)}{s} \,\middle|\, \mathcal{F}_t\right] = \frac{V(t)}{t}, \qquad 0 < s < t \le 1,$$

The conditional expected future value of $V(s)/s$ equals its current value $V(t)/t$. This is the martingale property, with the threshold running backward, hence the name *reverse martingale* [170].

For a first reading, it is enough to check the same calculation with only $V(t)$ conditioned on. If $V(t) = v$, then the $v$ true-null p-values that fell in $[0,t]$ are conditionally iid uniform on $[0,t]$. Therefore, for $s < t$,

$$V(s) \mid V(t) = v \sim \mathrm{Bin}\left(v, \frac{s}{t}\right), \qquad \mathbb{E}\left[\frac{V(s)}{s} \,\middle|\, V(t) = v\right] = \frac{v}{t}.$$

For example, if $V(0.4) = 20$, the expected number remaining below 0.1 is $20(0.1/0.4) = 5$. The expected count has fallen, but after dividing by the threshold both ratios have value 50. The same calculation remains valid after revealing the exact null p-values above 0.4. We may also condition on all nonnull p-values if their vector is independent of the nulls.

To use this identity at a threshold chosen from the data, the stopping decision must use only information already revealed in the search. Such a decision is called a *stopping time*. Additional boundedness or integrability conditions then allow an optional-stopping theorem to carry the

conditional expectation identity over to that random threshold. A threshold chosen by looking at information still hidden in this description need not satisfy the same identity.

Storey's procedure in the next section uses a cutoff $\lambda \in (0,1)$ chosen before examining the p-values. It estimates the null fraction by counting p-values above $\lambda$, then searches for a rejection threshold at or below $\lambda$. For example, with $\lambda = 0.5$, it first uses the count above 0.5 to form the estimate. Keeping that estimate fixed, it starts at 0.5 and lowers the rejection threshold until its estimated FDP meets the target. The proof below explains why this order of using the data permits the reverse-martingale calculation.

## 4. Estimating the Null Fraction

**Route: Core**

BH uses $m$ as a conservative upper bound on the number of true nulls. If many hypotheses are nonnull, this can be conservative. Storey's procedure estimates the null fraction $\pi_0 = m_0/m$ from the right side of the p-value histogram. Choose a cutoff $\lambda \in (0,1)$ in advance. A uniform null p-value exceeds $\lambda$ with probability $1-\lambda$, so the expected number of true-null p-values above the cutoff is $m_0(1-\lambda)$. Dividing the observed count above $\lambda$ by $1-\lambda$ therefore suggests an estimate of $m_0$. Dividing once more by $m$ gives an estimate of the null fraction. The version used here adds one to the count:

$$\widehat{\pi}_0^{\lambda} = \frac{1 + \#\{i : P_i > \lambda\}}{(1-\lambda)m}.$$

Alternative hypotheses often produce small p-values, so the upper-tail count can be mostly due to true nulls. Alternatives above $\lambda$ increase the estimate, making the resulting procedure less powerful. The theorem below does not require that this tail contain only nulls.

The plus one matters because the estimate will appear in the denominator of the rejection rule. It keeps that denominator positive even when no p-value exceeds $\lambda$, and it supplies the reciprocal-moment bound in the proof. This estimate can exceed one: for example, with $m = 10$, $\lambda = 0.5$, and six p-values above 0.5, it equals $7/5 = 1.4$. We use the estimate as written, without truncating it at one. Although the true fraction is at most one, retaining a larger estimate makes this particular rejection rule more conservative.

Figure 6.2 shows the tuning tradeoff directly. At small $\lambda$, more alternatives can enter the right-tail count and push the estimate upward. Near one, fewer null p-values remain, so small changes in the count can cause large changes in the estimate. This is a tradeoff between upward bias and variability, rather than a reason to choose whichever cutoff gives the most discoveries.

Given $\widehat{\pi}_0^{\lambda}$, an adaptive BH threshold replaces $m$ by $m\widehat{\pi}_0^{\lambda}$:

$$\widehat{T} = \sup\left\{0 \le t \le \lambda : \frac{m\widehat{\pi}_0^{\lambda} t}{R(t) \vee 1} \le q\right\}.$$

Reject all hypotheses with $P_i \le \widehat{T}$. If none lie below this threshold, make no rejections. The search can be implemented by sorting the observed p-values and finding the largest rank $k$ such that

$$p_{(k)} \le \min\left\{\lambda, \frac{qk}{m\widehat{\pi}_0^{\lambda}}\right\}.$$

Reject the first $k$ hypotheses; if no rank qualifies, reject none. The numerical cutoff from this rank calculation need not equal the supremum $\widehat{T}$, but the two descriptions give the same rejection set.

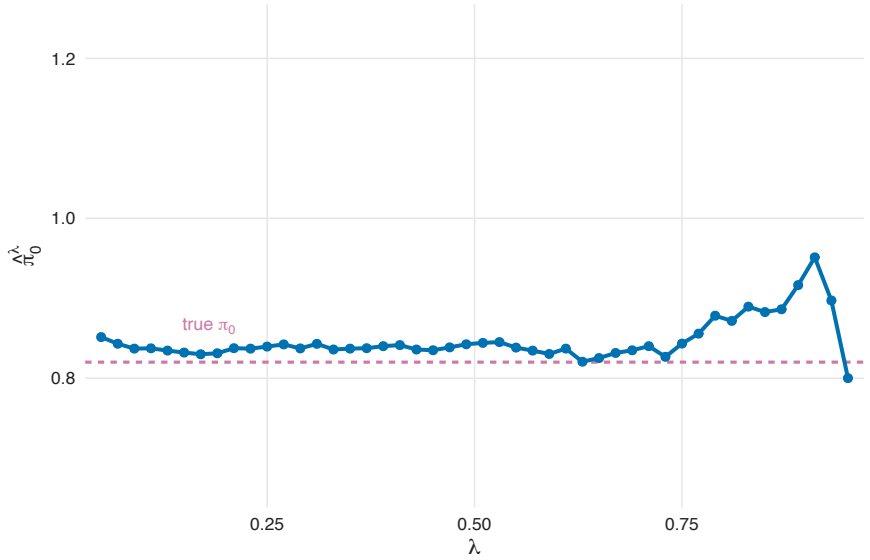



FIGURE 6.2. Storey's estimate $\widehat{\pi}_0^\lambda$ as $\lambda$ varies in one simulated mixture of $m = 2500$ tests with true null fraction $\pi_0 = 0.82$, null distribution $N(0,1)$, and alternative distribution $N(2.6,1)$. Small $\lambda$ can be biased by alternatives; large $\lambda$ can be noisy because few p-values remain in the tail.

The restriction $t \le \lambda$ keeps the search within the lower part of the p-value range after the upper-tail estimate has been fixed. Under the independence assumptions below, the number of true-null p-values contributing to the estimate satisfies

$$V_0(\lambda) = \#\{i \in \mathcal{I}_0 : P_i > \lambda\} \sim \mathrm{Bin}(m_0, 1-\lambda).$$

The proof bounds a ratio involving $1 + V_0(\lambda)$. It is this binomial calculation, rather than merely an upward bias of $\widehat{\pi}_0^\lambda$, that establishes the FDR guarantee.

**Theorem 6.9: Storey's finite-sample bound**

Assume the true-null p-values are independent uniforms and are independent of the vector of nonnull p-values. For fixed $q \in (0,1)$ and $\lambda \in (0,1)$, Storey's procedure above, with the untruncated $\widehat{\pi}_0^\lambda$, satisfies

$$\mathrm{FDR} \le q.$$

**Proof of Theorem 6.9**

Let $V(t) = \#\{i \in \mathcal{I}_0 : P_i \le t\}$. The value $\lambda$ is fixed in advance, and the threshold search is restricted to $0 < t \le \lambda$. Conditional on the nonnull p-values and on which true-null p-values lie above $\lambda$, the remaining true-null p-values in $[0,\lambda]$ are iid uniforms on $[0,\lambda]$. Call this conditioning the *tail information at* $\lambda$. After this information is fixed, $\widehat{\pi}_0^\lambda$ is fixed, and the only remaining randomness relevant to $V(t)$ comes from the unobserved null p-values below $\lambda$.

In the reverse filtration that reveals the lower tail as $t$ decreases, $V(t)/t$ is therefore a reverse martingale. The event $\{\widehat{T} \ge t\}$ can be decided from the p-values at thresholds $u \in [t,\lambda]$,

together with the fixed tail information at $\lambda$, so $\widehat{T}$ is a stopping time for this downward search. The supremum is attained: between successive observed p-values the ratio in the threshold rule is linear and increasing, and at a p-value its denominator can only increase. Also,

$$m\widehat{\pi}_0^{\lambda} \le \frac{m+1}{1-\lambda}, \qquad \widehat{T} \ge a := \min\left\{\lambda, \frac{q(1-\lambda)}{m+1}\right\} > 0.$$

Indeed, the estimated FDP at $a$ is at most $m\widehat{\pi}_0^{\lambda} a \le q$, even if $R(a) = 0$. Thus the process is stopped in the interval $[a, \lambda]$, where $V(t)/t \le m_0/a$. This deterministic bound justifies optional stopping without taking a limit at zero. Applying it conditionally on the tail information gives

$$\mathbb{E}\left[\frac{V(\widehat{T})}{\widehat{T}}\,\middle|\, \text{tail information at } \lambda\right] = \frac{V(\lambda)}{\lambda}.$$

Thus

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left[\frac{V(\widehat{T})}{R(\widehat{T}) \vee 1}\right] \\ &\le \frac{q}{m}\mathbb{E}\left[\frac{V(\widehat{T})}{\widehat{\pi}_0^{\lambda}\widehat{T}}\right] \\ &= \frac{q}{m}\mathbb{E}\left[\frac{m(1-\lambda)}{1+\#\{j : P_j > \lambda\}}\frac{V(\lambda)}{\lambda}\right].\end{aligned}$$

Since the denominator includes all p-values above $\lambda$,

$$1 + \#\{j : P_j > \lambda\} \ge 1 + m_0 - V(\lambda).$$

Here $m_0 - V(\lambda)$ is exactly the number of true-null p-values above $\lambda$. With $X = V(\lambda) \sim \mathrm{Bin}(m_0, \lambda)$, a direct calculation gives

$$\mathbb{E}\left[\frac{X}{1+m_0-X}\right] = \frac{\lambda(1-\lambda^{m_0})}{1-\lambda}.$$

Indeed, if $X \sim \mathrm{Bin}(n, p)$ and $Y = n - X \sim \mathrm{Bin}(n, 1-p)$, then

$$\frac{X}{1+n-X} = \frac{n-Y}{1+Y} = \frac{n+1}{1+Y} - 1,$$

and

$$\mathbb{E}\left[\frac{1}{1+Y}\right] = \sum_{y=0}^{n} \frac{1}{1+y}\binom{n}{y}(1-p)^y p^{n-y} = \frac{1-p^{n+1}}{(n+1)(1-p)}.$$

Substituting $n = m_0$, $p = \lambda$, and $X = V(\lambda)$ gives the displayed reciprocal moment. Substitution yields

$$\mathrm{FDR} \le q(1-\lambda^{m_0}) \le q.$$

□

Storey's estimate is not automatically valid under arbitrary dependence or arbitrary data-dependent choices of $\lambda$. The PRDS theorem for ordinary BH does not automatically transfer to Storey's adaptive denominator: dependence can make the right-tail estimate too small exactly on samples where the lower tail contains many null p-values. Positive dependence can therefore lead to FDR inflation for adaptive Storey procedures unless the dependence and adaptivity assumptions are controlled. In particular, searching over many values of $\lambda$ and retaining the one with the most rejections is a different procedure, for which the theorem above gives no guarantee.

Figure 6.3 makes this distinction visible by plotting both FDR and power as equicorrelation increases. The power panel records the attraction of estimating $\pi_0$, while the FDR panel shows why that gain must be interpreted together with the estimator's dependence assumptions.

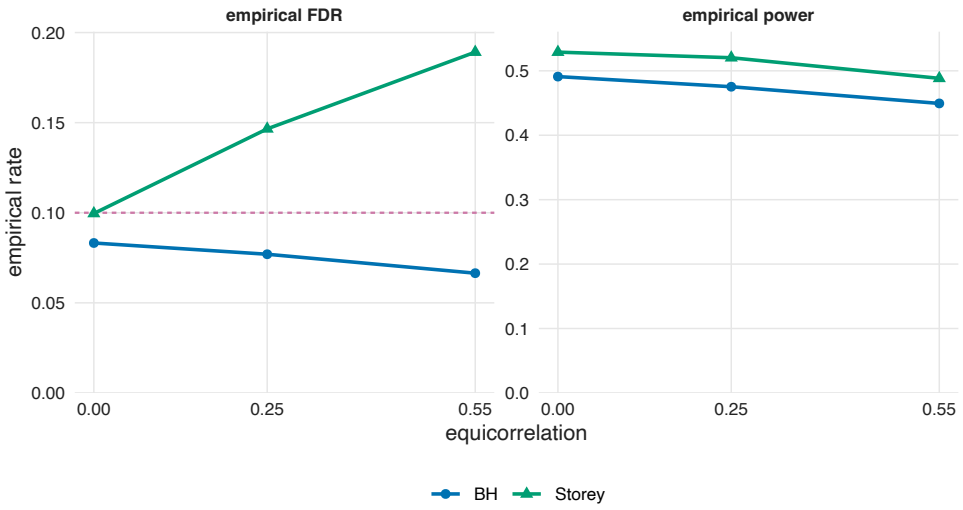



FIGURE 6.3. BH and the Storey procedure of Theorem 6.9 under equicorrelated one-sided Gaussian tests at target $q = 0.10$. Each configuration has $m = 400$ tests, including 65 nonnulls with mean 2.35; the Storey estimator uses $\lambda = 0.5$, retains values above one, and restricts rejection thresholds to at most $\lambda$. All Gaussian coordinates have variance one. Points show empirical FDR and empirical power, estimated from 900 independent simulation runs at correlations $\rho \in \{0, 0.25, 0.55\}$. The adaptive procedure gains power when the estimator is reliable, but this simulation also shows why dependence and adaptivity require explicit assumptions.

## 5. Q-Values

**Route: Core**

An adjusted value lets us report the result of a testing procedure without choosing just one nominal level. For example, a BH-adjusted p-value of 0.04 means that BH rejects that hypothesis at level $q = 0.04$ and at every larger level. Whether those rejections control FDR depends on the assumptions of the BH theorem.

Estimated Storey q-values use a similar calculation, with an estimate of the null fraction included. Order the observed p-values as $p_{(1)} \le \cdots \le p_{(m)}$, compute a positive estimate $\widehat{\pi}_0 > 0$ once, and hold it fixed as the nominal level varies. Define

$$\widehat{Q}_{(i)} = \min_{j \ge i} \left\{ \frac{\widehat{\pi}_0 m p_{(j)}}{j} \right\} \wedge 1.$$

Setting $\widehat{\pi}_0 = 1$ gives the usual BH-adjusted p-values. With an estimated null fraction, the values are commonly called estimated q-values. For $0 < q < 1$, rejecting when $\widehat{Q}_{(i)} \le q$ is equivalent to the adaptive step-up rule with critical values $qj/(m\widehat{\pi}_0)$. The formula describes its rejection sets; an arbitrary fitted $\widehat{\pi}_0$ does not, by itself, supply a finite-sample FDR guarantee.

A five-hypothesis calculation. Take $m = 5$, observed p-values $(0.005, 0.020, 0.025, 0.400, 0.800)$, and $\widehat{\pi}_0 = 0.8$. The middle row below first calculates $m\widehat{\pi}_0 p_{(i)}/i$; the last row takes the minimum from each column to the right-hand end.

| Rank $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $p_{(i)}$ | 0.005 | 0.020 | 0.025 | 0.400 | 0.800 |
| $m\widehat{\pi}_0 p_{(i)}/i$ | 0.020 | 0.040 | 1/30 | 0.400 | 0.640 |
| $\widehat{Q}_{(i)}$ | 0.020 | 1/30 | 1/30 | 0.400 | 0.640 |

Here $1/30 \approx 0.0333$. At level $q = 0.03$, the rule rejects only the first hypothesis. At $q = 0.035$, it rejects the first three, since the third rank passes the step-up comparison even though the second rank's raw value 0.04 exceeds $q$. At $q = 0.40$, it rejects the first four, and at $q = 0.64$, all five. More precisely, ranks 2 and 3 enter together at $q = 1/30$. The running minimum ensures that a smaller p-value never receives a larger adjusted value than a larger p-value. These rejection sets are *nested*: increasing the nominal level can add hypotheses but cannot remove a previously rejected hypothesis.

The threshold restriction $t \leq \lambda$ in Theorem 6.9 requires one further detail. To report adjusted values for that particular procedure, let $r_\lambda = \#\{i : p_i \leq \lambda\}$ and take the minimum only over $i \leq j \leq r_\lambda$, using $\widehat{\pi}_0^\lambda$. Hypotheses with $p_i > \lambda$ are ineligible for rejection. The unrestricted formula above does not encode this cap.

At the population level, Storey's q-value is defined using the positive FDR of a family of nested rejection regions; Section 10 gives that definition and its assumptions [169]. For now, the important distinction is that a q-value summarizes the error rate of a *set* of discoveries containing the hypothesis. For a one-sided statistic, such a set has the form $\{Z \geq c\}$. It may include observations less extreme than the current observation, because the q-value takes a minimum over all eligible regions containing it. A q-value is therefore different from the posterior probability that this one hypothesis is null. The two-groups model makes that difference precise.

## 6. The Two-Groups Model and Local FDR

**Route: Core**

The empirical-Bayes view starts by putting each hypothesis into a latent, or unobserved, state. Here $Z_i$ denotes the scalar evidence statistic for hypothesis $i$; in a one-sided Gaussian testing problem, larger $Z_i$ is stronger evidence against the null. Let $\theta_i = 0$ denote a null and $\theta_i = 1$ a nonnull. The two-groups model assumes

$$\mathbb{P}(\theta_i = 0) = \pi_0, \qquad Z_i \mid \theta_i = 0 \sim f_0, \qquad Z_i \mid \theta_i = 1 \sim f_1.$$

The marginal density is

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z).$$

These equations specify the marginal model for one hypothesis; they do not yet specify dependence between different hypotheses. Independence will be an additional assumption when we calculate a posterior given all of the observations below. The local FDR is the posterior null probability

$$\operatorname{lfdr}(z) = \mathbb{P}(\theta = 0 \mid Z = z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1 - \pi_0) f_1(z)}.$$

Figure 6.4 links this formula to the mixture geometry. The upper panel shows where the null and alternative densities contribute to $f$; the lower panel converts their pointwise shares into the posterior null probability $\operatorname{lfdr}(z)$. Regions dominated by the alternative therefore correspond to low local FDR.

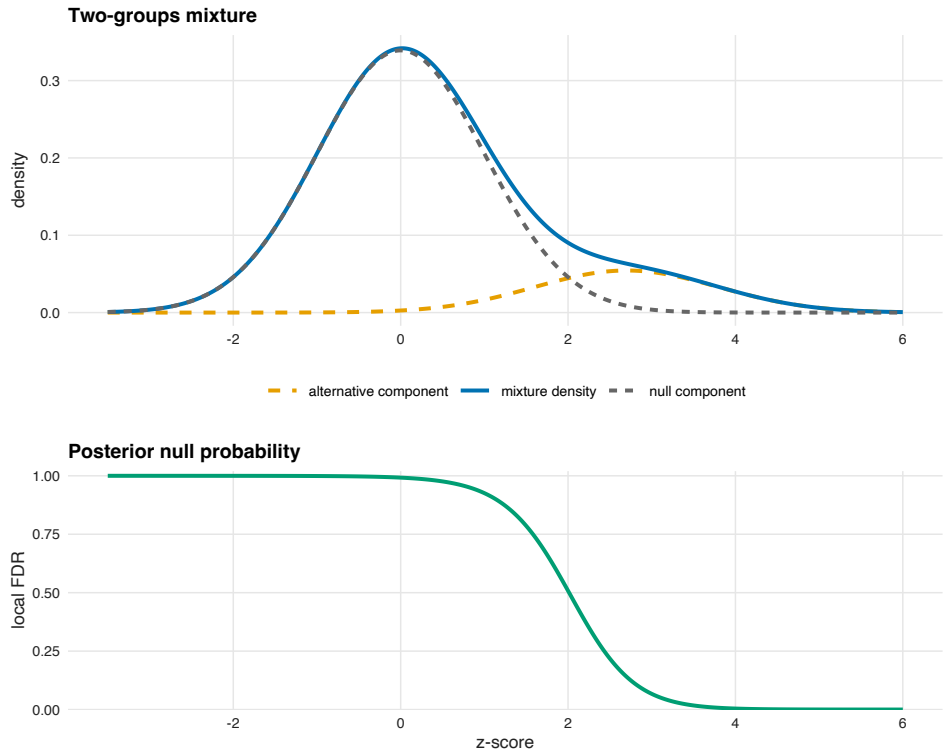



FIGURE 6.4. Two-groups model with $\pi_0 = 0.85$, $f_0 = N(0,1)$, and $f_1 = N(2.7, 1.1^2)$. The top panel shows the weighted null and alternative components, $\pi_0 f_0$ and $(1-\pi_0)f_1$, together with their mixture $f$. The bottom panel shows local FDR, the posterior probability that a case with z-score $z$ is null.

Local FDR is a pointwise quantity. A q-value is a tail-area quantity. In a one-sided problem, the posterior null probability for the tail $Z \geq z$ is

$$\mathbb{P}(\theta = 0 \mid Z \geq z).$$

This averages local FDR values over the whole rejection tail:

$$\mathbb{P}(\theta = 0 \mid Z \geq z) = \mathbb{E}\{\text{lfdr}(Z) \mid Z \geq z\}.$$

Here is the calculation. Let $I = \mathbf{1}\{\theta = 0\}$. Since $\text{lfdr}(Z) = \mathbb{P}(\theta = 0 \mid Z) = \mathbb{E}[I \mid Z]$, conditioning further on the tail event and applying iterated expectation gives

$$\begin{aligned}\mathbb{P}(\theta = 0 \mid Z \geq z) &= \mathbb{E}[I \mid Z \geq z] \\ &= \mathbb{E}\{\mathbb{E}[I \mid Z] \mid Z \geq z\} \\ &= \mathbb{E}\{\text{lfdr}(Z) \mid Z \geq z\}.\end{aligned}$$

Equivalently, the numerator of the tail probability is $\int_z^\infty \pi_0 f_0(u)\,du$, while averaging local FDR over the same tail gives

$$\frac{\int_z^\infty \text{lfdr}(u) f(u)\,du}{\int_z^\infty f(u)\,du} = \frac{\int_z^\infty \pi_0 f_0(u)\,du}{\int_z^\infty f(u)\,du}.$$

Thus $\text{lfdr}(z)$ is the posterior null probability for a case observed near $z$, while the tail-area quantity is the average posterior null probability among all cases at least as extreme as $z$. To connect this with q-values, let $p(z)$ be a right-tail p-value and consider the regions $\Gamma_u = \{z : p(z) \leq u\}$.

If $u < v$, then $\Gamma_u \subseteq \Gamma_v$, so increasing $u$ only adds observations. The population q-value of an observation is the smallest positive FDR over these regions that contain it; Section 10 gives the formal definition and the assumptions connecting positive FDR with the posterior tail probability calculated here. The two numbers can differ substantially, especially when the mixture density changes rapidly.

Figure 6.5 compares oracle local FDR with BH-adjusted p-values on the same ranked observations. Here *oracle* means that the true mixture parameters are supplied to the calculation. Each local FDR value is the posterior null probability for one observed case under that model. Each BH-adjusted p-value is calculated from the ranked sample, uses $m$ rather than an estimated null count, and takes a running minimum over ranks at or beyond the current rank. Consequently the difference between the plotted curves reflects the different targets, the BH null-count bound, and sampling variation. It is not explained by tail averaging alone.

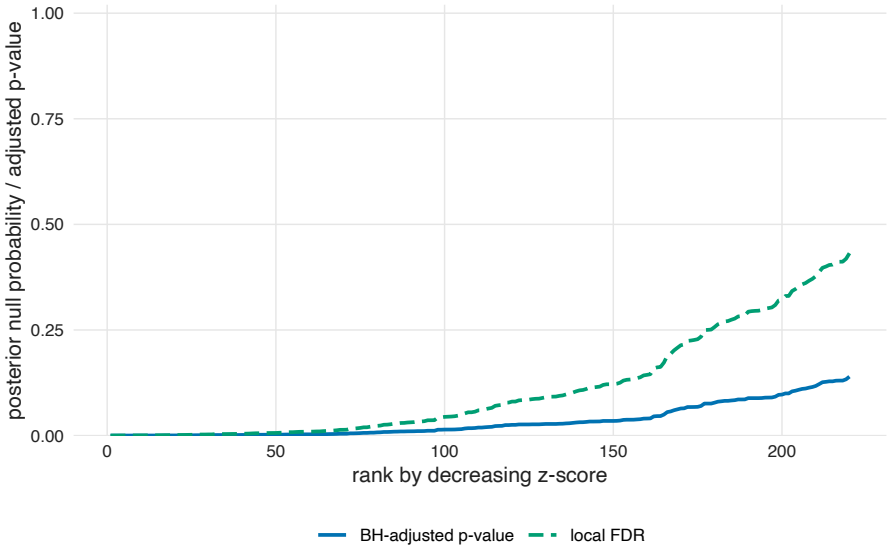



FIGURE 6.5. Oracle local FDR and BH-adjusted p-values for the 220 largest z-scores in one simulated sample of $m = 1800$ independent two-groups observations, with $\pi_0 = 0.85$, $f_0 = N(0, 1)$, and $f_1 = N(2.7, 1.1^2)$. Local FDR uses the known mixture model to calculate individual posterior null probabilities. BH-adjusted p-values give the nominal levels at which BH would reject the corresponding hypotheses; they are not individual posterior probabilities or exact population q-values.

## 7. Optimal Thresholding by Local FDR

**Route: Core**

The two-groups model also gives an optimization principle. This section treats the no-side-information setting: all hypotheses share the same null fraction $\pi_0$, null density $f_0$, and alternative density $f_1$. Suppose a rule rejects observations in a fixed region $\Gamma$. For $m$ hypotheses

with this marginal two-groups model, linearity of expectation gives the expected numbers of false and true discoveries:

$$\mathrm{EFP}(\Gamma) = m\pi_0 \int_\Gamma f_0(z)\,dz, \qquad \mathrm{ETP}(\Gamma) = m(1-\pi_0)\int_\Gamma f_1(z)\,dz.$$

The corresponding marginal false discovery rate (mFDR) is

$$\mathrm{mFDR}(\Gamma) = \frac{\mathrm{EFP}(\Gamma)}{\mathrm{EFP}(\Gamma)+\mathrm{ETP}(\Gamma)} = \frac{\mathbb{E}[V]}{\mathbb{E}[R]},$$

when $\mathbb{E}[R] > 0$. This ratio is not the same object as the frequentist FDR $\mathbb{E}[V/(R \vee 1)]$. Here it is the population error criterion for choosing a rejection region. An *oracle* rule is allowed to use the true $\pi_0$, $f_0$, and $f_1$, even though these quantities usually have to be estimated in an application. Studying the oracle rule first shows what a well-estimated procedure would try to approximate.

This objective can be written directly in terms of local FDR. Let

$$f(z) = \pi_0 f_0(z) + (1-\pi_0) f_1(z), \qquad \ell(z) = \mathrm{lfdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}.$$

Then

$$\mathrm{EFP}(\Gamma) = m\int_\Gamma \ell(z) f(z)\,dz, \qquad \mathbb{E}[R(\Gamma)] = m\int_\Gamma f(z)\,dz,$$

so $\mathrm{mFDR}(\Gamma)$ is the average value of $\ell(z)$ over the rejected region, with respect to the mixture density $f$. Thus a region containing mostly small local FDR values has a smaller average posterior null probability under the mixture distribution.

The following oracle statement is the continuous version of this sorting principle. To include boundary randomization, let a procedure be a measurable function $\delta(z) \in [0,1]$, where $\delta(z)$ is the probability of rejecting an observation with statistic $z$. A deterministic rejection region $\Gamma$ is the special case $\delta(z) = \mathbf{1}\{z \in \Gamma\}$. For such a rule,

$$\mathrm{mFDR}(\delta) = \frac{\int \delta(z)\ell(z)f(z)\,dz}{\int \delta(z) f(z)\,dz},$$

when the denominator is positive.

**Proposition 6.10: Oracle local FDR level sets**

Let $0 < q < 1$. Among all oracle procedures $\delta$ satisfying $\mathrm{mFDR}(\delta) \le q$, a procedure maximizing expected true discoveries can be chosen to reject a lower level set of local FDR:

$$\delta_c(z) = \mathbf{1}\{\ell(z) < c\} + \rho(z)\mathbf{1}\{\ell(z) = c\},$$

for some $c \in [0,1]$ and some boundary randomization $0 \le \rho(z) \le 1$. Include the no-rejection rule as feasible, with error criterion zero. If $\ell(Z)$ has a continuous distribution, no boundary randomization is needed: take $\Gamma_{c_*} = \{z : \ell(z) \le c_*\}$, where

$$c_* = \sup\left\{ c \in [0,1] : \int_{\{\ell(z)\le c\}} \{\ell(z) - q\} f(z)\,dz \le 0 \right\}.$$

For a region of positive probability, this constraint is equivalent to

$$\frac{\int_{\{\ell(z)\le c\}} \ell(z) f(z)\,dz}{\int_{\{\ell(z)\le c\}} f(z)\,dz} \le q.$$

Thus the rule includes observations in increasing order of local FDR until the average for the rejected region would exceed $q$. If no region of positive probability is feasible, it makes no rejections.

**Proof of Proposition 6.10**

Write integrals with respect to the mixture measure $d\nu(z) = f(z)\,dz$. For a rule $\delta$, define its rejection mass and false-discovery mass by

$$r(\delta) = \int \delta(z)\,d\nu(z), \qquad v(\delta) = \int \delta(z)\ell(z)\,d\nu(z).$$

Then $\mathrm{mFDR}(\delta) = v(\delta)/r(\delta)$ when $r(\delta) > 0$. If no positive-mass rule is feasible, the no-rejection rule is optimal and the claim is trivial. Hence assume below that the optimizer has positive rejection mass. The mFDR constraint is therefore

$$v(\delta) - qr(\delta) \le 0,$$

and the expected true-discovery mass is

$$\int \delta(z)\{1-\ell(z)\}\,d\nu(z) = r(\delta) - v(\delta).$$

The Lagrange calculation makes the shape of the optimum transparent. Attach a multiplier $\eta \ge 0$ to the constraint and consider the penalized objective

$$\begin{aligned}\mathcal{L}_\eta(\delta) &= \{r(\delta) - v(\delta)\} - \eta\{v(\delta) - qr(\delta)\} \\ &= \int \delta(z)\,[(1-\ell(z)) - \eta\{\ell(z) - q\}]\;d\nu(z) \\ &= \int \delta(z)\,[1 + \eta q - (1+\eta)\ell(z)]\;d\nu(z).\end{aligned}$$

For a fixed $\eta$, this integral is maximized pointwise: set $\delta(z) = 1$ where the coefficient of $\delta(z)$ is positive, set $\delta(z) = 0$ where it is negative, and allow randomization where it is zero. For this fixed penalty, the maximizing rule has the form

$$\delta(z) = 1\{\ell(z) < c_\eta\} + \rho(z)1\{\ell(z) = c_\eta\}, \qquad c_\eta = \frac{1+\eta q}{1+\eta}.$$

All boundary points therefore have the same local FDR value. The limiting case $\eta \to \infty$ gives $c_\eta \to q$. This penalty calculation suggests the form of the constrained optimum; the following argument justifies the form without requiring a finite optimizing multiplier.

The exchange argument below is the self-contained justification that the Lagrange form loses no generality. Fix the rejection mass $r(\delta) = r > 0$. Among rules with this same mass, maximizing expected true discoveries is the same as minimizing $v(\delta)$. If a rule puts positive rejection probability on a set $B$ with larger local FDR values while not fully rejecting a set $A$ with smaller values, take smaller subsets if necessary so the available $\nu$-masses match, and move a small amount of rejection probability from $B$ to $A$. The rejection mass $r$ is unchanged, but $v(\delta)$ decreases because the moved mass is charged a smaller value of $\ell$. Hence $r - v(\delta)$ increases. Repeating this exchange forces the rule, up to ties, to fill the smallest local FDR values first, with randomization only on a boundary set $\{\ell = c\}$.

Now take any mFDR-feasible rule and replace it by this lower-level-set rule with the same rejection mass. The replacement cannot increase $v(\delta)/r$ and cannot decrease $r - v(\delta)$, so an optimizer may be chosen from the family of lower level sets of local FDR. Within that

family, increasing the rejection mass cannot decrease expected true discoveries, and the average local FDR cannot decrease. We therefore include as much probability mass as the constraint allows. If $\ell(Z)$ has a continuous distribution, the integral in the definition of $c_*$ is continuous in $c$, so the limiting cutoff is feasible. This also handles an empty rejection region without dividing by zero. With an atom at the boundary, choose $\rho$ to include the largest feasible fraction of its mass. If rejecting everything is feasible, that rule is optimal. □

The finite-sample posterior version of the same argument treats local FDR as the probability of making an error when rejecting one observed case. For this calculation, assume the pairs $(\theta_i, Z_i)$, $i = 1, \ldots, m$, are independent and the mixture parameters are known. Then other hypotheses' statistics add no information about $\theta_i$ once $Z_i$ is known:

$$\ell_i = \mathrm{lfdr}(Z_i) = \mathbb{P}(\theta_i = 0 \mid Z_i) = \mathbb{P}(\theta_i = 0 \mid Z_1, \ldots, Z_m).$$

For a rejection set $A$ chosen from these observations, the posterior expected number of false discoveries is

$$\mathbb{E}[V(A) \mid Z_1, \ldots, Z_m] = \sum_{i \in A} \ell_i.$$

Since $A$, and hence its size, is fixed after conditioning on the data, the posterior expected FDP is

$$\mathbb{E}\left[\frac{V(A)}{|A| \vee 1} \,\middle|\, Z_1, \ldots, Z_m\right] = \frac{\sum_{i \in A} \ell_i}{|A| \vee 1}.$$

This conditional expectation is not the population ratio $\mathrm{mFDR} = \mathbb{E}[V]/\mathbb{E}[R]$ defined above. It evaluates the particular set selected after observing the data. A discovery with smaller $\ell_i$ has a smaller probability of being false, so the oracle rule starts with the smallest local FDR values.

**Proposition 6.11: Oracle local FDR step-up rule**

Under the independent two-groups model with known parameters just specified, suppose the oracle local FDR values $\ell_i = \mathrm{lfdr}(Z_i)$ are known, and order them as $\ell_{(1)} \le \ell_{(2)} \le \cdots \le \ell_{(m)}$. Among rejection sets of size $k$, the set containing the $k$ smallest $\ell_i$'s minimizes the posterior expected number of false discoveries. Among all sets with posterior expected FDP at most $q$, a set with the most rejections is obtained by taking the first $\widehat{k}$ hypotheses in this ordering, where

$$\widehat{k} = \max\left\{1 \le k \le m : \frac{1}{k}\sum_{j=1}^{k} \ell_{(j)} \le q\right\},$$

with the convention $\widehat{k} = 0$ if the set is empty. Its posterior expected FDP is at most $q$.

**Proof of Proposition 6.11**

Under the stated independence assumption, $\ell_i$ is the posterior probability that hypothesis $i$ is null given all the observed statistics. Therefore the posterior expected number of false discoveries in a rejection set $A$ is $\sum_{i \in A} \ell_i$, while the number of rejections is $|A|$. For any fixed size $k$, suppose $A$ contains an index $a$ with $\ell_a > \ell_b$ while omitting an index $b$. Swapping $a$ out and $b$ in changes the posterior expected false count by $\ell_b - \ell_a < 0$ and leaves the size equal to $k$. Repeating this exchange proves that the unique minimal set, up to ties, consists of the $k$ smallest $\ell_i$'s.

For the set containing the first $k$ hypotheses in the ordering, the posterior expected FDP is the average $a_k = k^{-1} \sum_{j=1}^{k} \ell_{(j)}$. These averages cannot decrease:

$$a_{k+1} - a_k = \frac{\ell_{(k+1)} - a_k}{k+1} \geq 0.$$

The displayed rule chooses the largest $k$ for which this average is at most $q$. Any other set of size $k$ has an average at least $a_k$, by the exchange argument. Thus no larger set can satisfy the same posterior expected FDP bound. □

Averaging the conditional FDP bound over the observations gives FDR $\leq q$ under this two-groups model. That expectation averages over the random null states as well as the data. It does not establish FDR control for every fixed configuration of true and false hypotheses. Moreover, if the pairs are dependent, the relevant scores for this posterior argument are $\mathbb{P}(\theta_i = 0 \mid Z_1, \ldots, Z_m)$; the marginal local FDR $\mathrm{lfdr}(Z_i)$ need not equal that probability.

In practice the $\ell_i$ are replaced by estimates $\widehat{\ell}_i$ from a fitted mixture model. A bound on their average is then only a bound on the *estimated* posterior error. Further assumptions and an error analysis are needed to transfer the oracle guarantee to the fitted procedure.

This explains why empirical-Bayes ranking can differ from p-value ranking. If the alternative distribution is asymmetric, contains subgroups with different variances, or has several peaks, observations with small posterior null probabilities need not be the ones with the smallest classical p-values.

In practice, $\pi_0$, $f_0$, and $f_1$ must be estimated. One approach fits a specified mixture by maximum likelihood. An expectation-maximization (EM) algorithm alternates between estimating group memberships and updating the mixture parameters. Another approach estimates the null distribution from the central part of the statistic histogram, where null cases are expected to be common. Each approach needs assumptions about the mixture and about which features of the data identify the null distribution. The resulting posterior probabilities describe evidence under the fitted model; the fitting method alone does not establish a finite-sample FDR bound.

## 8. Covariate-Assisted Local FDR

**Route: Advanced**

The previous section assumed that all hypotheses share the same mixture weights. Side information lets those weights, or the alternative distribution, depend on observed covariates. Let $X_i \in \mathcal{X}$ be a covariate observed before the testing decision, and let $P_i$ denote the random p-value with observed realization $p_i$. A covariate-assisted p-value mixture model has the form

$$\theta_i \mid X_i = x \sim \mathrm{Bernoulli}(1 - \pi_0(x)), \qquad P_i \mid X_i = x, \theta_i = 0 \sim \mathrm{Unif}[0,1],$$

and

$$P_i \mid X_i = x, \theta_i = 1 \sim f_1(\cdot \mid x).$$

The covariate is useful when either the conditional null fraction $\pi_0(x)$ or the conditional alternative density $f_1(\cdot \mid x)$ varies with $x$. For example, in a large genomic screen, genes or deoxyribonucleic acid (DNA) methylation sites with higher read depth, stronger baseline expression, known functional annotation, or corroborating evidence from a previous genome-wide association study (GWAS) may have different prior chances of being nonnull or different probabilities of producing small p-values under alternatives. Conditional null uniformity is a substantive assumption: it

must hold within each covariate stratum, and does not follow just from observing $X_i$ before the final testing decision. Bayes' rule gives the covariate local FDR

$$\mathrm{lfdr}(p, x) = \mathbb{P}(\theta_i = 0 \mid P_i = p, X_i = x) = \frac{\pi_0(x)}{\pi_0(x) + (1 - \pi_0(x)) f_1(p \mid x)}.$$

Thus the same p-value can carry different posterior evidence at different covariate values. For example, suppose two strata both have alternative density $f_1(0.02 \mid x) = 4$, but their null fractions are 0.8 and 0.5. The corresponding local FDR values at $p = 0.02$ are

$$\frac{0.8}{0.8 + 0.2 \times 4} = 0.50, \qquad \frac{0.5}{0.5 + 0.5 \times 4} = 0.20.$$

The observation is less likely to be null in the second stratum because nonnull cases are more common there. Changes in the alternative density also matter, but greater power does not guarantee smaller local FDR at every fixed p-value. Power concerns the probability that an alternative p-value falls in a rejection region; the posterior at a particular $p$ depends on the density $f_1(p \mid x)$ at that point.

**Proposition 6.12: Oracle covariate local FDR thresholding**

Assume that, conditional on the entire covariate vector $(X_1, \ldots, X_m)$, the pairs $(\theta_i, P_i)$ are independent, and that the conditional law of the $i$th pair is the displayed model with known functions $\pi_0(X_i)$ and $f_1(\cdot \mid X_i)$. The oracle local FDR values at the observed data are $\ell_i = \mathrm{lfdr}(p_i, X_i)$. Order them as $\ell_{(1)} \le \cdots \le \ell_{(m)}$. Define

$$\widehat{k} = \max\left\{1 \le k \le m : \frac{1}{k}\sum_{j=1}^{k} \ell_{(j)} \le q\right\},$$

with $\widehat{k} = 0$ if the set is empty, and reject the $\widehat{k}$ hypotheses with smallest $\ell_i$'s. Conditional on the observed data $(p_i, X_i)_{i=1}^m$, the posterior expected FDP of this rejection set is at most $q$. Among rejection sets of any fixed size, the set with the smallest posterior expected number of false discoveries rejects the smallest local FDR values. The sets consisting of the first $k$ hypotheses in this ordering grow by one hypothesis as $k$ increases. The displayed rule chooses the largest such set whose average local FDR is at most $q$, and no larger set satisfies the same posterior expected FDP bound.

**Proof of Proposition 6.12**

Write $\mathcal{D} = (P_1, \ldots, P_m, X_1, \ldots, X_m)$ for the full data. The conditional independence assumption implies that the other p-values and covariates add no information about $\theta_i$ once $(P_i, X_i)$ is known. Hence $\mathbb{P}(\theta_i = 0 \mid \mathcal{D}) = \mathrm{lfdr}(P_i, X_i)$, whose observed value is $\ell_i$. If a rejection set $A$ is chosen from the data, its rejection count is fixed at $R(A) = |A|$ after conditioning, while

$$V(A) = \sum_{i \in A} \mathbf{1}\{\theta_i = 0\}.$$

Taking posterior expectation conditional on the observed data gives

$$\mathbb{E}[V(A) \mid \mathcal{D}] = \sum_{i \in A} \mathbb{P}(\theta_i = 0 \mid \mathcal{D}) = \sum_{i \in A} \ell_i.$$

Therefore the posterior expected FDP, with the usual convention that the FDP is zero when $A$ is empty, is

$$\mathbb{E}\left[\frac{V(A)}{|A| \vee 1} \,\middle|\, \mathcal{D}\right] = \frac{1}{|A| \vee 1} \sum_{i \in A} \ell_i.$$

For fixed $|A| = k$, the denominator is fixed, so minimizing the posterior expected FDP is the same as minimizing $\sum_{i \in A} \ell_i$. If $A$ contains an index $a$ with $\ell_a > \ell_b$ and omits $b$, replacing $a$ by $b$ lowers the sum by $\ell_a - \ell_b$. Repeating this exchange yields the set of the $k$ smallest local FDR values, up to ties.

For that optimal size-$k$ set, the posterior expected FDP is $k^{-1} \sum_{j=1}^{k} \ell_{(j)}$. The step-up rule considers the sets containing the first $k$ hypotheses in the ordering and chooses the largest $k$ whose average is at most $q$. Thus its posterior expected FDP is at most $q$. For any $r > \widehat{k}$, a set of size $r$ has average local FDR at least $r^{-1} \sum_{j=1}^{r} \ell_{(j)}$. This last average exceeds $q$ by the definition of $\widehat{k}$, so no larger set can satisfy the bound. □

As in the no-covariate case, averaging this posterior FDP bound gives FDR control under the specified random-state model. The result requires the true posterior probabilities. In practice, $\pi_0(x)$ and $f_1(p \mid x)$ are estimated, for example by smoothing the null fraction across covariates and fitting conditional alternative densities. To justify the fitted rule asymptotically, one must show that the estimated local FDR averages approximate the true posterior averages accurately enough on the sets actually selected. Estimating the functions is not itself such a proof.

One danger is overfitting the p-values that will later be tested. A flexible model can treat an accidental cluster of small null p-values in one covariate region as evidence for many nonnulls there. It then assigns those same cases estimated local FDR values that are too small, and their average understates the error. AdaPT, developed next, allows a fitted model to guide the search while restricting the information it can use. For a hypothesis still eligible for rejection, the model sees the pair $\{p_i, 1 - p_i\}$ without learning which member is the observed p-value. It may use the covariates, permitted aggregate counts, and actual p-values of hypotheses already removed from consideration. This masking and the corresponding stopping rule provide the FDR argument under AdaPT's assumptions, without requiring the fitted local FDR values to be exact posterior probabilities.

## 9. AdaPT: Masking Guided by Local FDR

**Route: Advanced**

AdaPT uses a fitted model to decide which hypotheses look promising, while restricting the information available for making that decision. This allows it to learn different p-value thresholds for different covariate values and still control FDR in finite samples. The fitted model guides the search; a separate comparison of lower-tail and upper-tail counts decides when to stop. We describe the version with uniform null p-values. Throughout this section, $P_i$ is a random p-value and $p_i$ is its observed value.

What is hidden, and what is revealed? For each observation, write

$$a_i = \min(p_i, 1 - p_i), \qquad b_i = \mathbf{1}\{p_i \geq 1/2\}.$$

Knowing $a_i$ tells us that the p-value is either $a_i$ or $1 - a_i$. The indicator $b_i$ tells us which: $b_i = 0$ places it in the lower half, and $b_i = 1$ places it in the upper half. Initially, AdaPT supplies the covariates and these unordered pairs, but does not identify the actual p-value within each pair.

The subscript $t = 0, 1, \ldots$ below denotes the round of the algorithm. At round $t$, a threshold function $s_t : \mathcal{X} \to [0, 1/2]$ specifies a cutoff for each covariate value. The current rejection and mirror counts are

$$R_t = \#\{i : p_i \le s_t(X_i)\}, \qquad B_t = \#\{i : p_i \ge 1 - s_t(X_i)\}.$$

The lower tail contains the proposed discoveries, and the upper tail provides comparison observations of the same covariate-dependent width. The information supplied about the individual p-values is

$$\widetilde{p}_{t,i} = \begin{cases} p_i, & s_t(X_i) < p_i < 1 - s_t(X_i), \\ \{p_i, 1 - p_i\}, & \text{otherwise.} \end{cases}$$

Thus the central p-values are fully revealed. For an observation in either tail, the algorithm supplies the unordered pair. It also supplies the *aggregate counts* $R_t$ and $B_t$. These counts are available even though the individual tail assignments are not supplied. Let $\mathcal{F}_t$ denote all this information, including every covariate and the information supplied in earlier rounds. In probability terminology, $(\mathcal{F}_t)$ is the analyst's filtration: the record of what is known as the procedure progresses.

The sequence of decisions. Choose an initial surface $s_0$ using only the covariates and the unordered pairs; the simple choice $s_0(x) = 1/2$ is allowed. Fix a finite maximum number of rounds $T$. At each round, compute

$$\widehat{\mathrm{FDP}}_t = \frac{1 + B_t}{R_t \vee 1}.$$

If this quantity is at most the target $q$, stop and reject precisely $\{i : p_i \le s_t(X_i)\}$. If it exceeds $q$, choose a new surface using only $\mathcal{F}_t$, subject to

$$s_{t+1}(x) \le s_t(x) \qquad \text{for every } x \in \mathcal{X}.$$

The cutoffs may decrease by different amounts in different covariate regions. They cannot increase. If no round through $T$ meets the stopping condition, report no rejections. This finite limit is a convenient way to make the stopping rule completely specified.

The decreasing cutoffs explain how the procedure learns. Suppose a masked pair is $\{0.02, 0.98\}$ and its current cutoff is 0.05. Lowering that cutoff below 0.02 puts both possible p-values in the central region, so the actual value is revealed. At the same time, this hypothesis leaves the set eligible for rejection. The analyst can learn from its revealed value, but cannot later increase its cutoff to bring it back. All information already revealed remains available.

How local FDR can guide the next cutoff. The model in the preceding section gives a natural way to decide where to reduce the threshold. Fit a local FDR function $\widehat{\ell}_t(p, x)$ using the information available at round $t$. A masked pair must be treated as ambiguous when fitting the model; the fitting routine cannot inspect the underlying p-value. For example, one possible working model is

$$f(p \mid x) = \pi_0(x) + \{1 - \pi_0(x)\}\eta(x)p^{\eta(x)-1}, \qquad 0 < \eta(x) < 1.$$

Here the alternative p-value density is a beta density concentrated near zero. The corresponding fitted local FDR is

$$\widehat{\ell}_t(p, x) = \frac{\widehat{\pi}_{0,t}(x)}{\widehat{\pi}_{0,t}(x) + \{1 - \widehat{\pi}_{0,t}(x)\}\widehat{\eta}_t(x)p^{\widehat{\eta}_t(x)-1}}.$$

For fixed $x$, this score increases with $p$: a larger p-value is less convincing evidence against the null. For such a monotone fit, a possible proposal is

$$s_{t+1}(x) = \min\left\{s_t(x),\ \sup\{u \in (0, 1/2] : \widehat{\ell}_t(u, x) \le c_t\}\right\},$$

where an empty supremum is taken to be zero. Choose the tuning value $c_t$ from the permitted information to reduce at least one cutoff when further search is needed. The minimum with $s_t(x)$ enforces the rule that cutoffs only decrease. Other models and update rules are possible.

As a concrete illustration, suppose a fitted model has $\eta(x) = 1/2$, with $\pi_0(x) = 0.2$ in one covariate group and 0.8 in another. The same local FDR cutoff $c = 0.2$ gives p-value cutoffs 0.25 and approximately 0.00098, respectively, before imposing the current surface. The first group can retain larger p-values because its fitted fraction of null hypotheses is lower. These numbers illustrate how a model allocates the cutoffs; they do not themselves certify FDR control. In particular, $c_t$ is a tuning value for the model-based search and need not equal the target $q$. The algorithm still stops only when the mirror estimate $(1 + B_t)/(R_t \vee 1)$ is at most $q$.

**Proposition 6.13: Conditional symmetry under masking**

For a true-null hypothesis with $P \mid X \sim \mathrm{Unif}(0, 1)$, let $M = \min(P, 1 - P)$ and $B = \mathbf{1}\{P \geq 1/2\}$. Then $B$ is a Bernoulli(1/2) random variable independent of the pair $(M, X)$.

**Proof of Proposition 6.13**

For $0 \leq u \leq 1/2$, conditional uniformity gives

$$\mathbb{P}(M \leq u, B = 0 \mid X) = u, \qquad \mathbb{P}(M \leq u, B = 1 \mid X) = u.$$

The two events correspond to $P \in [0, u]$ and $P \in [1 - u, 1]$, respectively. Consequently $M \mid X$ is uniform on $[0, 1/2]$, and its conditional distribution is the same for either value of $B$. Moreover, each value of $B$ has conditional probability $1/2$ for almost every $X$. Thus knowing the masked value and the covariate does not change the fair-coin distribution of $B$. □

**Theorem 6.14: AdaPT FDR control under conditional independence**

Let $\mathbf{X} = (X_1, \ldots, X_m)$. Conditional on $\mathbf{X}$, assume that the true-null p-values are mutually independent, each has distribution $\mathrm{Unif}(0, 1)$, and their vector is independent of the non-null p-values. Run the finite-round procedure above: choose $s_0$ from covariates and unordered pairs, choose each $s_{t+1}$ from $\mathcal{F}_t$, and require $s_{t+1}(x) \leq s_t(x)$ for every $x$. Stop at the first round with $(1 + B_t)/(R_t \vee 1) \leq q$, and make no rejections if there is no such round by $T$. Then the reported rejection set has FDR at most $q$.

Why the stopping rule is valid. Let $V_t$ count the true-null p-values below $s_t(X_i)$, and let $U_t$ count the true-null p-values above $1 - s_t(X_i)$. Only $R_t$ and $B_t$ are observable, since the true nulls are unknown. We always have $U_t \leq B_t$. At a round $\tau$ where the procedure stops and rejects,

$$\mathrm{FDP} = \frac{V_\tau}{R_\tau \vee 1} = \frac{1 + U_\tau}{R_\tau \vee 1} \frac{V_\tau}{1 + U_\tau} \leq q \, \frac{V_\tau}{1 + U_\tau}.$$

If the procedure reaches $T$ without rejecting, set $\tau = T$; the same inequality holds because its reported FDP is zero.

Conditional on all initial masked values, the covariates, and the non-null p-values, the null tail indicators are independent fair coins. At later rounds, however, the analyst also knows aggregate tail counts. Conditioning on those counts can make the remaining indicators dependent, so it

would be incorrect to claim that they stay independent fair coins given everything the analyst has seen. The stopping-time argument in Lei and Fithian [117] accounts for exactly this additional information. For decreasing sets of masked hypotheses and updates based on the permitted information, it gives

$$\mathbb{E}\left[\frac{V_\tau}{1+U_\tau}\right] \le 1.$$

Combining this bound with the preceding FDP inequality proves the theorem. The added one in the denominator of this stopped comparison is the reason for the added one in AdaPT's mirror estimate.

A badly fitted local FDR model can lead to few discoveries, but it does not invalidate this guarantee if the assumptions and information restrictions hold. A fit using unrevealed p-values, or an update that raises a previous cutoff, falls outside the procedure just analyzed. Covariates must also satisfy the conditional null assumptions in the theorem; their being useful predictors of discoveries does not by itself establish those assumptions.

## 10. Positive FDR

**Route: Core**

The positive false discovery rate (pFDR) describes the expected error proportion conditional on making at least one discovery. When $\mathbb{P}(R > 0) > 0$, it is

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{V}{R}\,\middle|\, R > 0\right].$$

Since the FDP is defined to be zero when $R = 0$, the law of total expectation gives

$$\mathrm{FDR} = \mathrm{pFDR}\ \mathbb{P}(R > 0).$$

Thus FDR averages over all repetitions of an experiment, including those with no discoveries, whereas pFDR averages over repetitions that produce a nonempty list. For example, if a procedure makes discoveries with probability 0.1 and has pFDR 0.2, its FDR is 0.02. If it never makes a discovery, its FDR is zero and the conditional definition of pFDR is not applicable.

Under an independent two-groups model, pFDR also has a useful interpretation for a fixed rejection region.

**Proposition 6.15: Posterior interpretation of pFDR**

Suppose $(\theta_i, Z_i)$, $i = 1, \dots, m$, are independent and identically distributed copies of a pair $(\theta, Z)$ from the two-groups model, where $\theta = 0$ denotes a true null. Let $\Gamma$ be a fixed rejection region with $\mathbb{P}(Z \in \Gamma) > 0$, and reject $i$ precisely when $Z_i \in \Gamma$. Then

$$\mathrm{pFDR}(\Gamma) = \mathbb{P}(\theta = 0 \mid Z \in \Gamma).$$

**Proof of Proposition 6.15**

Write $r_\Gamma = \mathbb{P}(\theta = 0 \mid Z \in \Gamma)$. First condition on which observations fall in $\Gamma$. Independence of the pairs means that, for each selected observation, its null indicator has probability $r_\Gamma$ of being one; these indicators remain independent. Therefore, if exactly $k > 0$ observations

are selected, the expected number of nulls among them is $kr_\Gamma$. Averaging over all possible sets of $k$ selected indices gives

$$\mathbb{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\,\middle|\, R(\Gamma)=k\right]=r_\Gamma.$$

The condition $\mathbb{P}(Z\in\Gamma)>0$ ensures $\mathbb{P}(R(\Gamma)>0)>0$. Averaging the last display over the positive values of $k$ proves the result. □

This is a probability for a case selected into the whole region $\Gamma$. It averages over all statistic values in that region. Local FDR instead conditions on the individual observed value $Z=z$. Also, the proposition assumes an i.i.d. model and a fixed region. It does not automatically apply to a region estimated from the same observations, or to dependent observations.

To define a population q-value, specify a nested family of rejection regions $\{\Gamma_u : 0<u<1\}$, with the direction

$$u\le v \quad\Longrightarrow\quad \Gamma_u\subseteq\Gamma_v.$$

Larger $u$ therefore gives a less stringent rule and a larger rejection region. For a one-sided standard-normal null, a concrete example is

$$\Gamma_u=\{z : 1-\Phi(z)\le u\}=\{z : z\ge\Phi^{-1}(1-u)\}.$$

Here $u$ is the p-value cutoff. For an observation $z$ belonging to at least one region of positive probability, define

$$Q(z)=\inf_{\substack{0<u<1:\ z\in\Gamma_u,\\ \mathbb{P}(Z\in\Gamma_u)>0}}\mathrm{pFDR}(\Gamma_u).$$

The positive-probability restriction makes every conditional error rate in the infimum well-defined. The q-value is the smallest error level, in the sense of an infimum, among regions in this family that would include $z$. Taking an infimum also allows for the possibility that the tail-area error is not monotone in the cutoff. This population definition explains the tail-area interpretation of q-values discussed earlier; estimating it requires estimating the relevant mixture probabilities.

## 11. Conditional Calibration

**Route: Research frontier**

Known dependence can sometimes be used to obtain more discoveries than a correction designed for arbitrary dependence. Conditional calibration does this by examining each hypothesis's contribution to the FDR and choosing its rejection cutoff accordingly. It requires a dependence model that supports the necessary conditional calculations.

Calibrate contributions to the error rate. For any rejection set $\mathcal{R}$, let $R=|\mathcal{R}|$. Decompose its FDR as

$$\mathrm{FDR}=\sum_{i\in\mathcal{I}_0}\mathbb{E}\left[\frac{\mathbf{1}\{i\in\mathcal{R}\}}{R\vee 1}\right].$$

Each true null contributes zero when it is not rejected, and contributes $1/R$ when it is rejected. This denominator matters: rejecting a null in a list of two discoveries contributes more to the FDP than rejecting it in a list of two hundred.

For each $i$, choose a conditioning statistic $S_i$ that makes the relevant distribution under $H_i$ tractable. The statistics can differ across hypotheses. A sufficient target is

$$\mathbb{E}\left[\frac{\mathbf{1}\{i\in\mathcal{R}\}}{R\vee 1}\,\middle|\, S_i\right]\le\frac{q}{m}\qquad\text{whenever } H_i \text{ is true.}$$

This inequality must hold under each distribution in the specified model that satisfies $H_i$. Taking expectations removes the conditioning, and summing over the true nulls then gives $\mathrm{FDR} \le m_0 q/m \le q$. This short argument explains what is being calibrated. Constructing useful cutoffs that satisfy it is the difficult part, because the rejection event and the denominator both depend on the data.

A Gaussian example of the conditional calculation. Suppose

$$Z \sim N_m(\mu, \Sigma), \qquad \Sigma_{ii} = 1,$$

where $\Sigma$ is known and positive definite, and consider $H_i : \mu_i = 0$. The means of the other coordinates are unknown nuisance parameters. Conditioning directly on $Z_{-i}$ generally leaves the null law of $Z_i$ dependent on those unknown means. Instead, define

$$S_i = Z_{-i} - \Sigma_{-i,i} Z_i.$$

Its covariance with $Z_i$ is zero, since

$$\mathrm{Cov}(S_i, Z_i) = \Sigma_{-i,i} - \Sigma_{-i,i} \,\mathrm{Var}(Z_i) = 0.$$

Joint normality therefore makes $S_i$ independent of $Z_i$. Under $H_i$, the conditional distribution $Z_i \mid S_i$ is exactly $N(0,1)$. Once $S_i = s$ is fixed, a possible value $z_i$ reconstructs the entire data vector through

$$Z_i = z_i, \qquad Z_{-i} = s + \Sigma_{-i,i} z_i.$$

Consequently, a conditional expectation can be evaluated by varying $z_i$ under its standard-normal null law and recalculating the rejection rule on each reconstructed vector. This calculation tracks both whether hypothesis $i$ is rejected and how the number of discoveries changes. It does not require knowing the other means.

Connection with dependence-adjusted BH. The dependence-adjusted BH method of Fithian and Lei [62] starts with BH-adjusted p-values: the adjusted value for hypothesis $i$ is the smallest nominal level at which ordinary BH would reject it. Denote it by $p^{\mathrm{adj}}_{i,\mathrm{BH}}$, to distinguish it from the population q-value just defined. Conditional calibration chooses a separate cutoff $\widehat{c}_i$ for each hypothesis, leading to a proposed rejection set

$$\mathcal{R}_+ = \{i : p^{\mathrm{adj}}_{i,\mathrm{BH}} \le \widehat{c}_i\}.$$

A calibrated cutoff can be above or below $q$, according to the hypothesis's conditional contribution to the error rate.

The cutoffs cannot be chosen independently of the denominator issue. The construction in Fithian and Lei [62] uses a suitable estimate $\widehat{R}_i \ge 1$ of the eventual rejection count when calibrating each conditional contribution. If $\widehat{R}_i$ is a lower bound on the final number of rejections whenever $i$ is rejected, dividing by $\widehat{R}_i$ gives a conservative contribution. When the proposed set does not satisfy the required count bounds, the general construction includes a carefully specified randomized pruning step. This safeguard, and the conditional calibration itself, are part of the validity argument; the display for $\mathcal{R}_+$ alone is not a complete procedure.

Making individual p-values conditionally uniform is therefore not sufficient by itself to justify ordinary BH under dependence. The joint behavior of the rejection decisions and their count must also be controlled. Conditional calibration addresses that behavior directly. It offers a useful option when the dependence structure is known well enough for exact conditional calculations or valid conservative bounds. Merely fitting an approximate dependence model does not automatically preserve its finite-sample guarantee. BH under justified PRDS assumptions and BY under arbitrary dependence remain the simpler alternatives developed earlier in this chapter.

## 12. Assumptions in Plain Language

**Route: Core**

| Assumptions in practice | |
|---|---|
| **What these data cannot show** | A single vector of p-values cannot establish independence or PRDS, identify the true nulls, or validate a two-groups model. |
| **Sensitivity checks** | For Storey's procedure, plot $\widehat{\pi}_0^{\lambda}$ and the rejection set over several prespecified values of $\lambda$. Large changes show that the adaptive conclusion depends strongly on the right-tail cutoff. This comparison is descriptive; it does not justify choosing the value of $\lambda$ that gives the most rejections. |
| **Safer fallback** | If the p-values are valid but their dependence is not justified, use BY. If p-value validity itself is doubtful, neither BH nor BY supplies an FDR guarantee. |

PRDS describes how the whole p-value vector changes when we condition on one null p-value. Pairwise correlations alone generally do not establish it. For the one-sided Gaussian model in Theorem 6.2, however, the stated nonnegative covariances are sufficient. Under PRDS, BH controls FDR. BY permits arbitrary dependence, but its smaller thresholds may reduce the chance of detecting false nulls. The mirror counterexamples show why BH's dependence assumption cannot simply be transferred to another rule.

Storey's finite-sample theorem assumes independent uniform null p-values, independence of the null and nonnull vectors, and a cutoff $\lambda$ chosen in advance. Few alternatives above $\lambda$ can improve the estimate's accuracy, but the theorem does not require their absence. AdaPT also needs a specified null independence condition, as well as the masking, shrinking, and stopping rules in Section 9. A good fitted model does not replace these requirements.

Local FDR gives the probability that one hypothesis is null under a specified mixture model. Averaging the correct posterior probabilities over a selected set gives its posterior expected FDP. Estimated q-values instead summarize when hypotheses enter a specified family of rejection rules; their interpretation depends on how those rules are calibrated. Section 10 gives the corresponding population definition using pFDR. Keeping these quantities separate prevents a model-based probability for one case from being mistaken for a general FDR guarantee.

## 13. Bibliographic Notes

**Route: Core**

The PRDS framework and BH validity under positive dependence are due to Benjamini and Yekutieli [21]; related Simes-type positive-dependence results include Sarkar [158]. Failure of mirror and knockoff+ thresholds under PRDS and other dependent score laws is studied by Zhang [201]. Storey's estimator and q-values are developed in Storey [168], Storey [169], and Storey et al. [170]. The empirical-Bayes local FDR perspective originates with Efron [57] and is treated comprehensively in the book Efron [58]. The compound-decision formulation and oracle/adaptive local FDR thresholding are developed by Sun and Cai [173]. Covariate-assisted local FDR rules and their oracle optimality theory are developed by Zhang and Chen [202] and Cao et al. [39]. Related structured empirical-Bayes approaches, including local-index ideas for

dependent and spatial settings, are represented by Sun et al. [174]. AdaPT is due to Lei and Fithian [117], and the broader masking/interactive FDR framework is developed by Lei et al. [118]. General FDR reshaping and dependence ideas are discussed by Blanchard and Roquain [28]. Conditional calibration and dependence-adjusted BH are represented by Fithian and Lei [62]. A frontier direction with growing practical relevance under regulatory constraints such as the General Data Protection Regulation (GDPR) and the European Union (EU) AI Act is *differentially private* FDR control: the analyst adds calibrated Laplace or Gaussian noise to the log-p-values before selection, trading a small multiplicative inflation of the FDR bound for $(\varepsilon, \delta)$-differential privacy of the released rejection set. Dwork et al. [56] develops the theoretical framework.

## 14. Exercises

**Route: Core**

### Basic.

**Exercise 6.16: Increasing sets**

The text and Figure 6.1 give examples of increasing sets. Using the same p-value coordinate convention, draw two subsets of $[0,1]^2$ that are not increasing. For each set $A$, exhibit $p, p' \in [0,1]^2$ with $p \in A$, $p' \geq p$, and $p' \notin A$.

### Computational.

**Exercise 6.17: Storey under dependence**

Reproduce Figure 6.3, using the empirical average-power measure in Exercise 4.16. Vary $\lambda$, the signal strength, and the correlation. Report Monte Carlo SEs for the empirical FDR, and identify settings where it exceeds $q$ beyond simulation error.

**Exercise 6.18: BH-adjusted p-values and local FDR**

Simulate $m = 10{,}000$ z-scores from the two-groups model

$$Z \sim \pi_0 N(0,1) + (1-\pi_0) f_1, \qquad \pi_0 = 0.9,$$

where

$$f_1(z) = 0.7\,\varphi(z-2.5) + 0.3\,\varphi\{(z+1.5)/2\}/2$$

is an asymmetric Gaussian-mixture density with a wider component centered at $-1.5$. Use one-sided p-values $p = 1 - \Phi(Z)$. Compute BH-adjusted p-values and the oracle local FDR

$$\mathrm{lfdr}(z) = \frac{\pi_0 \varphi(z)}{\pi_0 \varphi(z) + (1-\pi_0) f_1(z)}.$$

Identify specific ranks where the two rankings disagree. For at least two such disagreements, use the simulated z-scores and the shape of $f_1$ to explain the different rankings.

**Exercise 6.19: Two-stage BKY**

Let $q \in (0,1)$ and set $q_1 = q/(1+q)$. The two-stage Benjamini–Krieger–Yekutieli (BKY) procedure of Benjamini et al. [23] first runs BH at level $q_1$, obtaining $R_1$ rejections. If $R_1 = 0$, reject nothing; if $R_1 = m$, reject all $m$ hypotheses. Otherwise run BH again at level

$$q_2 = q_1 \frac{m}{m - R_1},$$

equivalently using critical values $iq_1/(m - R_1)$. Implement this procedure and compare it with BH and the Storey procedure of Theorem 6.9. Use $m = 400$, $q = 0.10$, Storey's $\lambda = 0.5$, and $\pi_0 \in \{1, 0.9, 0.7\}$. Generate one-sided p-values $p_i = 1 - \Phi(Z_i)$ from independent unit-variance Gaussian statistics with mean zero for nulls and 2.5 for nonnulls. Then repeat with equicorrelation $\rho = 0.5$, keeping the marginal distributions fixed. Use at least 2,000 independent runs per setting. Report empirical FDR and, when $\pi_0 < 1$, empirical average power as defined in Exercise 4.16. The cited finite-sample BKY guarantee assumes independent p-values; the correlated simulations assess these particular settings and do not establish arbitrary-dependence control.

**Advanced.**

**Exercise 6.20: Dependence created by aggregate counts**

Proposition 6.13 gives the individual fair-coin symmetry behind AdaPT. Let $P_1, P_2$ be independent $\mathrm{Unif}(0,1)$ null p-values and set $B_i = \mathbf{1}\{P_i \geq 1/2\}$. Conditional on the information $B_1 + B_2 = 1$, compute the joint distribution of $(B_1, B_2)$ and both marginal distributions. Show that the indicators are no longer independent, even though each still has a Bernoulli(1/2) marginal distribution.

CHAPTER 7

# Structured and Hierarchical Multiple Testing

The BH and BY procedures of Chapters 5 and 6 treat the hypotheses as a flat list. They sort the p-values, compare them with a common boundary, and rely on broad assumptions about dependence among those p-values. They do not use the scientific map that says which hypotheses belong together. Real families are rarely flat. Genomic studies test variants inside genes and genes inside pathways. Clinical trials separate primary endpoints from secondary and tertiary endpoints. Benchmark suites for AI models pool prompts inside tasks, tasks inside capability domains, and domains inside an overall evaluation. When this structure is part of the question, ignoring it can waste power and can report discoveries at the wrong resolution.

This chapter develops procedures that use such structure without losing FDR control. We treat four cases: (i) hypotheses grouped into disjoint families with one layer of aggregation (group BH), (ii) several partition layers tested simultaneously (p-filter), (iii) hypotheses arranged on a rooted tree with selected-family FDR targets at each depth (TreeBH), and (iv) replicability across studies, where a discovery requires evidence in enough separate experiments through partial conjunction (PC) testing. A knockoff analogue also exists for structured feature selection; it is discussed after ordinary knockoffs are introduced in Chapter 9.

## 1. Why Flat BH Is Not Enough

**Route: Core**

**A plain-language roadmap.** Choose a method by asking two questions: what structure do the hypotheses have, and what unit will appear in the final scientific report?

**One collection of disjoint groups.:** Use group BH when the data are variants within genes, prompts within tasks, or another single grouping, and the reported discoveries are the *groups*. It controls group-level FDR; it does not by itself provide leaf-level error control within a selected group.

**Several groupings of the same leaves.:** Use the p-filter when every leaf belongs to one group in each of several prespecified partition layers. Groups from different layers may overlap, and the layers need not be nested. Use this method when the report should contain one jointly filtered leaf set together with the groups it touches at each layer. Its guarantee is simultaneous layerwise FDR for those induced group discoveries.

**A rooted hierarchy.:** Use TreeBH when discoveries must form paths down a tree, such as pathway $\to$ gene $\to$ variant, and the report should contain only leaves reached through selected ancestors. Its guarantee is a recursive selected-family FDR along those reported branches, rather than an unconditional FDR for all nodes at a given depth.

**Repeated studies.:** Use PC testing when the structure is across experiments rather than across resolutions, and the reported unit is a feature supported in at least $r$ of $K$ studies. The claim is "replicates often enough," not "is non-null in every study."

The procedures therefore answer different questions even when they start from the same leaf p-values. The rest of the chapter makes each target precise before giving its algorithm.

Suppose the $m$ hypotheses are partitioned into $G$ groups $\mathcal{G}_1, \ldots, \mathcal{G}_G$, where group $g$ contains $m_g$ hypotheses. Call group $g$ a *null group* if every hypothesis in $\mathcal{G}_g$ is a true null, and a *rejected group* if at least one hypothesis in $\mathcal{G}_g$ is rejected. Two natural error rates are

$$\mathrm{FDR}_{\mathrm{flat}} = \mathbb{E}\left[\frac{V}{R \vee 1}\right],$$

where $V$ is the number of false rejections across all hypotheses, and

$$\mathrm{FDR}_{\mathrm{group}} = \mathbb{E}\left[\frac{V^{\mathrm{group}}}{R^{\mathrm{group}} \vee 1}\right],$$

where $V^{\mathrm{group}}$ is the number of *rejected null groups* (i.e., null groups in which the procedure rejected at least one hypothesis) and $R^{\mathrm{group}}$ is the total number of rejected groups. Under this definition, a group that mixes true signal with one falsely rejected null is *not* counted as a false group discovery: the group's null hypothesis "all members are null" is genuinely false, so flagging the group is a true discovery at the group resolution.

These rates can differ dramatically. Consider a stylized genomic example: $G = 50$ genes, each with $m_g = 200$ tightly correlated single-nucleotide variants, so that the variants within a gene effectively measure the same underlying signal. Suppose five genes carry a real signal and BH at the variant level rejects 800 variants concentrated in those five real genes plus 40 variants spread across five null genes (where correlated noise produced small p-values). Then the realized flat FDP is $40/840 \approx 4.8\%$. This is one realized fraction, not the FDR, which averages over repeated data sets. The realized group-level FDP is $V^{\mathrm{group}}/R^{\mathrm{group}} = 5/10 = 50\%$: half the genes flagged are false positives at the gene resolution. A geneticist who is going to follow up at gene resolution needs the group-level guarantee, not the variant-level one.

> **Remark 7.1**
> The flat and group rates are not directly comparable, and control of one does not imply control of the other. Group BH targets group discoveries; the p-filter can control several declared resolutions simultaneously; TreeBH targets averages within selected branches. The error rate must match the unit of discovery that the analyst plans to report.

A second issue is heterogeneity of group sizes and group signal density. A flat BH procedure pools all p-values into a single ordering, so groups with many small p-values dominate the rejection set even when those p-values reflect correlated noise within a single null group. Procedures aware of the group structure can balance rejections across groups or apply group-level filtering before within-group testing. Either approach restores the connection between "what the procedure rejects" and "what the analyst will act on."

## 2. Group BH via Simes Aggregation

**Route: Core**

The simplest group-level procedure aggregates each group into a single super-uniform p-value and applies BH to the group representatives. Given within-group p-values $p_{g,1}, \ldots, p_{g,m_g}$, the Simes representative for group $g$ is

$$p_g^{\mathrm{Simes}} = \min_{1 \le k \le m_g} \frac{m_g\, p_{g,(k)}}{k},$$

where $p_{g,(k)}$ is the $k$th-smallest p-value in group $g$. Group BH then applies the ordinary BH procedure at level $q$ to $p_1^{\mathrm{Simes}}, \ldots, p_G^{\mathrm{Simes}}$ and rejects the corresponding groups. The plain-language idea is simple: a null group should not look significant unless at least one of its ordered p-values

crosses the Simes line. Under independence, and under the usual positive-dependence conditions for Simes, that crossing probability is no larger than the nominal level. Thus the whole group can be represented by one valid group-level p-value.

**Theorem 7.2: Simes super-uniformity inside a group**

Suppose group $g$ is null, meaning every constituent $H_{g,i}$ is true. Assume either that $p_{g,1}, \dots, p_{g,m_g}$ are mutually independent and super-uniform, or more generally that their joint null law satisfies the Simes inequality

$$\mathbb{P}\left( \min_{1 \le k \le m_g} \frac{m_g p_{g,(k)}}{k} \le t \right) \le t, \qquad 0 \le t \le 1.$$

For example, super-uniform p-values satisfying the PRDS condition of Section 1 of Chapter 6 satisfy this inequality: apply Theorem 6.5 under the global null and use the equivalence between a Simes rejection and at least one BH rejection. Then $\mathbb{P}(p_g^{\text{Simes}} \le t) \le t$ for every $t \in [0,1]$.

**Proof of Theorem 7.2**

First take all $m_g$ p-values to be independent and exactly uniform. The classical Simes calculation, proved in Theorem 3.1, states that for independent uniform $U_1, \dots, U_{m_g}$ with order statistics $U_{(1)} \le \cdots \le U_{(m_g)}$,

$$\mathbb{P}\left( \bigcup_{k=1}^{m_g} \{U_{(k)} \le tk/m_g\} \right) = t \qquad \text{for every } t \in [0,1],$$

and hence $\mathbb{P}(p_g^{\text{Simes}} \le t) = t$.

Now relax to independent super-uniform nulls. Let $F_i$ be the CDF of $p_{g,i}$, and write $F_i^{-1}(u) = \inf\{x : F_i(x) \ge u\}$ for its quantile function. Generate independent uniforms $U_1, \dots, U_{m_g}$ and set $p'_{g,i} = F_i^{-1}(U_i)$. These variables have the original independent p-value distributions. Super-uniformity means $F_i(x) \le x$, so $F_i^{-1}(u) \ge u$ for $0 < u < 1$; hence $p'_{g,i} \ge U_i$ almost surely. Coordinate-wise domination implies order-statistic domination, $p'_{g,(k)} \ge U_{(k)}$ for every $k$ path-wise, and since the Simes statistic is non-decreasing in each input, $\min_k m_g p'_{g,(k)}/k \ge \min_k m_g U_{(k)}/k$ path-wise. Because $(p'_{g,1}, \dots, p'_{g,m_g})$ has the same joint distribution as the original independent p-values, this gives $\mathbb{P}(p_g^{\text{Simes}} \le t) \le \mathbb{P}(\min_k m_g U_{(k)}/k \le t) = t$, as claimed. In the general case where the displayed Simes inequality is assumed directly, the conclusion is exactly that inequality rewritten in terms of $p_g^{\text{Simes}}$. The PRDS example is the positive-dependence Simes theorem of Sarkar [158]; see Chapter 6. □

The across-group condition in the next theorem is a condition on the group-level representatives, not merely on the original leaf-level p-values. Two safe situations are worth keeping in mind. If the groups are disjoint and the p-values in different groups are independent, then the Simes representatives are independent. More generally, if the group representatives themselves are one-sided Gaussian p-values whose underlying Gaussian vector has the positive-dependence structure of Theorem 6.2, then they are PRDS. What is *not* automatic is that a PRDS statement for all leaf-level p-values survives the nonlinear operation that maps each group to its Simes minimum.

**Theorem 7.3: Group BH FDR control**

Suppose that

(1) within each true-null group $\mathcal{G}_g$, the p-values satisfy the assumptions of Theorem 7.2, and
(2) the Simes representatives $p_1^{\text{Simes}}, \dots, p_G^{\text{Simes}}$ are mutually independent (or PRDS) across groups.

Then group BH at level $q$ applied to the Simes representatives $p_1^{\text{Simes}}, \dots, p_G^{\text{Simes}}$ controls group FDR at level $q$:

$$\mathbb{E}\left[\frac{V^{\text{group}}}{R^{\text{group}} \vee 1}\right] \leq q.$$

**Proof of Theorem 7.3**

By Theorem 7.2, each $p_g^{\text{Simes}}$ is super-uniform under its group null. Across-group independence (or PRDS) of the representatives is assumed. Applying Corollary 5.3 (or the PRDS variant from Chapter 6) to the family $\{p_g^{\text{Simes}}\}_{g=1}^{G}$ yields $\text{FDR}^{\text{group}} \leq q$. □

Interpretively, group BH differs from flat BH by changing the reported unit. A single strong signal in a group of size $m_g$ can produce many small correlated within-group p-values, inflating the variant-level discovery count even when the scientific signal lives at the gene level. Group BH collapses these into one calibrated representative, so the output is a group-level list rather than a long list of correlated within-group hits. If within-group dependence does not satisfy the Simes assumptions, Bonferroni aggregation or a BY-reshaped Simes value gives a valid group p-value. One must still check the across-group independence or PRDS assumption before invoking Theorem 7.3.

**Example 7.4: Two genes by hand**

Let $G = 2$ groups have within-group p-values

$$(p_{1,1}, p_{1,2}, p_{1,3}) = (0.028, 0.028, 0.028), \qquad (p_{2,1}, p_{2,2}, p_{2,3}) = (0.031, 0.40, 0.50).$$

The Simes representatives are

$$p_1^{\text{Simes}} = \min\left(\frac{3 \cdot 0.028}{1}, \frac{3 \cdot 0.028}{2}, \frac{3 \cdot 0.028}{3}\right) = 0.028,$$

$$p_2^{\text{Simes}} = \min\left(\frac{3 \cdot 0.031}{1}, \frac{3 \cdot 0.40}{2}, \frac{3 \cdot 0.50}{3}\right) = 0.093.$$

Group BH at $q = 0.05$ compares the sorted representatives $(0.028, 0.093)$ with the thresholds $qk/G = 0.025k$ for $k = 1, 2$, namely $(0.025, 0.05)$. Because $0.028 > 0.025$ and $0.093 > 0.05$, group BH rejects no group.

Compare with flat BH on the six pooled p-values, sorted as

$$0.028,\ 0.028,\ 0.028,\ 0.031,\ 0.40,\ 0.50,$$

against thresholds $qk/6 = 0.00833k$ for $k = 1, \dots, 6$,

$$(0.0083,\ 0.0167,\ 0.025,\ 0.0333,\ 0.0417,\ 0.05).$$

The step-up rule of BH rejects the largest $k$ for which $p_{(k)} \le qk/m$. Here $p_{(4)} = 0.031 \le 0.0333$, so flat BH rejects all four of the smallest p-values: the three from gene 1 *and* the first p-value of gene 2. Group BH does not: it has refused both groups, including the one for which flat BH would have rejected a (likely null) variant. The price of group-level interpretability is this conservatism: group BH refuses to rebrand a single gene's correlated evidence as multiple within-gene discoveries, and refuses to leak a variant-level rejection into a gene that fails its group-level test.

## 3. The p-Filter

**Route: Advanced**

Group BH controls FDR at a single layer at a time. Multi-resolution problems require simultaneous FDR control at several layers. The *p-filter* of Barber and Ramdas [10] solves this.

**Setup.** Let the hypotheses $\{1, \dots, m\}$ carry $L$ partition layers; layer $\ell$ is described by a many-to-one map $\pi_\ell : \{1, \dots, m\} \to \{1, \dots, G_\ell\}$ that sends each hypothesis to its group at layer $\ell$. Layer $\ell = 1$ is typically the finest (often $G_1 = m$ with one hypothesis per "group"); deeper layers are coarser. The layers need not be nested; the p-filter applies to multiple partitions of the hypotheses, with nested hierarchies as an important special case. Call a layer-$\ell$ group $g$ a *null group* if all of its constituent hypotheses are true nulls. For a rejection set $R \subseteq \{1, \dots, m\}$, define

$$R^{(\ell)} = |\pi_\ell(R)|, \qquad V^{(\ell)} = |\{g \in \pi_\ell(R) : g \text{ is a null group}\}|,$$

that is, the number of layer-$\ell$ groups containing at least one rejection, and the number of those that are null groups. The p-filter takes layer levels $q_1, \dots, q_L$ and returns a rejection set $\widehat{R}$ such that $\mathrm{FDR}^{(\ell)} = \mathbb{E}[V^{(\ell)}/(R^{(\ell)} \vee 1)] \le q_\ell$ simultaneously for every $\ell$.

**The iterative algorithm.** For each layer $\ell$ and each group $g$ at that layer, let $p_g^{(\ell)}$ be the Simes p-value computed directly from the original p-values of the hypotheses in that group (at $\ell = 1$ with $G_1 = m$, one has $p_h^{(1)} = p_h$ for each hypothesis $h$). Given a threshold vector $\boldsymbol{t} = (t_1, \dots, t_L)$, the joint hypothesis-level rejection set and its layer-$\ell$ image are

$$\widehat{R}(\boldsymbol{t}) = \{i : p^{(\ell)}_{\pi_\ell(i)} \le t_\ell \text{ for every } \ell\}, \qquad \widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}) = \pi_\ell(\widehat{R}(\boldsymbol{t})). \tag{2}$$

The p-filter of Barber and Ramdas [10] can be described through count-space coordinates. For a proposed vector $\boldsymbol{k} = (k_1, \dots, k_L)$, set

$$t_\ell(k_\ell) = \frac{q_\ell k_\ell}{G_\ell}, \qquad k_\ell \in \{0, 1, \dots, G_\ell\}.$$

The algorithm chooses the coordinate-wise largest feasible $\boldsymbol{k}$ such that

$$\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}(\boldsymbol{k})) \ge k_\ell \qquad \text{for every layer } \ell.$$

At the returned threshold vector $\boldsymbol{t}$, the BH-style boundary

$$\frac{G_\ell\, t_\ell}{\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}) \vee 1} \;\le\; q_\ell \tag{3}$$

holds at every layer $\ell$. The count formulation also sets $k_\ell = 0$ when no group survives at that layer. This fixes the otherwise irrelevant choice of threshold when the rejection set is empty.

**Algorithm 7.1: p-filter coordinate reduction**

1. Initialize $k_\ell = G_\ell$, equivalently $t_\ell = q_\ell$, for every layer.
2. Given the current $\boldsymbol{t}$, compute $\widehat{R}(\boldsymbol{t})$ and the layer counts $r_\ell = \#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t})$.
3. Replace $k_\ell$ by $\min(k_\ell, r_\ell)$ and $t_\ell$ by $q_\ell k_\ell / G_\ell$ for every layer, then repeat until $\boldsymbol{k}$ stabilizes.

The iteration is coordinate-wise monotone: each $k_\ell$ can only decrease, and each coordinate takes values in the finite set $\{0, 1, \ldots, G_\ell\}$. Hence the reduction iteration terminates in a finite number of steps at a self-consistent fixed point; convergence and uniqueness of the maximal feasible point are established in Barber and Ramdas [10].

**Validity.** The theorem below is the Simes version of Barber and Ramdas [10]. Its dependence assumption concerns the full vector of original, or *base*, p-values. Group p-values from different layers reuse these inputs and need not be independent. The proof uses their particular Simes construction; replacing them by arbitrary valid group p-values is not justified by this theorem.

**Theorem 7.5: p-filter FDR control, cited result [10]**

Fix the partition layers before observing the p-values. Suppose the base p-values are super-uniform for true nulls and their full vector satisfies PRDS on the true-null coordinates. Mutual independence of all base p-values is a sufficient special case. With Simes group p-values as defined above, the p-filter output at prespecified levels $q_1, \ldots, q_L \in (0, 1)$ satisfies $\mathrm{FDR}^{(\ell)} \le q_\ell$ for every $\ell$.

Proof idea. With one layer, the maximal feasible count is the number of BH rejections among the group p-values. Under independent base p-values, disjoint groups have independent Simes representatives, so Theorem 7.3 applies. With several layers, a group's survival also depends on thresholds at the other layers. Barber and Ramdas [10] handle this dependence by expressing a null group's contribution in terms of its base p-values and using PRDS to bound those contributions at the data-dependent thresholds. This yields a bound of $q_\ell$ at each layer. The same argument covers the single-layer PRDS case; the paper gives the full proof.

In the special case $L = 1$ with $G_1 = m$ (the trivial layering at the hypothesis level), the boundary (3) reduces to BH at level $q_1$ applied to $p_1, \ldots, p_m$. In the special case $L = 1$ with coarser groups and Simes aggregation, the layer-1 p-values are $p_g^{(1)} = p_g^{\mathrm{Simes}}$ and the p-filter reduces to the group BH of Section 2. The new content of the p-filter is that the same iteration controls FDR at multiple resolutions simultaneously without inflating the per-layer levels.

## 4. Trees of Hypotheses and TreeBH

**Route: Advanced**

We next consider a prespecified finite rooted tree $\mathcal{T}$. Let the leaves of $\mathcal{T}$ be the testable hypotheses $\{H_1, \ldots, H_m\}$ and let each internal node $v$ carry the intersection hypothesis

$$H_v = \bigcap_{i:\, i \preceq v} H_i,$$

where $i \preceq v$ means leaf $i$ is a descendant of $v$. A rejection set $R$ is *ancestor-closed* if, whenever a node is rejected, all of its tested ancestors are also rejected. Ancestor-closure is the natural constraint for hierarchical procedures because it makes scientific sense: one cannot claim a

gene-level discovery without also reporting the containing pathway as a discovery. At the level of hypotheses, a false gene null implies a false pathway null. The testing procedure respects this logical relation, although a rejection can of course be mistaken.

**Algorithm.** The TreeBH procedure of Bogomolov et al. [30] implements a top-down rejection rule. Let $C(v)$ denote the children of node $v$, and write $\rho$ for the formal root. We index depth from $d = 1$ at the topmost *tested* layer (the children of $\rho$) down to $d = D$ at the leaves; $\rho$ itself is only a container, not a tested family, so it does not consume a target FDR level. This is the opposite direction from the p-filter notation above: here depth 1 is the coarsest tested layer and depth $D$ is the leaf layer. Each tested depth carries its own target FDR level $q_d$ for $d = 1, \dots, D$.

For the version developed here, every leaf is at depth $D$, while the number of children may differ from node to node. This permits unequal group sizes without introducing a separate convention for branches that terminate at different depths.

Leaves carry their original p-values. Compute each internal-node p-value by applying Simes to the p-values of its immediate children, working upward from the leaves. Under the independent-leaf assumptions stated below, child p-values within a family depend on disjoint leaf sets and are independent. Theorem 7.2 therefore supplies valid internal-node p-values at each successive depth. Recursive Simes aggregation can differ from applying Simes directly to all descendant leaves; throughout this TreeBH construction we use the recursive rule.

For a selected parent $v$, write $\widehat{\mathcal{S}}(v) \subseteq C(v)$ for the children rejected by BH within that child family. The formal root $\rho$ is selected by convention, so its children form the first tested family.

**Algorithm 7.2: TreeBH top-down procedure**

1. At depth 1, apply BH to the p-values of the children of $\rho$ at level $q_1$; call the rejected children $\widehat{\mathcal{S}}(\rho)$.
2. At depth $d+1$, for every selected node $v$ at depth $d$, apply BH to the p-values of $C(v)$ at the path-specific level
$$q_v^{(d+1)} = q_{d+1} \prod_{r=1}^{d} \frac{|\widehat{\mathcal{S}}(v_{r-1})|}{|C(v_{r-1})|},$$
where $v_0 = \rho, v_1, \dots, v_d = v$ is the ancestor chain of $v$. Reject the children selected by that BH run and call the set $\widehat{\mathcal{S}}(v)$.
3. Recurse on rejected children until no further rejections occur or the leaves are reached.

The product of selected proportions along the ancestor chain is the key design choice. It reduces the level used within a family to account for selection of its ancestors using the same data. For a two-level hierarchy this product reduces to the selected-family factor $R^{(1)}/G_1$; in deeper trees it is path-specific, not a single global ratio at the previous depth.

**Validity.** For a two-level hierarchy, the selected-family target is the average FDP among the selected parent families. For deeper trees, TreeBH controls the recursive selected-family FDR, denoted $s\,\mathrm{FDR}^{\ell}$, rather than a flat average over all selected parents at the immediately previous depth.

Fix a target level $\ell$. For a selected node $u$ at depth $\ell$, let
$$A_u^{(\ell)} = \mathbf{1}\{H_u \text{ is a true null}\}.$$

For a selected node $v$ at depth $k < \ell$, define recursively

$$A_v^{(\ell)} = \frac{1}{|\widehat{\mathcal{S}}(v)| \vee 1} \sum_{u \in \widehat{\mathcal{S}}(v)} A_u^{(\ell)}.$$

Finally set

$$s\,\mathrm{FDR}^{\ell} \;=\; \mathbb{E}[A_\rho^{(\ell)}]. \tag{4}$$

The quantity $A_v^{(\ell)}$ is the false-discovery fraction passed upward from the selected descendants of $v$ at target depth $\ell$. At the formal root $\rho$, it is the recursively averaged FDP for the whole selected tree at depth $\ell$. At $\ell = 2$, this reduces to the usual average of within-family FDPs over selected depth-1 families. At $\ell = 3$, it first averages over selected children within each selected depth-1 family, then averages those quantities over selected depth-1 families. This recursive weighting differs from pooling all selected depth-2 parents together when selected branches have different numbers of selected children.

For example, suppose that two selected depth-1 nodes, $A$ and $B$, have respectively three and two selected children at depth 2. If one selected child under each node is a true null, then

$$A_A^{(2)} = \frac{1}{3}, \qquad A_B^{(2)} = \frac{1}{2}, \qquad A_\rho^{(2)} = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{2}\right) = \frac{5}{12}.$$

The pooled depth-2 FDP would instead be $2/5$. Recursive averaging gives the two selected parent families equal weight, whereas pooling gives more weight to $A$ simply because it has more selected children.

The following sufficient case of Bogomolov et al. [30, Corollary 1] matches the algorithm above. It assumes independence only at the leaves; parent and descendant p-values remain dependent because the former are computed from the latter.

**Theorem 7.6: TreeBH recursive selected-family FDR control, cited result**

Fix the tree and levels $q_1, \ldots, q_D \in (0, 1)$ in advance. Suppose all leaf p-values are mutually independent and every true-null leaf p-value is super-uniform. Let each internal hypothesis be the intersection of its children and form its p-value recursively by Simes. Apply BH to every selected parent's complete child family at the path-specific level displayed above. Then the TreeBH output satisfies

$$s\,\mathrm{FDR}^{\ell} \le q_\ell, \qquad \ell = 1, \ldots, D.$$

The cited result uses the fact that BH is a *simple selection rule*: within a family tested at a fixed level, varying one rejected p-value while it remains rejected does not change the number of rejections. Thus no extra selection condition needs to be checked for the BH steps specified here. The theorem permits different $q_d$'s in any order. If $q_1 \le \cdots \le q_D$, recursive Simes additionally makes the procedure *consonant*: every rejected internal node has at least one rejected child. Indeed, rejection of its Simes p-value means that some child crosses a BH boundary at the parent's rejection threshold, and the child's family is tested at a level at least that large. This is a coherence property, not an extra requirement for the independent-leaf FDR bound.

For dependent leaves, Bogomolov et al. [30, Theorem 1] also give simultaneous control under PRDS when all target levels are equal and internal p-values are formed by recursive Simes. Nondecreasing but unequal levels alone do not give that extension.

Why the rescaling has this form. Two levels suffice to see the idea: a root whose children are groups, whose own children are leaves. At depth $d = 1$, BH applied to the group-level Simes representatives at level $q_1$ gives $\mathrm{FDR}^{(1)} \le q_1$ by Theorems 7.2–7.3.

For depth $d + 1 = 2$, the standard mistake is to condition on "group $g$ was selected at depth 1": the selection event $\{p_g^{\mathrm{Simes}} \le q_1\, R^{(1)}/G_1\}$ is a function of the very leaf p-values inside $g$, so within-$g$ conditional super-uniformity is not available. The $q_1$ in this event is correct: it is the level used to select top-level groups. The later BH run inside each selected group uses the depth-2 level $q_2 R^{(1)}/G_1$. Instead we work unconditionally, noting that the within-$g$ FDP contributes to $s\,\mathrm{FDR}^2$ only when $g$ is selected:

$$s\,\mathrm{FDR}^2 \;=\; \mathbb{E}\left[\frac{1}{R^{(1)} \vee 1} \sum_g \mathbf{1}\left\{g \in \widehat{\mathcal{R}}^{(1)}\right\} \cdot \frac{V_g^{(2)}}{R_g^{(2)} \vee 1}\right].$$

The selected-family theorem of Bogomolov et al. [30] – of which TreeBH is the prototypical instance – shows that testing each selected family at the rescaled level $q_2 R^{(1)}/G_1$ controls the selected-family average in (4) at $q_2$. The factor $R^{(1)}/G_1$ is what compensates for choosing the family by looking at its own data: only a fraction $R^{(1)}/G_1$ of top-level families are selected, so the within-family level is reduced by that selected fraction. In deeper trees the same adjustment is applied recursively along the realized ancestor chain. A selected gene inside a selected pathway therefore inherits the pathway-level selected fraction and the selected-gene fraction inside that pathway; replacing this product by a single global previous-depth ratio is not the TreeBH rule.

**Remark 7.7**

TreeBH's per-level target (4) is a recursive selected-family FDR, not the unconditional "overall" FDR $\mathbb{E}[V^{(d+1)}/(R^{(d+1)} \vee 1)]$. Bounding the overall FDR at depth $d + 1$ directly would require simultaneous control across all rejected ancestors at depth $d$, and a naive per-rejected-parent BH does not deliver this without additional adjustment. "BH inside each rejected family" is the right tool for the recursive selected-family target, which is the natural multi-level error rate in genomics, neuroimaging, and benchmark hierarchies.

**Example 7.8: Three-level genomic hierarchy**

Suppose 20,000 single-variant p-values are nested inside 800 gene-level hypotheses, which are nested inside 50 pathway-level hypotheses, with target levels $(q_1, q_2, q_3) = (0.05, 0.10, 0.10)$ for pathway, gene, and variant FDR. At depth 1, BH is applied to the 50 pathway-level Simes representatives at level $q_1 = 0.05$; suppose this rejects $R^{(1)} = 6$ pathways. At depth 2, within each rejected pathway, BH is applied to the gene-level Simes representatives at the local testing level

$$q_{2,\mathrm{loc}} = q_2 \frac{R^{(1)}}{G_1} = 0.10 \cdot \frac{6}{50} = 0.012.$$

Here $q_2 = 0.10$ remains the global depth-2 selected-family FDR target, whereas $q_{2,\mathrm{loc}} = 0.012$ is the nominal level used by each BH run inside a selected pathway. The two quantities play different roles and should not share a symbol. Suppose this rejects genes inside those pathways. If a selected pathway $P$ has $R_{\mathrm{genes}}(P)$ rejected genes among $G_{\mathrm{genes}}(P)$ tested genes, then at depth 3, within each rejected gene in pathway $P$, BH is applied to the variant

p-values at the path-specific level

$$q_{3,P} = q_3 \cdot \frac{6}{50} \cdot \frac{R_{\text{genes}}(P)}{G_{\text{genes}}(P)}.$$

The variant threshold is therefore pathway-specific. The global ratio $0.10 \cdot R_{\text{genes,total}}/800$, for example $0.10 \cdot 80/800$ if 80 genes are rejected in total across all pathways, is not the TreeBH formula because it omits the top-level selected-pathway factor and replaces the selected-parent-specific gene ratio by a pooled count. TreeBH controls the recursive selected-family FDR $s\,\text{FDR}^{\ell}$ at each depth, not the ordinary depth-level FDR $\mathbb{E}[V^{(\ell)}/(R^{(\ell)} \vee 1)]$.

**The same pathway–gene–variant hierarchy, used in two ways**

**p-filter: joint, fine-to-coarse filtering**

Start with all variants and solve the thresholds at every requested layer together:

all variants

$\Downarrow$

$p_{\text{variant}} \le t_{\text{variant}},$
$p^{\text{Simes}}_{\text{gene}} \le t_{\text{gene}},$
$p^{\text{Simes}}_{\text{pathway}} \le t_{\text{pathway}}$

$\Downarrow$

one leaf set

A variant survives only if all of its layer constraints pass. The reported genes and pathways are the groups touched by that single jointly filtered variant set. The target is simultaneous FDR at every requested layer.

**TreeBH: coarse-to-fine branch selection**

Screen the tree from its coarsest tested layer and descend only along selected branches:

pathways

$\Downarrow$

genes within selected pathways

$\Downarrow$

variants within selected genes

Each within-branch BH run uses its path-specific local level. The report is an ancestor-closed set of selected pathways, genes, and variants. The target is recursive selected-family FDR along those selected branches.

The same original variant p-values can be inputs to both procedures, but the aggregation rules, error targets, thresholds, and resulting rejection sets generally differ.

The trade-off with flat BH is power for interpretability. TreeBH may gain or lose leaf-level power relative to flat BH, depending on how signal is distributed across branches and how strongly the parent levels screen the search. Its main advantage is interpretability: every rejected leaf is accompanied by rejected ancestors at each higher level, so a rejected variant lies in a rejected gene in a rejected pathway, and the chain of rejections is itself the report.

Structured versions of the knockoff filter address a related multi-resolution feature-selection problem using knockoff statistics rather than p-values. We defer that topic to Chapter 9, Section 10, after the ordinary knockoff construction has been introduced.

## 5. Replicability and Partial Conjunction

**Route: Advanced**

The discussion so far concerns multi-resolution error rates within a single study. A different structural problem appears when the analyst wants to report only findings that replicate across two or more separate studies.

**Definition 7.9: Replicability**

Let $H_{i,k}$, $k = 1, \dots, K$, be the null hypothesis for feature $i$ in study $k$. Feature $i$ is *$r$-out-of-$K$ replicated* if at least $r$ of the per-study nulls $H_{i,1}, \dots, H_{i,K}$ are false.

The *PC null* negates replicability. Writing $\mathcal{N}_i$ for the set of true-null per-study nulls of feature $i$,

$$H_i^{(r/K)} : \quad |\mathcal{N}_i| \geq K - r + 1,$$

that is, at most $r - 1$ of the per-study nulls $H_{i,1}, \dots, H_{i,K}$ are false. A finding is replicated if and only if this null is false.

The naive "$r$-th smallest p-value" statistic is not super-uniform. For instance, in the any-of-two case $r = 1$, $K = 2$ (one wants at least one of the two studies to detect a signal), the minimum of two independent uniforms has CDF $2t - t^2 > t$, so $p_{(1)}$ cannot serve as a global-null p-value for $H^{(1/2)}$ without correction. This is the global-null / any-of-two regime, not the two-study full-replication regime more commonly meant by "replication," which corresponds to $r = K = 2$ and is addressed separately below. The construction that handles general $r$ is due to Bogomolov and Heller [29] (with foundational antecedents in Benjamini and Heller's earlier work on PC testing) and uses a Bonferroni-type correction applied to the $r$-th smallest p-value.

**Definition 7.10: Partial conjunction p-value**

Let $p_{i,1}, \dots, p_{i,K}$ be per-study p-values for feature $i$ and let $p_{i,(1)} \leq \cdots \leq p_{i,(K)}$ be their order statistics. The $r$-out-of-$K$ Bonferroni-style PC p-value for feature $i$ is

$$p_i^{(r/K)} = (K - r + 1)\, p_{i,(r)} \wedge 1.$$

The intuition is that under $H_i^{(r/K)}$ at least $K - r + 1$ of the p-values are super-uniform nulls. When the procedure looks at the $r$-th smallest p-value, it is looking at the smallest among at least $K - r + 1$ null p-values (in the worst case where the $r - 1$ non-null p-values occupy the $r - 1$ smallest slots). Applying Bonferroni to the $K - r + 1$ nulls and using the smallest of them gives a multiplier of $K - r + 1$. Two boundary cases anchor the formula:

- $r = 1$ (any-of-$K$): $p_i^{(1/K)} = K p_{i,(1)}$, the Bonferroni global-null p-value applied to all $K$ studies. The PC null is "all studies are null," and the p-value is super-uniform by the union bound.
- $r = K$ (all-of-$K$): $p_i^{(K/K)} = p_{i,(K)}$, the largest p-value. The PC null is "at least one study is null," and the maximum p-value is at least any one fixed true-null p-value, so it is super-uniform.

**Theorem 7.11: Super-uniformity of the Bonferroni PC p-value**

Suppose that, under the PC null $H_i^{(r/K)}$, each true-null per-study p-value is super-uniform. No independence assumption on the joint distribution of the per-study p-values is required. Then

$$\mathbb{P}\Big(p_i^{(r/K)} \le t\Big) \le t, \qquad t \in [0,1].$$

**Proof of Theorem 7.11**

The cleanest case is $K = 2$, $r = 2$: $H_i^{(2/2)}$ says at least one of the two nulls is true, and the PC p-value is $p_i^{(2/2)} = p_{i,(2)} = \max(p_{i,1}, p_{i,2})$. If $p_{i,1}$ is the super-uniform null, then $p_{i,(2)} \ge p_{i,1}$, so $\mathbb{P}(p_{i,(2)} \le t) \le \mathbb{P}(p_{i,1} \le t) \le t$. This inheritance is unconditional on the joint distribution: the inequality $\max(p_1, p_2) \ge p_1$ holds path-wise.

The general case has the same flavor. Under $H_i^{(r/K)}$, at least $K - r + 1$ of the per-study p-values are super-uniform nulls. Choose a subset $\mathcal{N}_i^\star \subseteq \mathcal{N}_i$ of true-null study indices with size $K - r + 1$. For any subset $S$ of size $K - r + 1$, the path-wise inequality

$$p_{i,(r)} \ge \min_{k \in S} p_{i,k}$$

holds: the smallest element of $S$ can be beaten in rank by at most the $r - 1$ elements outside $S$, so its rank among all $K$ p-values is at most $r$. Applying this to $S = \mathcal{N}_i^\star$, the union bound over those $K - r + 1$ super-uniform nulls gives $\mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \le s) \le (K - r + 1)\, s$ without any independence assumption, because the union bound is dependence-free. For $0 \le t < 1$, therefore, $\mathbb{P}(p_i^{(r/K)} \le t) \le \mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \le t/(K - r + 1)) \le t$. At $t = 1$ the claim is immediate, since the p-value is capped at one. □

**Remark 7.12: Bonferroni PC vs. Simes/Fisher/Stouffer PC**

The proof above gives the Bonferroni-style PC p-value of Definition 7.10 and is dependence-robust: it remains valid under *arbitrary* joint dependence among the per-study p-values $p_{i,1}, \ldots, p_{i,K}$ for a given feature $i$, because the union bound does not require independence. Several other PC p-values trade dependence-robustness for sharper rejection thresholds:

- The *Simes-style PC p-value*

$$p_i^{(r/K),\text{Simes}} = \left[\min_{r \le j \le K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1}\right] \wedge 1$$

  sharpens Bonferroni under conditions ensuring the Simes inequality for the relevant true-null per-study p-values, such as independence or suitable PRDS [158]; it is well-defined and super-uniform under those conditions, and uniformly no larger than the Bonferroni PC p-value at every realization. The cap at 1 is harmless and kept for parallelism with the Bonferroni formula, although the minimum is already at most one because the $j = K$ term equals $p_{i,(K)}$. Outside the PRDS class, Simes can be anti-conservative, depending on the specific dependence structure.
- Combination-based global-null p-values (Fisher's $-2\sum_k \log p_k$, Stouffer's combined z-score) are typically used in the *global-null* regime $r = 1$, where they combine all $K$ per-study p-values into a single p-value for the global null "all studies are null."

Their usual calibration relies on independence (or on specific, justified dependence-aware variance/covariance calibration); positive dependence can inflate the variance of the combined statistic relative to the independent-study calibration and can make Fisher's and Stouffer's p-values anti-conservative. For example, with two perfectly dependent global-null p-values $p_1 = p_2 = U$, the Stouffer statistic uses $\sqrt{2}\,\Phi^{-1}(1-U)$, so its nominal 0.05 rejection probability is about 0.122. Fisher's independent $\chi^2_4$ calibration similarly rejects with probability $e^{-9.488/4} \approx 0.093$ under $p_1 = p_2 = U$.

This is the relevant distinction for replication across studies that share samples, share calibration sets, or otherwise induce dependence among the per-study p-values; see Heller and Bogomolov [81] for a unified treatment.

**Remark 7.13**

For $r = K = 2$, Definition 7.10 gives $p_i^{(2/2)} = p_{i,(2)}$, the larger of the two per-study p-values. This recovers the elementary "max p-value" rule for two-study full replication, a simple and common operational choice.

**Theorem 7.14: FDR control on replicated findings**

Let $\mathcal{R}^{(r/K)}$ be the set of features for which $H_i^{(r/K)}$ is false (i.e., the $r$-out-of-$K$ replicated features). Assume that for each feature $i$ with $H_i^{(r/K)}$ true, each true-null per-study p-value $p_{i,k}$ is super-uniform – no independence across studies is required for the Bonferroni PC p-value – and that the PC p-values $\{p_i^{(r/K)}\}_{i=1}^m$ are independent (or PRDS) across features. Then applying BH at level $q$ to $\{p_i^{(r/K)}\}_{i=1}^m$ controls FDR on the discovery set with respect to the target $\mathcal{R}^{(r/K)}$:

$$\mathbb{E}\left[\frac{|\widehat{R} \setminus \mathcal{R}^{(r/K)}|}{|\widehat{R}| \vee 1}\right] \leq q.$$

**Proof of Theorem 7.14**

By Theorem 7.11, each PC p-value $p_i^{(r/K)}$ is super-uniform under its PC null, with no cross-study independence needed. Across-feature independence (or PRDS) is assumed. Corollary 5.3 (or the PRDS variant from Chapter 6) now applies directly to the family $\{p_i^{(r/K)}\}_{i=1}^m$. □

**Example 7.15: Two studies by hand**

Two studies report p-values for four features:

$$(p_{1,1}, \ldots, p_{4,1}) = (0.001,\ 0.002,\ 0.03,\ 0.45), \qquad (p_{1,2}, \ldots, p_{4,2}) = (0.0005,\ 0.5,\ 0.02,\ 0.03).$$

The 2-out-of-2 PC p-values are the maxima per feature:

$$(p_1^{(2/2)}, p_2^{(2/2)}, p_3^{(2/2)}, p_4^{(2/2)}) = (0.001,\ 0.5,\ 0.03,\ 0.45).$$

Sorted, the PC p-values are $0.001, 0.03, 0.45, 0.5$, to be compared with the BH thresholds $0.05 \cdot k/4$ for $k = 1, 2, 3, 4$, namely $0.0125, 0.025, 0.0375, 0.05$. Only $0.001 \leq 0.0125$ crosses (since $0.03 > 0.025$, $0.45 > 0.0375$, and $0.5 > 0.05$), so feature 1 is the unique replicated discovery.
For contrast, BH applied separately within each study at $q = 0.05$ rejects features $\{1, 2, 3\}$ in study 1 (since the sorted study-1 p-values $0.001, 0.002, 0.03, 0.45$ compared with thresholds $0.0125, 0.025, 0.0375, 0.05$ yield rejections at $k = 1, 2, 3$) and features $\{1, 3, 4\}$ in study 2 (the sorted study-2 p-values $0.0005, 0.02, 0.03, 0.5$ compared with the same thresholds yield rejections at $k = 1, 2, 3$, which correspond to features $1, 3, 4$ by original index). The naive intersection of single-study rejections is $\{1, 3\}$; the PC procedure returns only $\{1\}$. The contrast is not that PC is "more conservative" in a loose sense; rather, the two procedures target different nulls. The naive intersection is not calibrated for the PC null. A feature can be truly associated in one study and null in another, yet appear in both single-study rejection lists because of a false positive in the null study. The PC procedure instead tests the explicit $r$-out-of-$K$ replication target. Intersecting separately valid discovery lists is a different operation and need not be calibrated for the replication FDP. The max of feature 3's p-values, $\max(0.03, 0.02) = 0.03$, exceeds the joint BH threshold of 0.025, so feature 3 is not declared replicated even though both studies individually rejected feature 3.

## 6. Connections to Regulatory Practice

**Route: Core**

The closure principle of Chapter 4 gives strong FWER control on hierarchical families through graphical procedures. TreeBH gives FDR control on the same families, but for a different inferential target. Confirmatory regulatory clinical trials primarily use FWER-controlling gatekeeping, because the trial is asking whether *any* primary or secondary endpoint shows a true effect with a small tolerance for false positives over the entire endpoint hierarchy. Genomic, neuroimaging, and benchmark-evaluation pipelines typically prefer FDR-controlling hierarchical procedures, because the goal is to identify many promising findings with a controlled FDP at each resolution. The choice between FWER and FDR hierarchical procedures is not a methodological dispute; it is a difference in inferential target.

The relevant regulatory documents include International Council for Harmonisation [95] (ICH E9 addendum on estimands), the EMA *Multiplicity Issues in Clinical Trials* guideline [59], and the FDA *Multiple Endpoints in Clinical Trials* guidance [179]. These documents specify the structural language (primary/secondary/tertiary endpoints, intersection-union and union-intersection tests, gatekeeping) used to organize multi-endpoint confirmatory testing. Their preferred mechanism is FWER-controlling gatekeeping, not FDR-controlling hierarchical procedures: hierarchical FDR procedures of the kind developed in this chapter are appropriate for exploratory and biomarker-discovery contexts, where the analyst trades a tight per-family error rate for a tunable average FDP across many findings. Treating these documents as endorsements of hierarchical FDR would overstate their scope.

## 7. Assumptions in Plain Language

**Route: Core**

Group BH assumes that the within-null-group p-values make the chosen aggregator super-uniform, and that the resulting group representatives are independent or PRDS across groups; PRDS of the original leaf p-values alone does not automatically imply PRDS of the Simes

representatives. The p-filter theorem uses PRDS of the base p-value vector together with direct Simes aggregation within each group, allowing the cross-layer dependencies created by shared inputs.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Before outcomes are examined, define the reporting unit and the groups. Verify that every group or node p-value was computed by the stated Simes or Bonferroni formula. |
| **What these data cannot show** | The p-values cannot show that the grouping is scientifically appropriate, that the dependence conditions for Simes and BH hold, or that a group-level FDR guarantee also holds for individual hypotheses. |
| **Sensitivity checks** | Choose scientifically plausible groupings before examining outcomes, repeat the analysis for each, and report how the rejection lists change. If Simes validity is uncertain, compare Simes with Bonferroni group p-values. |
| **Safer fallback** | Under unrestricted within-group dependence, use Bonferroni group p-values. At one reporting layer, apply BY to those valid group p-values for arbitrary dependence, or Holm for strong group FWER. No automatic substitute preserves every p-filter or TreeBH guarantee. |

The TreeBH theorem stated here assumes independent leaf p-values and recursive Simes aggregation. These assumptions make internal-node p-values valid and supply the conditions for BH with the ancestor-selection adjustments. Its target is the recursive selected-family FDR. The cited PRDS extension uses equal target levels at all depths. Replacing an aggregator by a dependence-robust one does not, by itself, verify the selection and dependence conditions of a hierarchical FDR theorem.

PC tests in the Bonferroni version of Definition 7.10 require only that each true-null per-study p-value is super-uniform; the Bonferroni PC p-value is dependence-robust across studies, by Theorem 7.11. The Simes-style PC p-value is sharper than Bonferroni under conditions ensuring the Simes inequality for the relevant true-null per-study p-values, such as independence or suitable PRDS [158], but loses validity under dependence outside that class. Combination-based global-null p-values (Fisher, Stouffer), typically used in the global-null regime $r = 1$, assume independence across studies in their usual calibration; positive dependence among the combined per-study p-values can inflate the variance of the combined statistic relative to the independent-study calibration and can make these p-values anti-conservative. In replication analyses across genomic consortia or AI benchmarks where study-level p-values may be correlated through shared samples or calibration sets, Bonferroni PC is the safer default; Simes PC is acceptable when PRDS is plausible; Fisher or Stouffer should be used with their usual independent-study calibration only when across-study independence is verified.

## 8. Bibliographic Notes

**Route: Core**

Hierarchical FDR control was introduced by Yekutieli [197]. The modern tree-structured procedure with multi-resolution FDR control is from Bogomolov et al. [30]. The p-filter and its multilayer generalization are due to Barber and Ramdas [10]. Adaptive group-aware procedures

for partially ordered side information are from Li and Barber [120]. The original PC framework is Benjamini and Heller [19]; the replicability extension to multiple features and two-study FDR control is developed by Bogomolov and Heller [29] and Heller and Bogomolov [81]. The connection to regulatory practice in clinical trials is documented in International Council for Harmonisation [95], European Medicines Agency [59], and U.S. Food and Drug Administration [179].

## 9. Exercises

**Route: Core**

### Basic.

**Exercise 7.16: Tree-respecting rejection**

Consider a tree whose untested formal root has two tested children $A$ and $B$, where $A$ has leaves $\{1, 2\}$ and $B$ has leaves $\{3, 4\}$. Suppose a tree-aware procedure rejects $A$ but not $B$, while flat BH on the four leaves rejects $\{1, 3\}$. Under the rule that a leaf may be reported only if every tested ancestor has been rejected, which flat-BH rejections satisfy this ancestor condition, and which violate it? This asks only about eligibility under the ancestor rule; the information given does not determine the TreeBH leaf decisions.

**Exercise 7.17: Discrete super-uniform p-values**

Theorem 7.2 applies to independent super-uniform p-values. Let $U_1, \dots, U_m$ be independent $\mathrm{Unif}(0, 1)$ variables, and set $p_i = \lceil KU_i \rceil / K$ for a fixed integer $K \geq 2$. Compute $\mathbb{P}(p_i \leq t)$ for $t \in [0, 1]$, and use this calculation to verify super-uniformity.

### Intermediate.

**Exercise 7.18: Simes-style PC p-value: validity and sharpness**

The Bonferroni PC p-value of Definition 7.10 can be sharpened under conditions ensuring the Simes inequality for the relevant true-null per-study p-values by the Simes-style PC p-value

$$p_i^{(r/K),\mathrm{Simes}} = \left[ \min_{r \leq j \leq K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1} \right] \wedge 1.$$

Show that this is super-uniform under $H_i^{(r/K)}$, for example when the per-study p-values are independent (or suitable PRDS [158]) and the true-null p-values are super-uniform. Hint: choose a fixed subset of $K - r + 1$ true-null studies, as in the proof of Theorem 7.11, and compare the full order statistics with the order statistics of this subset before applying the Simes inequality. Then show algebraically that the Simes-style PC p-value is no larger than the Bonferroni-style PC p-value at every realization.

**Exercise 7.19: p-filter as BH special case**

Verify that the p-filter with $L = 1$ layer at the hypothesis level reduces to ordinary BH. Then verify that the p-filter with $L = 1$ layer using Simes-aggregated group representatives reduces, at the group level, to group BH of Section 2.

## Computational.

### Exercise 7.20: Genomic pathway simulation

Simulate a three-level hierarchy: 20 pathways, each containing 20 genes (400 genes total), each gene containing 25 variants (10,000 variants total). Generate each variant p-value as $p_i = \Phi(-Z_i)$ where $Z_i \sim N(\mu_i, 1)$, with $\mu_i = 2.5$ for signal variants and $\mu_i = 0$ for nulls. Place signal in 4 pathways, 5 signal genes per signal pathway (out of 20 genes per pathway), and 5 signal variants per signal gene (out of 25 variants per gene); all other variants are nulls. Generate the $Z_i$'s independently. For TreeBH, compute internal-node p-values recursively from immediate children using Simes; for the p-filter, apply Simes directly to all variant p-values in each layer group. Implement (a) flat BH at the variant level at $q = 0.05$; (b) TreeBH with per-level targets $(q_{\text{pathway}}, q_{\text{gene}}, q_{\text{variant}}) = (0.05, 0.10, 0.10)$, corresponding to $(q_1, q_2, q_3)$ in the coarsest-to-finest TreeBH order; and (c) the p-filter with layer levels mapped to its finest-to-coarsest convention: $q_{\text{variant}} = 0.10$, $q_{\text{gene}} = 0.10$, and $q_{\text{pathway}} = 0.05$. Average over $\geq 200$ replicates and report variant-, gene-, and pathway-level FDR for flat BH and the p-filter. For TreeBH, report recursive selected-family FDR at each tested depth, using $s\,\text{FDR}^1, s\,\text{FDR}^2, s\,\text{FDR}^3$ in (4). Also report average power at each resolution for all three procedures. For flat BH and the p-filter, count a gene or pathway as discovered when it contains a rejected variant; for TreeBH, use its rejected nodes at each depth. A gene or pathway is nonnull when it contains at least one signal variant. At each resolution, use the average-power definition of Exercise 4.16, with nonnull units and discoveries counted at that resolution.

### Exercise 7.21: Replication study

Generate three studies with $m = 1000$ features each. For feature $i$ in study $k$, use $p_{i,k} = \Phi(-Z_{i,k})$, with independent $Z_{i,k} \sim N(\mu_{i,k}, 1)$ across all features and studies. Set $\mu_{i,k} = 0$ for nulls and $\mu_{i,k} = 2.5$ for signals. Of the 100 replicated-truth features, let 50 be signals in exactly two studies, assigning study pairs $(1,2), (1,3), (2,3)$ cyclically, and let 50 be signals in all three. Another 100 features are signals in exactly one study, assigning studies $1, 2, 3$ cyclically; the remaining 800 features are null in all three studies.

Compare three procedures targeting $r = 2$-out-of-$K = 3$ replication: (a) BH at $q = 0.05$ in each study separately, declaring a feature replicated if it is rejected in at least two studies; (b) BH at $q = 0.05$ applied to the Bonferroni PC p-values; and (c) BH at $q = 0.05$ applied to the Simes-style PC p-values. Use at least 200 independent replicates. Count false discoveries as selected features with fewer than two nonnull studies and true discoveries as selected features among the 100 replicated truths. Report empirical replication FDR, the mean of $V/(R \vee 1)$ across replicates, and empirical average power, the mean fraction of the 100 replicated truths discovered.

### Exercise 7.22: Tree depth and power

Generate a balanced tree with three tested levels, branching factor 4, and 64 leaves labeled from left to right. Let leaves $1, \ldots, 16$, which lie below the leftmost child of the formal root, be signals; all other leaves are null. Use $p_i = \Phi(-Z_i)$, with independent $Z_i \sim N(\mu_i, 1)$, where $\mu_i = 2.5$ for signals and $\mu_i = 0$ for nulls. Form internal-node p-values recursively from immediate children using Simes. Apply TreeBH with top-tested-layer target $q_1 = 0.05$ and $q_d = 0.10$ for every lower tested depth $d \geq 2$.

Repeat with 64 leaves arranged as six tested levels with branching factor 2, again labeled from left to right. Keep exactly the same signal leaves $1, \ldots, 16$; they now lie below the leftmost depth-two node. Keep the marginal distributions, recursive Simes aggregation, and per-depth target convention unchanged. Use at least 200 independent replicates per design. Report empirical leaf-level average power and leaf-level recursive selected-family FDR: $s\,\text{FDR}^3$ for the first tree and $s\,\text{FDR}^6$ for the second, using (4). Discuss how the added depth affects these quantities.

**Advanced.**

**Exercise 7.23: TreeBH versus flat BH for high-density signal**

Use a four-level binary tree with 16 leaves. Put signals in all four leaves below the leftmost depth-two internal node and set all other leaves to null. Generate independent one-sided p-values from $Z_i \sim N(2.5, 1)$ for signal leaves and $Z_i \sim N(0, 1)$ for null leaves, with internal-node p-values formed recursively by Simes combinations of immediate-child p-values, matching Section 4. Compare flat BH on the 16 leaves at $q = 0.10$ with TreeBH using $(q_1, q_2, q_3, q_4) = (0.05, 0.10, 0.10, 0.10)$ over at least 500 independent simulation runs. Report empirical leaf-level average power and explain the interpretation gained when a rejected leaf is accompanied by rejected ancestors.

**Exercise 7.24: Restricted PC with verified side information**

Heller and Bogomolov [81] discuss sharper procedures that use information about $|\mathcal{N}_i|$ beyond the worst-case bound. Suppose a rule fixed by the study design before looking at the p-values identifies a subset $S_i$ of $K - r + 1$ studies that is guaranteed to consist of true nulls whenever $H_i^{(r/K)}$ holds. The restricted PC p-value is

$$p_i^S = \left[ (K - r + 1) \min_{k \in S_i} p_{i,k} \right] \wedge 1.$$

Its super-uniformity follows from the union-bound step in the proof of Theorem 7.11; under independence across features, the BH guarantee follows as in Theorem 7.14. Explain what can go wrong if $S_i$ is inferred from the same p-values, or if the side information is only an unverified belief. Why can the resulting procedure fail to control FDR uniformly over the PC null?

CHAPTER 8

# E-Values, Safe Testing, and Multiple Testing

Suppose a laboratory is running many experiments at once. Some outcomes are observed today, more data may arrive next week, and the same subjects, batches, or model outputs may be reused across several hypotheses. The analyst wants to make discoveries while the evidence is still being monitored. This is the setting in which ordinary fixed-sample p-values become fragile: if we repeatedly look at the data and then decide whether to continue, the final fixed-sample p-value no longer has the error guarantee suggested by its formula. Here *fixed-sample* means that the p-value formula was calibrated for one prespecified sample size and one planned analysis. It was not calibrated for a workflow in which the sample size or the number of looks is chosen after seeing interim evidence.

This chapter introduces e-values as an alternative evidence scale. The main idea is deliberately simple. A p-value controls a tail probability under the null; an e-value controls an expectation under the null. This expectation budget can be spent in ways that are difficult for p-values: e-values can be multiplied over time when each new factor is conditionally valid, averaged across procedures, pooled across data sources, and used in the e-BH procedure to control FDR under arbitrary dependence.

The chapter is organized around a running question. How should we build an evidence measure that remains valid after monitoring, combining, or selecting among many hypotheses? The answer has three parts:

(1) construct evidence whose null expectation is controlled;
(2) prove that the construction did not spend too much null-evidence budget;
(3) apply a rejection rule whose proof uses only that budget.

Appendix A develops the more advanced theory. It explains how to choose evidence that grows quickly under a specified alternative, when approximate distributional calculations justify approximate e-values, and how to test hypotheses as they arrive over time. It also constructs confidence sets that cover the true parameter simultaneously at all observation times, so their coverage remains valid when the data determine when to stop.

## 1. Setup and Notation

**Route: Core**

This chapter uses the multiple-testing notation introduced in Chapters 5–6. We recall the formulas needed below, and then introduce the new sequential and e-value notation for this chapter.

For one hypothesis, let $X$ denote the observed data and let $\mathcal{H}_0$ be the null model, represented as a set of probability distributions for $X$. A simple null has the form $\mathcal{H}_0 = \{P_0\}$. A composite null has the form

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\},$$

where the null parameter set $\Theta_0$ contains more than one value. When densities exist, $p_0(x)$ denotes the density of $P_0$, $p_\theta(x)$ denotes the density of $P_\theta$, and $g(x)$ denotes the density of an

alternative distribution $Q$. The letter $g$, rather than $q$, keeps an alternative density distinct from the book-wide FDR target. These are density functions, not p-values.

Recall that for multiple testing there are $m$ hypotheses $H_1, \dots, H_m$. The set of true null indices is $\mathcal{I}_0 \subset \{1, \dots, m\}$, with $m_0 = |\mathcal{I}_0|$. For a rejection set $\mathcal{R}$, we write

$$R = |\mathcal{R}|, \qquad V = |\mathcal{R} \cap \mathcal{I}_0|, \qquad \mathrm{FDP} = \frac{V}{R \vee 1}, \qquad \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

The convention $R \vee 1$ makes $\mathrm{FDP} = 0$ when there are no rejections.

For sequential problems, $\mathcal{F}_t$ is the information available after the $t$th look or batch; the increasing family $(\mathcal{F}_t)_{t \geq 0}$, with $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$, is called a *filtration*. A quantity is *adapted* if its value at time $t$ is $\mathcal{F}_t$-measurable, that is, computable from the information available at time $t$. A rule is *predictable* at time $t$ if it is $\mathcal{F}_{t-1}$-measurable; in plain language, it is chosen using only past data. A stopping time $\tau$ is a time at which the decision to stop is made using the information available so far. A deterministic finite horizon is denoted by $K$. We reserve $t$ for time or a generic threshold variable, and write $\widehat{t}$ for data-dependent thresholds.

## 2. Evidence on an Expectation Scale

**Route: Core**

**Definition 8.1: E-value**

A nonnegative random variable $E = E(X)$ is an e-value for $\mathcal{H}_0$ if

$$\sup_{P \in \mathcal{H}_0} \mathbb{E}_P[E] \leq 1.$$

Large values of $E$ are evidence against the null.

Two conventions will be useful later. First, we allow $E$ to take the value $+\infty$, provided that $\mathbb{P}_P(E = \infty) = 0$ for every $P \in \mathcal{H}_0$. An infinite e-value plays the same role as a p-value equal to zero: it may occur only on events that are impossible under the null. Second, we call an e-value *exact* if

$$\mathbb{E}_P[E] = 1 \qquad \text{for every } P \in \mathcal{H}_0.$$

For a simple null, exactness means the entire null-evidence budget is spent. If $\mathbb{E}_{P_0}[E] < 1$, the e-value $E + 1 - \mathbb{E}_{P_0}[E]$ is at least as large as $E$ and is exact; the likelihood ratios constructed below are also exact. For a composite null, exactness at every null distribution is usually too restrictive, and a useful e-value may meet the budget with equality only at a least favorable null distribution; the one-sided Gaussian example later in this chapter has this property.

Recall from Chapter 2 that a p-value, denoted by the lowercase letter $p$, is a $[0, 1]$-valued statistic satisfying

$$\sup_{P \in \mathcal{H}_0} \mathbb{P}_P(p \leq u) \leq u, \qquad 0 \leq u \leq 1.$$

Small p-values are evidence against the null. A statistic satisfying this tail inequality is called *super-uniform*: it is stochastically no smaller than a $\mathrm{Unif}(0, 1)$ random variable. Although $p$ is computed from the data, we avoid writing $p(X)$ here because symbols such as $p_0(x)$ and $p_\theta(x)$ denote density functions in this chapter.

The p-value and e-value definitions use different currencies. A p-value promises that small values are rare under the null. An e-value promises that the average amount of null evidence is at most one. The second promise is weaker in some fixed sample problems, but it is algebraically useful because expectations add and conditional expectations multiply.

**Proposition 8.2: Markov calibration**

If $E$ is an e-value for $\mathcal{H}_0$, then

$$p_E = \min\{1, 1/E\}$$

is a valid p-value for $\mathcal{H}_0$.

**Proof of Proposition 8.2**

Fix $P \in \mathcal{H}_0$, using $1/0 = \infty$ and $1/\infty = 0$. For $0 < u < 1$,

$$\mathbb{P}_P(p_E \le u) = \mathbb{P}_P(E \ge 1/u) \le u\,\mathbb{E}_P[E] \le u,$$

where the inequality is Markov's inequality. For $u = 0$, the event is $\{E = \infty\}$, which has null probability zero. For $u = 1$, the probability is one because $p_E \le 1$. The truncation at one only keeps the calibrated p-value inside $[0, 1]$. □

Figure 8.1 makes the reciprocal geometry visible: strong evidence maps to a small calibrated p-value, whereas every e-value below one is truncated to the uninformative value $p_E = 1$.

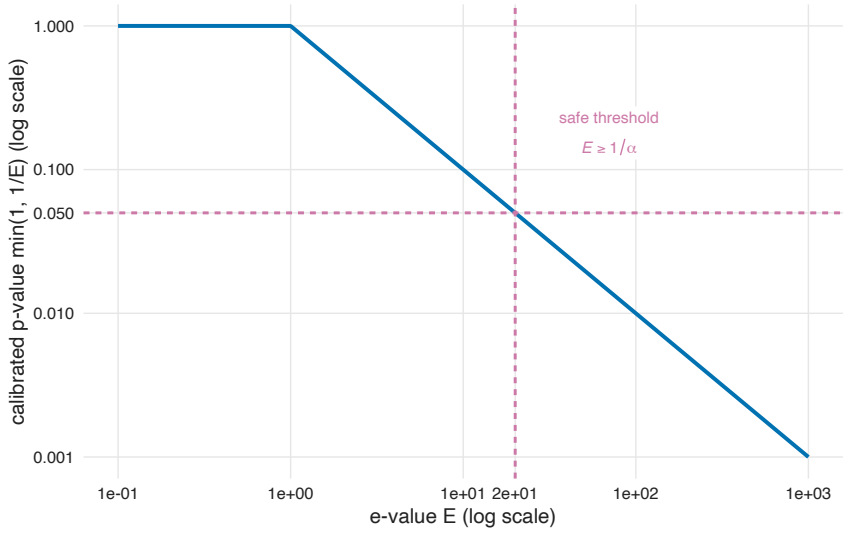



FIGURE 8.1. Calibration from e-values to p-values on log-log axes. The conversion $p_E = \min\{1, 1/E\}$ is always valid by Markov's inequality. Weak e-values with $E < 1$ all map to $p_E = 1$, so calibration is valid but can lose information. The dashed lines mark the level $\alpha = 0.05$ and its rejection threshold $E = 20$.

The reverse direction is more delicate, and it is where the two evidence scales genuinely differ.

**Calibrating p-values into e-values.** The naive conversion $1/p$ fails. If $p \sim \text{Unif}(0, 1)$, then

$$\mathbb{E}[1/p] = \int_0^1 \frac{1}{u}\,du = \infty,$$

so $1/p$ is very far from being an e-value. A valid conversion rule must be declared before the data are seen.

**Definition 8.3: Calibrator**

A *calibrator* is a nonincreasing function $f : [0,1] \to [0,\infty]$ such that $f(p)$ is an e-value whenever $p$ is a p-value, for every null model $\mathcal{H}_0$.

To compare calibrators we need two more terms, which will be reused several times in this chapter. A calibrator $f$ *dominates* a calibrator $g$ if $f \geq g$ pointwise, so that $f$ never reports less evidence than $g$ from the same p-value; the domination is *strict* if in addition $f \neq g$. A calibrator is *admissible* if no calibrator strictly dominates it. Admissibility means the rule cannot be improved for free: any attempt to report more evidence somewhere breaks validity.

**Proposition 8.4: Characterization of calibrators**

A nonincreasing function $f : [0,1] \to [0,\infty]$ is a calibrator if and only if

$$\int_0^1 f(u)\,du \leq 1.$$

It is admissible if and only if, in addition, $f$ is left-continuous on $(0,1]$, $f(0) = \infty$, and $\int_0^1 f(u)\,du = 1$.

**Proof of the first statement**

Suppose $\int_0^1 f(u)\,du \leq 1$, and let $p$ be a p-value under $P \in \mathcal{H}_0$. Fix $x \geq 0$. Because $f$ is nonincreasing, the set $\{u \in [0,1] : f(u) > x\}$ is an interval starting at 0; let $u_x$ be its right endpoint, taking $u_x = 0$ if the set is empty. The event $\{f(p) > x\}$ is contained in $\{p \leq u_x\}$ or $\{p < u_x\}$, so the p-value property gives $\mathbb{P}_P(f(p) > x) \leq u_x$. For $U \sim \mathrm{Unif}(0,1)$, the same interval has probability exactly $u_x$, so

$$\mathbb{P}_P(f(p) > x) \leq u_x = \mathbb{P}(f(U) > x).$$

Integrating the tail probabilities over $x \in [0,\infty)$,

$$\mathbb{E}_P[f(p)] \leq \mathbb{E}[f(U)] = \int_0^1 f(u)\,du \leq 1.$$

Conversely, an exactly uniform p-value is available in the models used in this chapter, and for $p \sim \mathrm{Unif}(0,1)$ the requirement $\mathbb{E}[f(p)] \leq 1$ is precisely $\int_0^1 f(u)\,du \leq 1$. For the proof of the admissibility statement, see Vovk and Wang [188] and Chapter 2 of Ramdas and Wang [147]. □

The integral condition makes examples easy to generate and easy to check. Each of the following functions integrates to one on $[0,1]$, takes the value $\infty$ at zero, and is left-continuous, so each is an admissible calibrator:

$$f_\kappa(p) = \kappa p^{\kappa-1}, \qquad 0 < \kappa < 1,$$

$$f_{\mathrm{mix}}(p) = \frac{1 - p + p\log p}{p(\log p)^2}, \qquad 0 < p < 1,$$

$$f_{\mathrm{sqrt}}(p) = p^{-1/2} - 1, \qquad f_{\log}(p) = -\log p.$$

For every displayed calibrator, set $f(0) = \infty$. The mixture calibrator averages the power calibrators with equal weight over $\kappa \in (0, 1)$. Holding the observed p-value fixed and integrating over $\kappa$ gives

$$f_{\mathrm{mix}}(p) = \int_0^1 \kappa p^{\kappa-1}\, d\kappa = \frac{1 - p + p \log p}{p(\log p)^2}, \qquad 0 < p < 1.$$

The integral form also explains the endpoint: at $p = 1$, it equals $\int_0^1 \kappa\, d\kappa = 1/2$. We therefore define $f_{\mathrm{mix}}(1) = 1/2$, the continuous extension of the displayed formula. Each power calibrator has integral one over p-values, so their average does too. This mixture is useful when one does not want to commit to a single $\kappa$. The function equal to $2(1 - p)$ for $p > 0$ and to $\infty$ at $p = 0$ is also an admissible calibrator; away from zero it never reports evidence larger than 2, which limits its use for single tests but makes it stable inside products. As a concrete number, $f(p) = p^{-1/2} - 1$ converts $p = 0.01$ into $e = 9$: an honest pre-declared conversion turns a p-value of 0.01 into roughly ninefold evidence, not hundredfold. Figure 8.2 plots these calibrators together: all sit below the hindsight envelope $1/p$, and none dominates the others, which is exactly what admissibility permits.

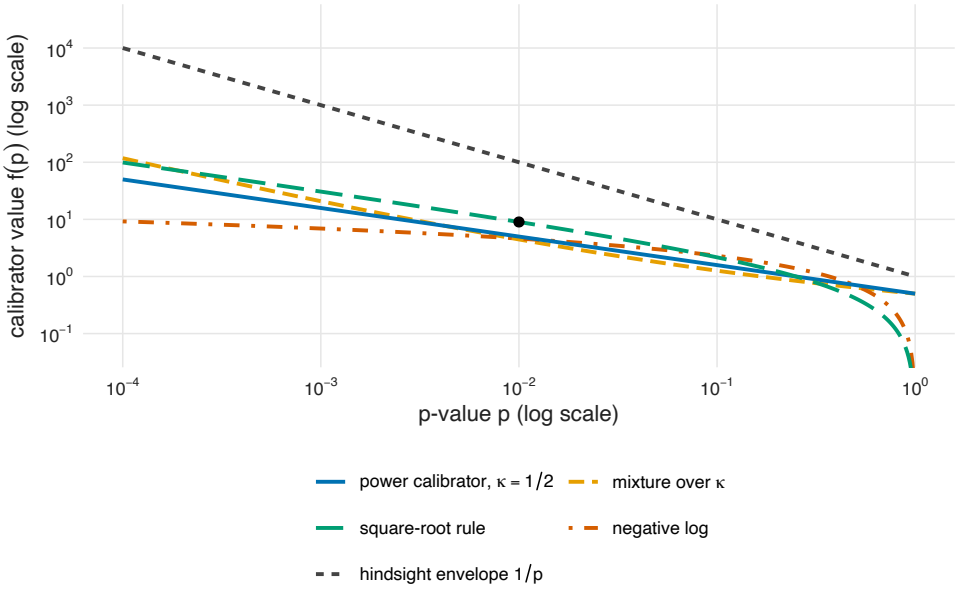



FIGURE 8.2. Four admissible calibrators on log-log axes, with the hindsight envelope $1/p$ as a dashed line. Every calibrator lies below the envelope, and no calibrator dominates another: the mixture is largest for very small p-values, the square-root rule near $p = 0.01$, the negative logarithm for moderate p-values, and the power calibrator for large ones. The marked point is the conversion $p = 0.01 \mapsto e = 9$ under the square-root rule $f(p) = p^{-1/2} - 1$. The square-root and negative-log rules both equal zero at $p = 1$; these zero endpoints cannot be shown on a logarithmic axis. The power calibrator with $\kappa = 1/2$ and the mixture calibrator both equal $1/2$ at $p = 1$.

Two facts sharpen the comparison with $1/p$. First, every calibrator satisfies

$$f(p) \le \frac{1}{p}, \qquad 0 < p \le 1.$$

Indeed, if $f(p_0) > 1/p_0$ for some $p_0$, then, since $f$ is nonincreasing,

$$\int_0^1 f(u)\,du \geq \int_0^{p_0} f(u)\,du \geq p_0 f(p_0) > 1.$$

Equality $f(p_0) = 1/p_0$ at a fixed point $p_0$ forces $f(u) = \mathbf{1}\{u \leq p_0\}/p_0$ for all $u \in (0,1]$: up to its value at zero, $f$ must be the calibrator that stakes everything on the single threshold $p_0$. Second, the failure of $1/p$ has a betting interpretation. For each fixed $r \in (0,1]$, the all-or-nothing function

$$f_r(u) = \frac{\mathbf{1}\{u \leq r\}}{r}$$

is a calibrator, because $\int_0^1 f_r(u)\,du = 1$. It is a pre-declared bet that pays $1/r$ if the p-value lands below $r$ and zero otherwise. The realized value $1/p$ is exactly the payoff of the best such bet chosen with hindsight, $r = p$. Reporting $1/p$ as evidence is double counting: the data are used once to choose the bet and again to settle it.

In the reverse direction there is essentially no choice. Among all nonincreasing functions $h$ such that $h(E)$ is a p-value for every e-value $E$, the Markov calibration $h(e) = \min\{1, 1/e\}$ of Proposition 8.2 dominates every other in the sense appropriate for p-values, where smaller is more informative: it reports a p-value no larger than any valid competitor's, and it is the unique admissible choice [147, 188]. It also preserves decisions at every level simultaneously: for $\alpha \in (0,1)$,

$$E \geq \frac{1}{\alpha} \quad \Longleftrightarrow \quad \min\{1, 1/E\} \leq \alpha,$$

so thresholding the e-value and thresholding its calibrated p-value are the same test at every $\alpha$. No p-to-e calibrator has this property: by the bound $f(p) \leq 1/p$, converting $p$ to $f(p)$ and then testing $f(p) \geq 1/\alpha$ can only remove rejections from the rule $p \leq \alpha$. The all-or-nothing calibrator at a chosen threshold preserves that particular test, but no single calibrator preserves all p-value thresholds. Calibration in both directions loses evidence: converting $p = 0.01$ to $e = 9$ and back yields $1/9 \approx 0.11$. The practical advice is to choose the evidence scale once, at design time, based on whether the analysis needs the algebra of expectations developed in the rest of this chapter. For a single fixed null distribution, one can also represent any e-value as $f(p)$ for a suitable calibrator and an exactly uniform p-value, allowing independent randomization when needed [147]. This representation concerns one null law; it is not a claim that every e-value for a composite null has the same distribution under all null laws.

**How large is a convincing e-value?** Students meeting e-values for the first time reasonably ask what a value such as $E = 7$ means. For e-values built from likelihood ratios and their generalizations, which is most of this chapter, a useful rough grid, in the spirit of Jeffreys's scale for Bayes factors (ratios of marginal likelihoods, introduced in the next section), is given in Table 8.1.

TABLE 8.1. A rough grid for reading the size of likelihood-ratio-type e-values. It does not apply to all-or-nothing conversions.

| Observed e-value | Reading |
|---|---|
| $E < 1$ | favors the null model over the specified alternative |
| $E = 1$ | neutral |
| $1 < E \leq 3.2$ | barely worth mentioning |
| $3.2 < E \leq 10$ | substantial for the specified likelihood comparison |
| $10 < E \leq 32$ | strong for the specified likelihood comparison |
| $32 < E \leq 100$ | very strong |
| $E > 100$ | decisive |

Two caveats keep the grid honest. First, it applies to naturally constructed e-values such as likelihood ratios, not to all-or-nothing e-values of the form $\mathbf{1}\{p \le \alpha\}/\alpha$; for the latter, $E = 10$ only records that $p \le 0.1$. Second, rejecting when $E \ge 1/\alpha$ gives a valid predeclared level-$\alpha$ e-test by Markov's inequality. The descriptive labels in the table are not error guarantees; the level-$\alpha$ threshold is justified by the expectation bound. It can be conservative for a single prespecified test; the value of the e-value scale is what it buys later in this chapter, namely monitoring, multiplication, combination, and e-BH.

A final warning: a p-value and an e-value computed from the same data can point in opposite directions. Let $X \sim N(\mu, 1)$, test $H_0 : \mu = 0$ against the alternative $\mu = 5$, and observe $X = 2$. The one-sided p-value is $1 - \Phi(2) \approx 0.023$, which most analysts read as evidence against the null. The likelihood-ratio e-value is

$$E = \exp\{5X - 25/2\} = \exp\{-2.5\} \approx 0.08,$$

which supports the null: the observation $X = 2$ is unlikely under $N(0, 1)$ but far more unlikely under $N(5, 1)$. There is no contradiction. The p-value measures rarity under the null alone; the likelihood-ratio e-value compares the null against a specific alternative. Only the calibrated conversions of this section are guaranteed to preserve direction; separately constructed p-values and e-values need not agree.

## 3. Likelihood Ratios, Bayes Factors, and Composite Nulls

**Route: Core**

The most basic way to construct an e-value is to compare two probability models.

**Proposition 8.5: Likelihood-ratio e-value**

Suppose $\mathcal{H}_0 = \{P_0\}$ is simple and $Q$ is absolutely continuous with respect to $P_0$. Then

$$E(X) = \frac{dQ}{dP_0}(X)$$

is an e-value for $P_0$.

**Proof of Proposition 8.5**

By the definition of the Radon-Nikodym derivative,

$$\mathbb{E}_{P_0}\left[\frac{dQ}{dP_0}(X)\right] = \int \frac{dQ}{dP_0}\, dP_0 = \int dQ = 1.$$

□

The computation shows more than validity: the null expectation equals one exactly, so likelihood-ratio e-values are exact in the sense of the previous section. For the simple null $\mathcal{H}_0 = \{P_0\}$, say that an e-value $E'$ *dominates* $E$ if $E' \ge E$ $P_0$-almost surely, and that the domination is strict if $P_0(E' > E) > 0$. Call $E$ *admissible* if no other e-value strictly dominates it. With this almost-sure definition, an e-value for a simple null is admissible if and only if it is exact, if and only if it equals $dS/dP_0$ for some probability distribution $S$ absolutely continuous with respect to $P_0$ [147]. Thus choosing an admissible e-value for a simple null amounts to choosing a distribution $S$ whose likelihood ratio will be used as evidence.

If densities are available, the same e-value is $E(X) = g(X)/p_0(X)$. If $Q$ is a mixture over alternatives,

$$g(x) = \int p_\theta(x)\, dG_1(\theta),$$

then $g(X)/p_0(X)$ is a Bayes factor and also an e-value for the simple null. The phrase "simple null" is doing real work. The denominator $p_0$ is the density of the one null distribution whose expectation must be controlled. More generally, for any prior distribution $G$ on a parameter set, write

$$p_G(x) = \int p_\theta(x)\, dG(\theta),$$

and let $Q_G$ denote the probability distribution with density $p_G$. This notation is used below for both alternative mixtures and null mixtures.

**Example 8.6: Gaussian likelihood ratio**

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$. We test

$$H_0 : \mu = 0 \qquad \text{against} \qquad H_1 : \mu = \mu_1 > 0.$$

The joint density under $\mu$ is

$$p_\mu(x_1, \dots, x_n) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} (x_j - \mu)^2 \right\}.$$

Therefore

$$\begin{aligned} \frac{p_{\mu_1}(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} &= \exp\left\{ -\frac{1}{2} \sum_{j=1}^{n} (X_j - \mu_1)^2 + \frac{1}{2} \sum_{j=1}^{n} X_j^2 \right\} \\ &= \exp\left\{ \mu_1 \sum_{j=1}^{n} X_j - \frac{n\mu_1^2}{2} \right\}. \end{aligned}$$

This is an e-value for $H_0 : \mu = 0$. Under the alternative $\mu = \mu_1$, its logarithm has positive mean $n\mu_1^2/2$, so the e-value tends to grow.

Composite nulls require more care. Suppose $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$, and assume that the alternative mixture is absolutely continuous with respect to the null mixture, $Q_{G_1} \ll Q_{G_0}$. The Bayes factor

$$\frac{p_{G_1}(X)}{p_{G_0}(X)}$$

is then a version of $dQ_{G_1}/dQ_{G_0}$. It averages the null over a prior $G_0$ on $\Theta_0$, but it is not automatically an e-value for the whole composite null. To be an e-value, it must satisfy

$$\mathbb{E}_{P_\theta}\left[ \frac{p_{G_1}(X)}{p_{G_0}(X)} \right] \le 1 \qquad \text{for every } \theta \in \Theta_0.$$

Averaging over the null prior gives only

$$\mathbb{E}_{\theta \sim G_0} \mathbb{E}_{P_\theta}\left[ \frac{p_{G_1}(X)}{p_{G_0}(X)} \right] = 1,$$

which does not control the worst null value of $\theta$.

One useful design problem is therefore: given an alternative distribution $Q$, find the e-value for the composite null that grows fastest under $Q$. To make "fastest" precise, we compare e-values

on the logarithmic scale; the next section explains why the logarithm is the right scale when evidence is accumulated multiplicatively. Say that an e-value $E^*$ for $\mathcal{H}_0$, positive $Q$-almost surely, is *log-optimal* against $Q$ if

$$\mathbb{E}_Q\left[\log \frac{E}{E^*}\right] \le 0 \qquad \text{for every e-value } E \text{ for } \mathcal{H}_0.$$

The ratio inside the logarithm matters: it keeps the comparison meaningful even in problems where $\mathbb{E}_Q[\log E]$ is infinite for some e-values. For these comparisons, use $x/\infty = 0$ and $\infty/x = \infty$ for finite $x \ge 0$, and $\infty/\infty = 1$; also $\log 0 = -\infty$. The last ratio convention treats two infinite evidence values as equal [110].

A benchmark for comparing evidence. A *numeraire* is a benchmark e-value against which we measure other e-values. Fix the null family $\mathcal{H}_0$ and an alternative $Q$. An e-value $E^*$, positive $Q$-almost surely, is a numeraire if

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] \le 1 \qquad \text{for every e-value } E \text{ for } \mathcal{H}_0. \tag{5}$$

There are two different expectations here. To be a valid e-value, $E^*$ must have mean at most one under every null distribution. The displayed comparison instead averages under the specified alternative $Q$: after dividing a competing e-value by $E^*$, its mean is at most one. Jensen's inequality then gives

$$\mathbb{E}_Q[\log(E/E^*)] \le \log \mathbb{E}_Q[E/E^*] \le 0,$$

so a numeraire is log-optimal. Appendix A develops its existence and uniqueness and its connection to reverse information projection, including cases where the optimal e-value is a likelihood ratio against a suitable null mixture. These optimality results help us choose an e-value; ordinary validity still requires only the null expectation bound.

**Universal inference: validity without regularity conditions.** Finding a valid likelihood ratio against a composite null can be difficult. Universal inference is a general-purpose construction that trades optimality for applicability: it produces a valid e-value for essentially any composite null where maximum likelihood can be computed, with no regularity conditions on the model [192].

Split the sample into two independent parts $D_0$ and $D_1$; for i.i.d. data any predetermined split works. Using $D_1$ only, construct any probability density $\widehat{g}_1$ intended to approximate the alternative: a maximum likelihood fit, a Bayes predictive density, or even a fixed guess. Using $D_0$, compute the null maximum likelihood.

**Proposition 8.7: Split likelihood-ratio e-value**

Let $D_0$ and $D_1$ be independent blocks of observations, i.i.d. from a common distribution. Let $\widehat{g}_1$ be a probability density built from $D_1$ alone, and let $\widehat{p}_0$ maximize the null likelihood on $D_0$:

$$\prod_{X_j \in D_0} \widehat{p}_0(X_j) = \max_{\theta \in \Theta_0} \prod_{X_j \in D_0} p_\theta(X_j).$$

Then

$$E = \frac{\prod_{X_j \in D_0} \widehat{g}_1(X_j)}{\prod_{X_j \in D_0} \widehat{p}_0(X_j)}$$

is an e-value for $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$. For each $\theta$, define $\widehat{g}_1(x)/p_\theta(x) = 0$ whenever $p_\theta(x) = 0$.

**Proof of Proposition 8.7**

Fix $\theta \in \Theta_0$. The null maximum likelihood is at least the likelihood at $\theta$, so, $P_\theta$-almost surely,

$$E \le \prod_{X_j \in D_0} \frac{\widehat{g}_1(X_j)}{p_\theta(X_j)}.$$

Condition on $D_1$. Given $D_1$, the function $\widehat{g}_1$ is a fixed probability density, and the observations in $D_0$ are independent draws from $P_\theta$. Each factor then has conditional expectation

$$\mathbb{E}_{P_\theta}\left[\frac{\widehat{g}_1(X_j)}{p_\theta(X_j)} \,\middle|\, D_1\right] = \int_{\{p_\theta > 0\}} \widehat{g}_1(x)\, dx \le 1,$$

with equality precisely when $\widehat{P}_1 \ll P_\theta$. Here $\widehat{P}_1$ is the distribution with density $\widehat{g}_1$. The factors are conditionally independent, so $\mathbb{E}_{P_\theta}[E \mid D_1] \le 1$. Averaging over $D_1$ gives $\mathbb{E}_{P_\theta}[E] \le 1$. □

If the null maximum is not attained, replace the maximum by the supremum; this only decreases $E$, so validity is preserved. The remarkable feature of the proposition is what it does not assume: no smoothness of the model in $\theta$, no dimension conditions, no asymptotics. It applies to irregular problems, such as mixture models with an unknown number of components, where the classical generalized-likelihood-ratio theory has no usable distributional limit. The price is conservativeness: universal inference is generally not exact. For a specified alternative, the numeraire is log-optimal, so a split likelihood ratio has expected log evidence relative to the numeraire at most zero. This compares average relative log evidence under the specified alternative: on a particular dataset, the split likelihood ratio may still be larger than the numeraire [110]. Data splitting also costs power, which can be partly recovered by swapping the roles of $D_0$ and $D_1$ and averaging the two e-values [55, 192].

It is worth placing the constructions of this section side by side.

- The frequentist generalized likelihood ratio, which maximizes the numerator over the alternative parameter set on the observed data, is not an e-value in general: optimizing the numerator after seeing the data spends more than the null-evidence budget.
- A Bayes factor with an alternative numerator and a null-mixture denominator must satisfy the expectation bound under every null parameter. If $E^* = g/p_{G_0}$ is valid and positive under the alternative $Q$, then for any competing e-value $E$, $\mathbb{E}_Q[E/E^*] = \int_{\{g>0\}} E p_{G_0} \le \mathbb{E}_{Q_{G_0}} E \le 1$. Thus, when competing e-values are divided by $E^*$, their means under $Q$ are at most one. This is exactly the numeraire condition (5), so the valid ratio $E^*$ is also log-optimal. Appendix A gives the full guess-and-verify argument and discusses the support assumptions.
- Universal inference replaces both the projection and the verification by a data split; it is always valid, and typically conservative.

**A permutation e-value.** Some composite nulls are defined by a symmetry rather than by a parameter set. The null of exchangeability states that the joint distribution of the data is invariant under permutations of the observations; it contains every i.i.d. model, and it is the null behind permutation tests.

**Example 8.8: Soft-rank e-value**

Let $L_0 = T(X)$ be a test statistic computed from the data, and let $L_1, \dots, L_B$ be recomputations of the same statistic after applying independent uniformly random permutations to the observations. Under the exchangeability null, the vector $(L_0, L_1, \dots, L_B)$ is exchangeable: its distribution is unchanged when its coordinates are permuted. Form nonnegative scores by subtracting the minimum,

$$R_b = L_b - \min_{0 \le b' \le B} L_{b'}, \qquad S = \sum_{b=0}^{B} R_b,$$

and define

$$E = \frac{(B+1)R_0}{S},$$

with $E = 0$ when $S = 0$. Then $E$ is an e-value. To see this, condition on the unordered multiset of score values $\{R_0, R_1, \dots, R_B\}$. Exchangeability makes every assignment of these values to the positions $0, 1, \dots, B$ equally likely, so

$$\mathbb{E}[R_0 \mid R_0 + \cdots + R_B] = \frac{S}{B+1},$$

and therefore

$$\mathbb{E}[E] = \mathbb{E}\left[\mathbf{1}\{S > 0\}\, \frac{B+1}{S}\, \mathbb{E}(R_0 \mid S)\right] = \mathbb{P}(S > 0) \le 1,$$

using that $S$ is determined by the conditioning information and that $E = 0$ when $S = 0$. Compare with the standard permutation p-value $p_{\text{perm}} = (B+1)^{-1} \sum_{b=0}^{B} \mathbf{1}\{L_b \ge L_0\}$. Since the scores preserve order,

$$S \ge \sum_{b : R_b \ge R_0} R_b \ge R_0 \cdot \#\{b : L_b \ge L_0\} = R_0 (B+1) p_{\text{perm}},$$

so $E \le 1/p_{\text{perm}}$. For a single fixed-sample test, the permutation p-value therefore always beats the calibrated conversion of this e-value. The e-value earns its keep exactly in the settings this chapter develops: independent batches multiply, laboratories combine, and many hypotheses enter e-BH.

Exchangeability nulls of this type reappear in Chapter 9 for conditional randomization and knockoff methods and in Chapter 10 for conformal prediction.

## 4. Betting Scores and Safe Tests

**Route: Core**

The betting language is only a bookkeeping device. No actual money is needed. Imagine that we keep an evidence account whose starting balance is one. The number one means "no accumulated evidence yet," so all later values are measured relative to this neutral starting point. At time $t$, after seeing the next piece of data, we multiply the current balance by a nonnegative factor $E_t$. If $E_t = 2$, the account doubles; if $E_t = 1/2$, the account is cut in half; if $E_t = 1$, the new data leave the account unchanged. The cumulative evidence, also called the capital process, is therefore

$$M_t = \prod_{s=1}^{t} E_s, \qquad M_0 = 1.$$

The statistical constraint is that, under the null, this account is not allowed to grow in expectation. Informally, the null model must view the game as fair or unfavorable to the analyst. Thus each new multiplier should have conditional null expectation at most one, given the past information. Then the capital can still go up on some sample paths, but it cannot have positive expected drift under the null. Under a well-chosen alternative, the multipliers tend to be larger than one often enough that the product may grow quickly.

For independent observations with a simple null and a fixed alternative,

$$E_t = \frac{g(X_t)}{p_0(X_t)}$$

is a per-observation likelihood-ratio e-value. The cumulative likelihood ratio is

$$M_t = \prod_{s=1}^{t} \frac{g(X_s)}{p_0(X_s)}.$$

Under $P_0$, $\mathbb{E}_{P_0}[M_t] = 1$ for each fixed $t$. A level-$\alpha$ safe test rejects when

$$M_t \geq \frac{1}{\alpha}.$$

For a fixed $t$, Markov's inequality gives

$$\mathbb{P}_{P_0}\left(M_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

The next section explains why, with the right sequential construction, the same boundary can be monitored over time.

Figure 8.3 contrasts representative capital paths under the null and under the alternative: null paths may cross upward temporarily, while a well-chosen likelihood ratio has sustained multiplicative growth under the alternative.




FIGURE 8.3. Eight simulated paths per panel of $\log_{10} M_t$ for the Gaussian likelihood-ratio capital process with $\alpha = 0.05$, $K = 120$ looks, and alternative mean $\mu_1 = 0.30$. The horizontal line is $\log_{10}(1/\alpha)$. The x-axis is the look index $t$, which equals the sample size in this one-observation-per-look simulation.

The logarithm of $M_t$ explains what a good e-value is trying to optimize. For a fixed evidence function $E_s = E(X_s)$, if the observations are i.i.d. from an alternative $Q$ and $\mathbb{E}_Q[|\log E(X_1)|] < \infty$, then

$$\frac{1}{t}\log M_t = \frac{1}{t}\sum_{s=1}^{t}\log E_s \xrightarrow{Q\text{-a.s.}} \mathbb{E}_Q[\log E_1].$$

So the long-run growth rate is expected log evidence under the alternative, subject to the null constraint $\mathbb{E}_{P_0}[E_1] \le 1$. This is the statistical content behind betting terminology: we want evidence that cannot grow in expectation under the null but grows multiplicatively under alternatives that matter. The quantity in the limit deserves a name.

**Definition 8.9: E-power**

The *e-power* of an e-value $E$ against an alternative $Q$ is $\mathbb{E}_Q[\log E]$, whenever this expectation is well defined; it may be $-\infty$ or $+\infty$, but not an undefined difference of infinities.

By Jensen's inequality, $\mathbb{E}_Q[\log E] \le \log \mathbb{E}_Q[E]$, so positive e-power is a strictly stronger requirement than $\mathbb{E}_Q[E] > 1$: the evidence must grow geometrically, not merely on average. The next proposition explains why e-power, and not ordinary power at a fixed sample size, is the right design criterion for evidence that will be multiplied.

**Proposition 8.10: E-power decides the long run**

Let $E_1, E_2, \ldots$ be i.i.d. under $Q$, each an e-value for $\mathcal{H}_0$, and let $M_t = \prod_{s=1}^{t} E_s$. Suppose the e-power $\mathbb{E}_Q[\log E_1]$ is well defined, possibly infinite. If $\mathbb{E}_Q[\log E_1] > 0$, then for every $\alpha \in (0,1)$,

$$\mathbb{P}_Q\left(M_t \ge \frac{1}{\alpha}\right) \to 1 \qquad \text{as } t \to \infty.$$

If $\mathbb{E}_Q[\log E_1] < 0$, the same probability tends to zero.

**Proof of Proposition 8.10**

By the strong law of large numbers, which applies whenever the mean exists in $[-\infty, \infty]$, $t^{-1}\log M_t \to \mathbb{E}_Q[\log E_1]$ $Q$-almost surely. If the limit is positive, then $\log M_t \to \infty$ almost surely, so $M_t$ eventually exceeds any fixed boundary; if the limit is negative, $\log M_t \to -\infty$ and the crossing probability at time $t$ vanishes. □

Recall from Section 3 that an e-value $E^*$ is log-optimal against $Q$ when $\mathbb{E}_Q[\log(E/E^*)] \le 0$ for every competing e-value $E$, and that the ratio form is needed because $\sup_E \mathbb{E}_Q[\log E]$ can be infinite. For a simple null the log-optimal choice is exactly the likelihood ratio.

**Theorem 8.11: Log-optimality of the likelihood ratio**

Let $\mathcal{H}_0 = \{P_0\}$ be simple, and let $Q$ have density $g$ with $p_0 > 0$ wherever $g > 0$. Then $E^* = g(X)/p_0(X)$ is log-optimal against $Q$, it is the unique log-optimal e-value up to $Q$-null events, and its e-power equals the KL divergence:

$$\mathbb{E}_Q[\log E^*] = D(Q \,\|\, P_0).$$

**Proof of Theorem 8.11**

For any competing e-value $E$, changing measure on the support of $Q$ gives

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] = \int_{\{g>0\}} E(x)p_0(x)\,dx \le \mathbb{E}_{P_0}[E] \le 1.$$

The inequality allows $P_0$ to put mass outside the support of $Q$. Jensen's inequality now yields $\mathbb{E}_Q[\log(E/E^*)] \le 0$, proving log-optimality. The e-power identity is the definition of $D(Q \,\|\, P_0)$.

For uniqueness, suppose $E$ is also log-optimal and put $Z = E/E^*$. Log-optimality of $E$, tested against $E^*$, gives $\mathbb{E}_Q[-\log Z] \le 0$, while the preceding calculation gives $\mathbb{E}_Q[Z] \le 1$. The positive part of $\log Z$ is integrable because $(\log Z)_+ \le Z$; the first inequality then also makes the negative part integrable. Thus

$$0 \le \mathbb{E}_Q[\log Z] \le \log \mathbb{E}_Q[Z] \le 0.$$

Equality in Jensen's inequality for the strictly concave logarithm forces $Z$ to be constant $Q$-almost surely, and the equalities force that constant to be one. □

The proof also applies when $D(Q \,\|\, P_0) = \infty$. In that case, $\mathbb{E}_Q[\log E^*] = \infty$, and a competing e-value can also have infinite e-power. Subtracting these two expectations would give the undefined expression $\infty - \infty$. Instead, we form the ratio before taking the logarithm and expectation. The bound $\mathbb{E}_Q[Z] \le 1$ ensures that the positive part of $\log Z$ has finite expectation, so $\mathbb{E}_Q[\log Z]$ is well defined, although it may equal $-\infty$. Here the likelihood ratio $E^* = g/p_0$ is finite and positive $Q$-almost surely; it is its expected logarithm that can be infinite. For example, $E^*/2$ is a valid e-value and, when $E^*$ has infinite e-power, so does $E^*/2$. Their relative log evidence nevertheless is

$$\mathbb{E}_Q\left[\log \frac{E^*/2}{E^*}\right] = -\log 2.$$

The ratio comparison therefore distinguishes them even though their separate e-powers are both infinite.

Two cautions keep e-power in perspective. First, an all-or-nothing e-value $\mathbf{1}\{p \le \alpha\}/\alpha$ equals zero with positive probability under any nondegenerate alternative, so its e-power is $-\infty$; e-power is a criterion for evidence that will be multiplied repeatedly, and it treats a single zero as fatal because a product with a zero factor never recovers. Second, for the same reason, e-power can disagree with single-use usefulness: there exist e-values that exceed 100 with probability 0.99 under $Q$ and yet have e-power $-\infty$ [147]. A useful bridge between the two scales is the fact that $\mathbb{E}_Q[E] > 1$ if and only if the shrunken e-value $1 - \lambda + \lambda E$ has positive e-power for some $\lambda \in [0, 1]$ [147]; shrinkage toward one converts average growth into geometric growth, an idea that returns twice below, in hedged products and in betting confidence sequences.

**Unknown alternatives: mixtures and plug-ins.** Example 8.6 fixes a known alternative mean $\mu_1$. In practice the alternative is unknown, and it may be two-sided. Two standard recipes adapt the likelihood ratio.

The first recipe is the *mixture method*. Choose a prior $\nu$ on the alternative parameter and define

$$E_n = \int \prod_{j=1}^{n} \frac{p_\theta(X_j)}{p_0(X_j)}\,\nu(d\theta).$$

For each fixed $\theta$ the product is an e-value for $P_0$, and an average of e-values is an e-value, so $E_n$ is valid. For the two-sided Gaussian problem, taking $\nu$ to place mass $1/2$ on each of $\pm\delta$ gives

$$E_n = \frac{1}{2}\exp\left\{\delta S_n - \frac{n\delta^2}{2}\right\} + \frac{1}{2}\exp\left\{-\delta S_n - \frac{n\delta^2}{2}\right\}, \qquad S_n = \sum_{j=1}^{n} X_j,$$

which can accumulate evidence against the null for either alternative $\mu = \delta$ or $\mu = -\delta$. A fixed $\delta$ can still fit other effect sizes poorly.

The second recipe is the *predictable plug-in*. Estimate the alternative parameter from the past only, and use the estimate in the next factor:

$$E_t = \prod_{s=1}^{t} \frac{p_{\widehat{\theta}_{s-1}}(X_s)}{p_0(X_s)},$$

where $\widehat{\theta}_{s-1}$ is computed from $X_1, \ldots, X_{s-1}$. Given the past, each factor is an ordinary likelihood ratio with a fixed parameter value, so its conditional null expectation is one; the next section turns exactly this observation into a general theory of sequential evidence. Under suitable support and moment conditions, both recipes can achieve the oracle growth rate $n^{-1}\mathbb{E}_Q[\log E_n] \to D(Q \,\|\, P_0)$. The mixture prior must put mass near the true alternative. For a plug-in, consistent parameter estimates alone do not guarantee convergence of expected log evidence: their predictive log loss must also converge in expectation [74, 193]. Here the oracle growth rate means the rate $D(Q \,\|\, P_0)$ achieved by the likelihood ratio that knows the true alternative. Figure 8.4 illustrates both recipes on Gaussian data: adapting to an unknown alternative costs only a vanishing fraction of the growth rate, while committing to a badly chosen fixed alternative costs a constant fraction forever.

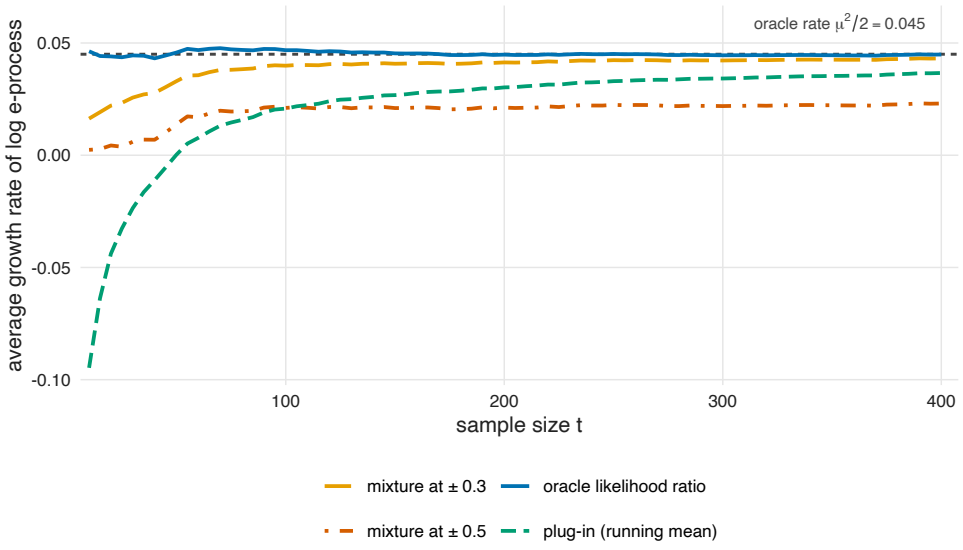



FIGURE 8.4. Average growth rate $t^{-1}\mathbb{E}_Q[\log M_t]$ of four e-processes for $H_0 : \mu = 0$, estimated from 300 simulated $N(0.3, 1)$ sequences of length 400. The oracle likelihood ratio that knows $\mu = 0.3$ sits at the rate $\mu^2/2 = 0.045$; the predictable plug-in with the running-mean estimate climbs toward the same rate; the two-point mixture at $\pm 0.3$ approaches it; and the mixture at $\pm 0.5$ settles near its theoretical rate $\delta\mu - \delta^2/2 = 0.025$.

**Comparison with the fixed-sample test.** The safe threshold and the fixed-sample Neyman-Pearson threshold solve different problems. In Example 8.6, the fixed-sample Neyman-Pearson test at sample size $n$ rejects when

$$\sum_{j=1}^{n} X_j \geq \sqrt{n}\, z_{1-\alpha}.$$

In terms of the e-value $M_n$, this is

$$M_n \geq \exp\left\{\mu_1 \sqrt{n}\, z_{1-\alpha} - \frac{n\mu_1^2}{2}\right\}.$$

The Neyman-Pearson threshold is optimal for the prespecified sample size. The safe threshold $1/\alpha$ is designed for monitoring and continuation. It may be more conservative at a fixed sample size because it is protecting a larger workflow.

Three remarks quantify and qualify this conservativeness. First, under a fixed null law $P$, the level guarantee uses two inequalities:

$$\mathbb{P}_P(E \geq 1/\alpha) \leq \alpha \mathbb{E}_P[E] \leq \alpha.$$

Markov's first inequality follows by averaging the pointwise bound $\mathbf{1}\{E \geq 1/\alpha\} \leq \alpha E$. Below the rejection threshold, equality requires $E = 0$; at or above it, equality requires $E = 1/\alpha$. Thus equality in this first step holds exactly when $E$ takes only those two values, up to events of null probability. To attain the final bound $\alpha$, we also need $\mathbb{E}_P[E] = 1$. Together, these conditions require $E = 1/\alpha$ with probability $\alpha$ and $E = 0$ otherwise. Gaussian likelihood-ratio e-values do not have this two-point form, so their fixed-sample rejection probability at the safe threshold is strictly below $\alpha$, and can be much smaller. Second, when the null distribution of the e-value is fully known, as for a simple null, nothing forbids treating $E$ as an ordinary test statistic and rejecting beyond its null $(1-\alpha)$-quantile, which can be far below $1/\alpha$; the threshold $1/\alpha$ is the price of using nothing about $E$ beyond the expectation bound. A randomized compromise needs no distributional knowledge at all: draw $U \sim \mathrm{Unif}(0,1)$ independent of the data and reject when $E \geq U/\alpha$. Then

$$\mathbb{P}_P(E \geq U/\alpha) = \mathbb{E}_P[\min\{\alpha E, 1\}] \leq \alpha\, \mathbb{E}_P[E] \leq \alpha,$$

and the first inequality is an equality whenever $\alpha E \leq 1$ almost surely; for an exact e-value bounded by $1/\alpha$, randomization therefore makes the type-I error exactly $\alpha$ [146]. Third, no expressibility is lost by working on the e-value scale: any level-$\alpha$ test with rejection region $A$ is recovered by thresholding the all-or-nothing e-value $\mathbf{1}\{X \in A\}/\alpha$ at $1/\alpha$. The difference between the two scales is not which tests are reachable but which optimality criterion is targeted, power at a prespecified sample size or growth under monitoring, and which algebra the reported evidence supports afterwards. Intermediate thresholds below $1/\alpha$ are also valid under shape restrictions on the null distribution of $E$; see Chapter 15 of Ramdas and Wang [147].

**Post-hoc validity: choosing the level after the data.** Classical testing fixes the level $\alpha$ before the data are seen, and the guarantee evaporates if $\alpha$ is chosen afterwards. E-values support a genuinely post-hoc guarantee. For each $\alpha \in (0,1)$, let $\phi_\alpha \in \{0,1\}$ denote the decision of a level-$\alpha$ test, with $\phi_\alpha = 1$ meaning rejection, and suppose $\phi_\alpha$ is nondecreasing as $\alpha$ grows.

**Definition 8.12: Post-hoc valid family**

The family $(\phi_\alpha)_{\alpha\in(0,1)}$ is *post-hoc valid* for $\mathcal{H}_0$ if

$$\mathbb{E}_P\left[\sup_{\alpha\in(0,1)} \frac{\phi_\alpha}{\alpha}\right] \le 1 \qquad \text{for every } P \in \mathcal{H}_0.$$

The ratio $\phi_\alpha/\alpha$ is the Type I error budget spent if one rejects at level $\alpha$; the supremum charges the analyst for the most aggressive level they could choose after seeing the data. For a fixed $\alpha$, the definition recovers the usual guarantee, since $\mathbb{P}_P(\phi_\alpha = 1) = \mathbb{E}_P[\phi_\alpha] \le \alpha$.

**Proposition 8.13: E-value tests are post-hoc valid**

Let $E$ be an e-value for $\mathcal{H}_0$ and let $\phi_\alpha = \mathbf{1}\{E \ge 1/\alpha\}$. Then

$$\sup_{\alpha\in(0,1)} \frac{\phi_\alpha}{\alpha} \le E \quad \text{pointwise,}$$

so the family is post-hoc valid.

**Proof of Proposition 8.13**

Fix a data realization. If $E \le 1$, then $\phi_\alpha = 0$ for every $\alpha \in (0,1)$ and the supremum is zero. If $1 < E < \infty$, then $\phi_\alpha/\alpha = 1/\alpha$ exactly when $\alpha \ge 1/E$, and the supremum over such $\alpha$ equals $E$, attained at $\alpha = 1/E$. In the remaining case $E = \infty$, the ratio is $1/\alpha$ for every $\alpha$, and its supremum is $\infty = E$. Thus the supremum is at most $E$ in every case, and taking expectations under any null $P$ finishes the proof. □

A converse also holds: up to pointwise domination, every post-hoc valid family arises from an e-value in exactly this way, with $E = \sup_\alpha \phi_\alpha/\alpha$, so testing with e-values is essentially the only route to post-hoc validity [73, 107]. The practical reading deserves emphasis. For any data-dependent choice $\widehat{\alpha} \in (0,1)$, the definition gives

$$\mathbb{E}_P\left[\frac{\phi_{\widehat{\alpha}}}{\widehat{\alpha}}\right] \le 1.$$

For example, $E = 25$ leads to rejection at either $\alpha = 0.05$ or $\alpha = 0.04$, even if the analyst chooses between these levels after observing the evidence. The displayed guarantee charges each rejection by the reciprocal of its chosen level. It does not assert that the conditional Type I error given $\widehat{\alpha} = a$ is at most $a$. An ordinary super-uniform p-value supplies fixed-level tail bounds but does not, by itself, supply this random-level normalized bound. The calibrated p-value $\min\{1, 1/E\}$ inherits it from its underlying e-value. A multiple-testing version appears in the e-BH section below.

## 5. Optional continuation and e-processes

**Route: Core**

A fixed-time e-value remains valid when its sampling plan is fixed in advance. For continuous monitoring, the process itself must be valid. An *e-process* is an adapted nonnegative process

with $M_0 \leq 1$ whose value is an e-value at every stopping time: for every null law $P$,

$$\mathbb{E}_P[M_\tau] \leq 1 \qquad \text{for every stopping time } \tau.$$

At an infinite stopping time, use $M_\infty = \liminf_{t\to\infty} M_t$. Appendix A gives the formal definition and general results. Here we establish the main construction, with $M_0 = 1$.

Suppose $E_t$ is nonnegative, available at time $t$, and satisfies the conditional e-value bound under every $P \in \mathcal{H}_0$,

$$\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \leq 1.$$

Let $\lambda_t \in [0,1]$ be $\mathcal{F}_{t-1}$-measurable. Thus $\lambda_t$ is the fraction of current capital chosen from past information and staked on the new evidence $E_t$: $\lambda_t = 0$ leaves the capital unchanged, while $\lambda_t = 1$ makes the full bet. Predictability gives

$$\mathbb{E}[1 - \lambda_t + \lambda_t E_t \,|\, \mathcal{F}_{t-1}] = 1 - \lambda_t + \lambda_t \mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \leq 1.$$

Consequently,

$$M_t = \prod_{s=1}^{t} \{1 - \lambda_s + \lambda_s E_s\}$$

satisfies the explicit conditional calculation

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}\mathbb{E}[1 - \lambda_t + \lambda_t E_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}.$$

These expectations are under an arbitrary null law $P$. Thus $M_t$ is a nonnegative test supermartingale under every null law. Ville's inequality (Theorem A.8) gives

$$\mathbb{P}_P\left(\sup_{t\geq 0} M_t \geq 1/\alpha\right) \leq \alpha, \qquad P \in \mathcal{H}_0.$$

Thus a reported stopping time may depend on the observed past, but the bet at time $t$ may not use future information or hindsight about $E_t$. The general e-process definition, its completeness for sequential tests, mixtures, and the optional-continuation derivations are in Appendix A.

## 6. E-BH: FDR Control by Aggregate Null Evidence

**Route: Core**

Now return to $m$ hypotheses. Suppose hypothesis $H_i$ has a nonnegative evidence score $E_i$. For individual e-values one often assumes $\mathbb{E}[E_i] \leq 1$ for every true null $i$. For multiple testing, the proof only needs the aggregate null-evidence condition

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \leq m. \tag{A}$$

Vectors $(E_1, \ldots, E_m)$ satisfying (A) are called *compound e-values*. The name records that validity is a property of the vector as a whole, not of its coordinates. One point of care: in a composite model, both the set of true nulls $\mathcal{I}_0 = \mathcal{I}_0(P)$ and the expectation depend on the unknown data-generating distribution $P$, and condition (A) is required for every candidate $P$. Individual e-values imply (A), since $m_0 \leq m$. So do deterministically weighted e-values $w_i E_i$ with $w_i \geq 0$ and $\sum_i w_i \leq m$. But the compound class is strictly richer than either: condition (A) constrains only the sum, so a single coordinate may have null expectation far above one, provided the excess is paid for by the other coordinates under the same $P$. The systematic study of compound e-values, and the name, are due to Ignatiadis et al. [92]; the condition itself powers the FDR theorem of Wang and Ramdas [191].

**Definition 8.14: The e-BH procedure**

Order the evidence scores from largest to smallest:

$$E_{(1)} \geq E_{(2)} \geq \cdots \geq E_{(m)}.$$

At FDR level $q$, define

$$\widehat{k} = \max\left\{k \in \{1, \ldots, m\} : E_{(k)} \geq \frac{m}{qk}\right\},$$

with $\widehat{k} = 0$ if the set is empty. Reject the hypotheses attached to the $\widehat{k}$ largest e-values.

**Theorem 8.15: E-BH FDR control**

If $E_1, \ldots, E_m$ are nonnegative and satisfy (A), then e-BH controls FDR:

$$\mathrm{FDR} \leq q.$$

In particular, if each true null $i \in \mathcal{I}_0$ satisfies $\mathbb{E}[E_i] \leq 1$, then $\mathrm{FDR} \leq (m_0/m)q \leq q$.

**Proof of Theorem 8.15**

Let $\mathcal{R}$ be the e-BH rejection set and $R = |\mathcal{R}|$. If $R = 0$, then $\mathrm{FDP} = 0$. Suppose $R > 0$. Every rejected hypothesis $i$ satisfies

$$E_i \geq E_{(R)} \geq \frac{m}{qR}.$$

Therefore, path by path,

$$\frac{\mathbf{1}\{i \in \mathcal{R}\}}{R} \leq \frac{qE_i}{m}.$$

Summing over true nulls gives

$$\mathrm{FDP} = \frac{\sum_{i \in \mathcal{I}_0} \mathbf{1}\{i \in \mathcal{R}\}}{R \vee 1} \leq \frac{q}{m} \sum_{i \in \mathcal{I}_0} E_i.$$

Taking expectations and applying (A),

$$\mathrm{FDR} \leq \frac{q}{m} \mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \leq q.$$

If $\mathbb{E}[E_i] \leq 1$ for every true null, the aggregate expectation is at most $m_0$, giving the sharper bound $(m_0/m)q$. □

The proof does not condition on other p-values, does not require independence, and does not use the PRDS positive-dependence condition of Chapter 6. The random rejection set may depend on all e-values in an arbitrary way; the pathwise inequality above absorbs that dependence before expectations are taken. This is the main reason e-BH is useful when evidence streams share subjects, features, prompts, batches, or model outputs.

The proof yields two strengthenings at no extra cost. First, it never used the exact e-BH rule, only the property that every rejected hypothesis satisfies $E_i \geq m/(qR)$. Call a procedure with this property *self-consistent* at level $q$. Every self-consistent procedure controls FDR at level $q$ by the same argument, and e-BH is the largest one: it rejects the largest set compatible with the self-consistency inequality. Smaller self-consistent sets are useful when the rejections

must also satisfy structural constraints, for example respecting a hierarchy or a graph among hypotheses. Second, the pathwise inequality in the proof is uniform in $q$, which upgrades the FDR guarantee to a post-hoc one in the sense of Proposition 8.13.

**Corollary 8.16: Post-hoc FDR control**

For each $q \in (0,1)$, let $\mathcal{R}_q$ be the rejection set of e-BH at level $q$, or of any self-consistent procedure at level $q$, applied to compound e-values $E_1, \dots, E_m$, and let $R_q = |\mathcal{R}_q|$ and $V_q = |\mathcal{R}_q \cap \mathcal{I}_0|$. Then

$$\mathbb{E}\left[\sup_{q\in(0,1)} \frac{V_q}{q(R_q \vee 1)}\right] \leq 1.$$

In particular, $\mathbb{E}[V_{\widehat{q}}/(\widehat{q}(R_{\widehat{q}} \vee 1))] \leq 1$ for every data-dependent level $\widehat{q}$: in this budget form, with the level inside the expectation, the guarantee survives choosing the level after seeing all the e-values.

**Proof of Corollary 8.16**

For each fixed $q$, the pathwise inequality established in the proof of Theorem 8.15 states

$$\frac{V_q}{R_q \vee 1} \leq \frac{q}{m} \sum_{i\in\mathcal{I}_0} E_i.$$

Divide both sides by $q$. The right-hand side no longer depends on $q$, so the same bound holds after taking the supremum over $q \in (0,1)$. Taking expectations and applying (A) completes the proof. □

Read the corollary the way a practitioner would: after computing the e-values, one may look at the entire path of e-BH rejection sets across levels and report the level whose trade-off looks best, and the reported false discovery guarantee, in the budget form $\mathbb{E}[V_{\widehat{q}}/(\widehat{q}(R_{\widehat{q}} \vee 1))] \leq 1$, still holds. The ordinary BH theorem for p-values does not provide this post-hoc guarantee; choosing its level from the observed p-values requires a separate analysis.

Figure 8.5 shows the e-BH crossing in rank space, while Figure 8.6 compares p-BH and e-BH under equicorrelated Gaussian evidence. Together they separate the algorithmic threshold rule from the moment condition that supplies dependence robustness.

## 7. How Threshold Rules Create E-Values

**Route: Advanced**

Universality. The aggregate-null-evidence proof above is the principal finite-sample guarantee. A converse theory shows that, every FDR-controlling procedure can be represented as e-BH applied to suitable compound e-values. These scores need not be individually valid e-values. The formal statement and proof are deferred to Appendix A.

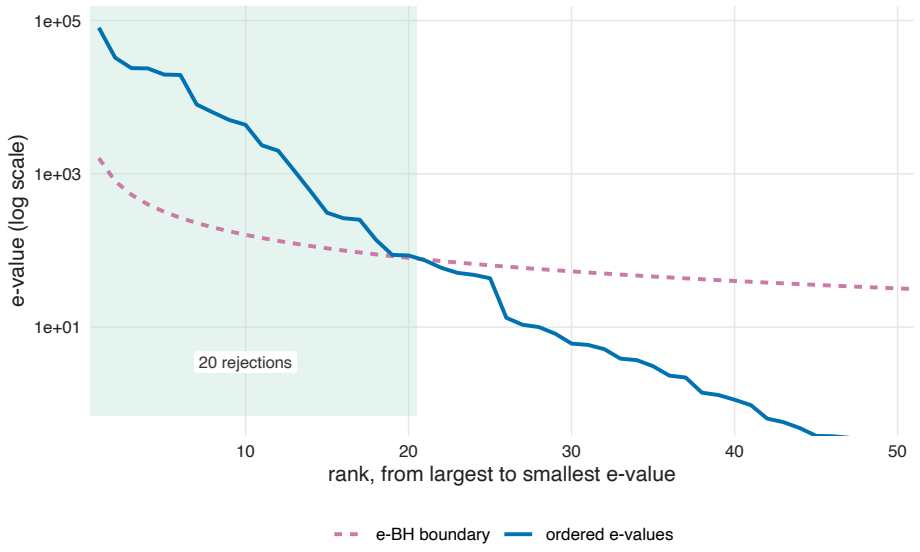



FIGURE 8.5. The e-BH rule in one simulated family of $m = 160$ hypotheses. Each hypothesis has $n = 12$ independent $N(\mu_i, 1)$ observations; data are also independent across hypotheses. The 124 null means are zero and the 36 nonnull means are $\theta = 0.9$. Writing $Z_i = \sqrt{n}\,\bar{X}_i$, we compute the likelihood-ratio e-value $E_i = \exp\{\theta\sqrt{n}\,Z_i - n\theta^2/2\}$ for testing $\mu_i = 0$ against $\mu_i = \theta$. At target $q = 0.10$, e-BH compares the ordered e-values with $m/(qk)$ over all 160 ranks; the plot shows the first 50. The largest crossing rank is $\widehat{k} = 20$. All top 20 hypotheses are rejected, including any earlier ranks whose values lie below their own rankwise boundaries.

Many p-value procedures work by choosing a data-dependent threshold and rejecting everything below it. The e-value viewpoint turns this into a research recipe:

$$\begin{array}{c}\text{threshold rule}\\ \Downarrow\\ \text{induced evidence vector}\\ \Downarrow\\ \text{aggregate null-evidence proof}\\ \Downarrow\\ \text{e-BH FDR control}\end{array}$$

This construction identifies the aggregate null-evidence bound needed for FDR control.

**BH as the first example.** Recall from Chapter 5 that BH can be written in threshold form. Let $p_1, \ldots, p_m$ be p-values. For a threshold $s \in [0, 1]$, define

$$R(s) = \sum_{i=1}^{m} \mathbf{1}\{p_i \le s\}.$$

The BH threshold can be written as

$$\widehat{t}_{\mathrm{BH}} = \sup\left\{0 < s \le 1 : \frac{ms}{R(s) \vee 1} \le q\right\}.$$

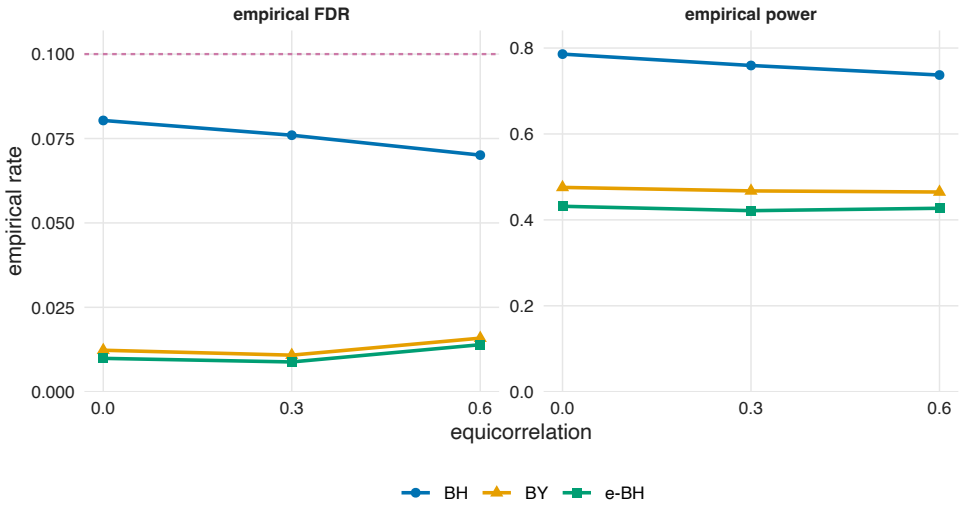



FIGURE 8.6. Empirical FDR and empirical power with $m = 400$ hypotheses, 80 nonnulls, equicorrelated Gaussian test statistics based on $n = 20$ observations with working effect $\theta = 0.65$, 650 independent simulation runs, correlations $\rho \in \{0, 0.3, 0.6\}$, and $q = 0.10$. The e-values are Gaussian likelihood ratios $\exp\{\theta\sqrt{n}\,Z - n\theta^2/2\}$. The e-BH guarantee applies at each correlation because these likelihood ratios satisfy the null expectation bound.

When the set is empty, set $\widehat{t}_{\mathrm{BH}} = 0$. The BH rejection set is

$$\mathcal{R}_{\mathrm{BH}} = \{i : p_i \le \widehat{t}_{\mathrm{BH}}\}.$$

On the event $\widehat{t}_{\mathrm{BH}} > 0$, define the induced evidence score

$$E_i^{\mathrm{BH}} = \frac{1}{\widehat{t}_{\mathrm{BH}}}\mathbf{1}\{p_i \le \widehat{t}_{\mathrm{BH}}\}.$$

If $\widehat{t}_{\mathrm{BH}} = 0$, define $E_i^{\mathrm{BH}} = 0$ for all $i$.

**Proposition 8.17: The BH rejection set as e-BH**

Applying e-BH at level $q$ to the induced vector $(E_i^{\mathrm{BH}})_{i=1}^m$ gives the same rejection set as BH.

**Proof of Proposition 8.17**

If $\widehat{t}_{\mathrm{BH}} = 0$, both procedures reject nothing. Suppose $\widehat{t}_{\mathrm{BH}} > 0$ and let $\widehat{k} = R(\widehat{t}_{\mathrm{BH}})$. If $\widehat{k} = 0$, then all induced e-values are zero and both procedures reject nothing. Now assume $\widehat{k} > 0$. By the BH threshold definition,

$$\frac{m\widehat{t}_{\mathrm{BH}}}{\widehat{k}} \le q, \qquad \text{so} \qquad \frac{1}{\widehat{t}_{\mathrm{BH}}} \ge \frac{m}{q\widehat{k}}.$$

The $\widehat{k}$ BH rejections have evidence $1/\widehat{t}_{\mathrm{BH}}$, and all non-rejections have evidence 0. Hence the $\widehat{k}$th largest induced e-value crosses the e-BH boundary, while no non-rejected hypothesis can be added because its induced e-value is zero. Thus e-BH selects exactly $\mathcal{R}_{\mathrm{BH}}$. □

This proposition is algebraic. It does not prove BH validity. Validity comes from the budget question:

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i^{\mathrm{BH}}\right] \le m?$$

Under the usual independent-null assumptions for BH, this aggregate bound can be proved by a reverse-time martingale or leave-one-out argument. The intuition is as follows. If a true-null p-value is rejected at a data-dependent threshold $\widehat{t}_{\mathrm{BH}}$, it contributes $1/\widehat{t}_{\mathrm{BH}}$. For a fixed threshold $s$, a super-uniform null p-value has expected contribution at most $\mathbb{P}(p_i \le s)/s \le 1$. The technical argument shows that the same bound survives when $s$ is the BH threshold chosen from the data, because replacing one rejected true-null p-value by zero does not change the final number of BH rejections in the leave-one-out comparison.

There is a second route to the same aggregate bound that assumes positive dependence instead of independence. Recall from Chapter 6 that PRDS is the positive-dependence condition under which BH controls FDR. On the event $\widehat{k} \ge 1$, the BH threshold is $\widehat{t}_{\mathrm{BH}} = q\widehat{k}/m$, so the induced evidence of Proposition 8.17 can be written

$$E_i^{\mathrm{BH}} = \frac{m}{q\widehat{k}}\,\mathbf{1}\left\{p_i \le \frac{q\widehat{k}}{m}\right\}.$$

The dependency-control lemma of Blanchard and Roquain [27] can be stated in the form needed here. Let $p_i$ be a super-uniform true-null p-value and let $R = R(\boldsymbol{p}) \in \{0, 1, \dots, m\}$ be coordinatewise nonincreasing in the p-value vector. Under independence or PRDS, for every $c > 0$,

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le cR\}}{R \vee 1}\right] \le c.$$

The fixed-threshold inequality $\mathbb{P}(p_i \le s)/s \le 1$ is therefore preserved when the threshold is the random, monotone quantity $cR$; PRDS is exactly what controls this dependence between $p_i$ and $R$. For BH take $R = \widehat{k}$ and $c = q/m$. The lemma gives

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le q\widehat{k}/m\}}{\widehat{k} \vee 1}\right] \le \frac{q}{m}.$$

Multiplying by $m/q$, and using the convention $E_i^{\mathrm{BH}} = 0$ when $\widehat{k} = 0$, yields $\mathbb{E}[E_i^{\mathrm{BH}}] \le 1$. Thus the induced vector consists of individual e-values, the aggregate bound holds with room to spare, and Theorem 8.15 returns FDR $\le (m_0/m)q$ for BH under PRDS. A classical theorem becomes a short consequence of the e-value viewpoint plus one dependence lemma.

**Benjamini-Yekutieli as e-BH on calibrated p-values.** The second example needs no dependence assumption at all, and it connects the calibrators of this chapter to the arbitrary-dependence correction of Chapter 5. Recall Theorem 5.6: the BY procedure, which is BH run at the reduced level $q/H_m$ with $H_m = \sum_{k=1}^m 1/k$ the harmonic number, controls FDR at level $q$ under arbitrary dependence. Fix $q$ and $m$, and define a step function that rounds each p-value onto the BY grid:

$$f_{\mathrm{BY}}(p) = \frac{m}{qk} \quad \text{for } p \in \left(\frac{q(k-1)}{mH_m}, \frac{qk}{mH_m}\right], \quad k = 1, \dots, m,$$

with $f_{\mathrm{BY}}(0) = m/q$ and $f_{\mathrm{BY}}(p) = 0$ for $p > q/H_m$. This is a nonincreasing function with

$$\int_0^1 f_{\mathrm{BY}}(u)\,du = \sum_{k=1}^m \frac{m}{qk}\cdot\frac{q}{mH_m} = \frac{1}{H_m}\sum_{k=1}^m \frac{1}{k} = 1,$$

so by Proposition 8.4 it is a calibrator: $E_i = f_{\mathrm{BY}}(p_i)$ is an e-value for each hypothesis whose p-value is valid, with no assumption on the dependence among the p-values.

**Proposition 8.18: BY as e-BH**

e-BH at level $q$ applied to $E_i = f_{\mathrm{BY}}(p_i)$, $i = 1, \dots, m$, rejects exactly the BY rejection set at level $q$. In particular, Theorem 8.15 implies FDR $\le (m_0/m)q$ under arbitrary dependence, recovering Theorem 5.6.

**Proof of Proposition 8.18**

By construction, $f_{\mathrm{BY}}(p) \ge m/(qk)$ if and only if $p \le qk/(mH_m)$. Since $f_{\mathrm{BY}}$ is nonincreasing, the $k$th largest e-value corresponds to the $k$th smallest p-value, so

$$E_{(k)} \ge \frac{m}{qk} \quad \Longleftrightarrow \quad p_{(k)} \le \frac{qk}{mH_m},$$

where $p_{(k)}$ here denotes the $k$th smallest p-value. The largest $k$ satisfying the left condition, which defines e-BH, therefore equals the largest $k$ satisfying the right condition, which defines BY, and the rejected sets coincide. The FDR bound is Theorem 8.15 applied to individual e-values. □

The factor $H_m \approx \log m$ arises from calibrating p-values into e-values in this construction. Directly constructed e-values can be used in e-BH under arbitrary dependence without this factor. When only p-values are available, BY is optimal within this calibration construction [147, 191].

**The general threshold template.** The BH example suggests a broader template. Suppose a procedure has candidate thresholds $s \in \mathcal{T}$. At threshold $s$, let

$$R_i(s) \in \{0, 1\}$$

indicate whether hypothesis $i$ would be rejected, and let

$$R(s) = \sum_{i=1}^{m} R_i(s).$$

Let $\widehat{M}(s)$ be the procedure's estimate or upper bound for the number of false discoveries at threshold $s$. The procedure chooses

$$\hat{t} = \sup\left\{ s \in \mathcal{T} : \frac{\widehat{M}(s)}{R(s) \vee 1} \le q \right\}$$

and rejects all $i$ with $R_i(\hat{t}) = 1$. We assume throughout that the feasible set has a largest element; if it is empty, the procedure makes no rejections. One convenient way to ensure attainment is to use a finite candidate grid, as in the generalized BH construction of Chapter 5. Right-continuity alone is not sufficient for an arbitrary threshold rule.

The induced evidence formula requires a positive false-discovery budget when there is at least one rejection. We therefore assume

$$R(\hat{t}) > 0 \quad \Longrightarrow \quad \widehat{M}(\hat{t}) > 0.$$

If a candidate rule can have $\widehat{M}(\hat{t}) = 0$ and $R(\hat{t}) > 0$, it must be modified before using this template; common fixes are to add a $+1$ correction to the budget or to disallow thresholds with zero budget and positive rejections.

The induced evidence vector is

$$E_i = \frac{mR_i(\widehat{t})}{\widehat{M}(\widehat{t})},$$

with the convention $E_i = 0$ for all $i$ when $R(\widehat{t}) = 0$. A rejected hypothesis receives evidence equal to $m$ divided by the estimated false-discovery budget; a non-rejected hypothesis receives zero.

**Proposition 8.19: Threshold rule to e-BH**

For a threshold rule of the form above, e-BH applied to the induced vector $E_i = mR_i(\widehat{t})/\widehat{M}(\widehat{t})$ has exactly the same rejection set as the threshold rule. If the induced vector also satisfies

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m,$$

then e-BH controls FDR at level $q$.

**Proof of Proposition 8.19**

Let $\widehat{r} = R(\widehat{t})$. If $\widehat{r} = 0$, there is nothing to prove. If $i$ is rejected by the threshold rule, then

$$E_i = \frac{m}{\widehat{M}(\widehat{t})}.$$

The threshold definition gives

$$\frac{\widehat{M}(\widehat{t})}{\widehat{r}} \le q, \qquad \text{so} \qquad E_i \ge \frac{m}{q\widehat{r}}.$$

Thus at least $\widehat{r}$ hypotheses cross the e-BH boundary at rank $\widehat{r}$. By definition, the remaining hypotheses have $R_i(\widehat{t}) = 0$ and hence induced evidence zero, so e-BH cannot reject beyond rank $\widehat{r}$. The rejection sets are equal. The FDR statement is exactly Theorem 8.15. □

Table 8.2 gives examples of the template. Each row should be read in the same way: the middle column supplies the false-discovery budget $\widehat{M}(s)$, the last column supplies the fixed-threshold rejection rule $R_i(s)$, and Proposition 8.19 then gives the induced e-values once the data choose $\widehat{t}$.

Raw p-values are not always the best ranking scale. A researcher may want to borrow information across hypotheses, as in empirical-Bayes or local-FDR ranking, or use external information such as covariates, prior biological knowledge, group structure, or weights learned from independent data. A simple way to represent this is to rank hypothesis $i$ by a transformed score. Before using the p-value $p_i$, choose a function $\psi_i$ and define

$$Z_i = \psi_i(p_i).$$

Assume $\psi_i$ is nondecreasing. Then a smaller p-value gives a smaller, or at least no larger, score, so small $Z_i$ still means stronger evidence. A score-threshold rule therefore rejects $i$ when $Z_i \le s$, that is, when $R_i(s) = \mathbf{1}\{\psi_i(p_i) \le s\}$.

Validity depends on how the score transformation is chosen. In this simple template, $\psi_i$ is fixed before the procedure uses $p_i$. If the score transformation is learned from other p-values or from the whole dataset, then the validity proof must account for that learning step, for example through sample splitting, masking, independence, or a direct proof of the aggregate null-evidence bound.

After the ranking score is fixed, we still need a budget $\widehat{M}(s)$ for the number of false rejections. One simple choice is $mg(s)$, where the curve $g$ is chosen before testing. Here $g(s)$ is a proposed allowance per hypothesis for the probability of a false rejection at a fixed cutoff $s$; multiplying by $m$ gives a budget for the expected total number of false rejections. For example, with raw p-values, each true null satisfies $\mathbb{P}(p_i \le s) \le s$, so

$$\mathbb{E}[V(s)] = \sum_{i \in \mathcal{I}_0} \mathbb{P}(p_i \le s) \le m_0 s \le ms,$$

where $V(s)$ counts the false rejections at cutoff $s$. Thus BH uses $g(s) = s$. For transformed scores, the chosen curve must reflect their null rejection probabilities. A bound at each fixed cutoff does not by itself justify choosing the cutoff from the data: the induced evidence must still satisfy the aggregate null-evidence bound.

Another choice uses the mirror idea from Chapter 5: estimate the left-tail null count by looking at very large p-values. For an exactly uniform null p-value, intervals of equal width near zero and one have equal probability. Under assumptions that preserve an appropriate comparison between these tails, the right-tail count can therefore estimate the left-tail null count. A large p-value $p_j$ becomes small after the transformation $1 - p_j$, so the count

$$1 + \sum_{j=1}^{m} \mathbf{1}\{\psi_j(1 - p_j) \le s\}$$

counts right-tail observations on the same score scale used for left-tail rejections. The $+1$ is a conservative correction that keeps the estimated budget from being zero. This right-tail idea is useful only when the model or assumptions justify the comparison between the two tails; it is not a free validity guarantee. The raw-p-value right-tail row in the table is the special case $\psi_i(u) = u$.

TABLE 8.2. Self-contained examples of the threshold template. The customized rows use the score $Z_i = \psi_i(p_i)$ defined in the text above the table. In the Storey row, $\widehat{\pi}_0^{\lambda}$ is the null-proportion estimator at tuning parameter $\lambda$, recalled from Chapter 6.

| Procedure | $\widehat{M}(s)$ | $R_i(s)$ |
|---|---|---|
| BH | $ms$ | $\mathbf{1}\{p_i \le s\}$ |
| BY | $mH_m s$ | $\mathbf{1}\{p_i \le s\}$ |
| Storey | $m\widehat{\pi}_0^{\lambda} s$ | $\mathbf{1}\{p_i \le s\}$, $s \le \lambda$ |
| Raw p-values, right-tail budget | $1 + \sum_{j=1}^{m} \mathbf{1}\{p_j \ge 1 - s\}$ | $\mathbf{1}\{p_i \le s\}$, $s < 1/2$ |
| Transformed score, fixed budget | $mg(s)$ | $\mathbf{1}\{\psi_i(p_i) \le s\}$ |
| Transformed score, right-tail budget | $1 + \sum_{j=1}^{m} \mathbf{1}\{\psi_j(1 - p_j) \le s\}$ | $\mathbf{1}\{\psi_i(p_i) \le s\}$ |

That such a translation must exist, for any FDR-controlling rule whatsoever, is the content of the universality theorem, Theorem A.12 [92, 191]. The useful message, developed by Li and Zhang, is that for threshold rules the translation is explicit: the induced e-values come in closed form once the procedure is written through $R_i(s)$ and $\widehat{M}(s)$ [121, 122]. The remaining mathematical task is to prove the aggregate null-evidence bound

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \le m.$$

Different constructions need different arguments: independent-null leave-one-out arguments for BH-like rules, symmetry or conservativeness arguments for right-tail estimates, and problem-specific arguments for transformed scores.

## 8. Combining and Assembling E-Values

**Route: Advanced**

The e-BH proof uses only aggregate null evidence. This makes it possible to combine several evidence sources without redoing the whole FDR proof each time.

In one sentence, *combining* merges several evidence values for the same hypotheses, whereas *assembling* rescales and pools evidence from disjoint hypothesis families into one final rejection list.

Before turning to multiple testing, consider the one-hypothesis version of the question. Several teams hand us e-values $E_1, \dots, E_L$ for the same null, with unknown dependence; the analyses may share data, subjects, or models. Which functions of them are again e-values?

Averaging always works. For fixed weights $\lambda_\ell \geq 0$ with $\sum_\ell \lambda_\ell \leq 1$,

$$\mathbb{E}_P\left[\sum_{\ell=1}^{L} \lambda_\ell E_\ell\right] = \sum_{\ell=1}^{L} \lambda_\ell \, \mathbb{E}_P[E_\ell] \leq 1$$

by linearity, no matter how the $E_\ell$ depend on one another. Multiplication does not. The product of arbitrarily dependent e-values is not an e-value, and the failure is immediate: take $E_1 = E_2 = E$, where $E$ equals 2 with probability $1/2$ and 0 otherwise, so that each factor is an e-value, while $\mathbb{E}[E_1 E_2] = \mathbb{E}[E^2] = 2$. Products require independence, under which $\mathbb{E}[\prod_\ell E_\ell] = \prod_\ell \mathbb{E}[E_\ell] \leq 1$, or the sequential structure of Proposition A.6. Taking the maximum also fails, for the same reason that $1/p$ failed earlier: reporting the best of several evidence numbers, chosen after seeing them, is a hindsight bet.

Under arbitrary dependence, weighted averages are the canonical safe choice. Call a Borel-measurable, coordinatewise nondecreasing function $F : [0, \infty)^L \to [0, \infty)$ an *e-merging function* if $F(E_1, \dots, E_L)$ is an e-value whenever $E_1, \dots, E_L$ are arbitrarily dependent e-values. Every e-merging function is dominated pointwise by one of the form

$$F_{\boldsymbol{\lambda}}(e_1, \dots, e_L) = \lambda_0 + \sum_{\ell=1}^{L} \lambda_\ell e_\ell, \qquad \lambda_\ell \geq 0, \quad \lambda_0 + \sum_{\ell=1}^{L} \lambda_\ell = 1,$$

a weighted mean that may place some weight $\lambda_0$ on the constant one, and the admissible e-merging functions are exactly these $F_{\boldsymbol{\lambda}}$ [147, 188].

A merging rule is *symmetric* if reordering its inputs leaves its output unchanged. If we want such a rule, the arithmetic mean $A(\boldsymbol{e}) = L^{-1} \sum_{\ell=1}^{L} e_\ell$ has a useful guarantee: whenever another symmetric e-merging rule reports a value greater than one, the arithmetic mean reports at least that much evidence. This property is called *essential domination*. The restriction to outputs above one matters. For example, the rule $G = (1 + A)/2$ is a valid symmetric e-merging rule, but if all the inputs equal 0.2, it reports 0.6, which exceeds their arithmetic mean 0.2. Both outputs are below one, so this does not contradict the guarantee. The geometric and harmonic means satisfy a stronger comparison: neither ever exceeds the arithmetic mean, for any nonnegative inputs. For the harmonic mean, use the value zero if any input is zero.

For multiple testing there are two different problems. The practical question is whether the same hypothesis appears in every evidence source.

Before separating the two problems, define the error rates used when hypotheses are split into groups. Suppose

$$\mathcal{G}_1, \dots, \mathcal{G}_L$$

are disjoint groups of hypotheses. If a procedure returns a rejection set $\mathcal{R}_\ell \subseteq \mathcal{G}_\ell$ inside group $\ell$, define

$$R_\ell = |\mathcal{R}_\ell|, \qquad V_\ell = |\mathcal{R}_\ell \cap \mathcal{I}_0|, \qquad \mathrm{FDR}_\ell = \mathbb{E}\left[\frac{V_\ell}{R_\ell \vee 1}\right].$$

Here *groupwise* refers to false discoveries *within* each family. It differs from the group-level FDR of Chapter 7, where each rejected group counts as one discovery. Controlling groupwise FDR at level $q$ means $\mathrm{FDR}_\ell \le q$ for every group $\ell$. If a procedure instead returns one pooled rejection set $\mathcal{R}^{\mathrm{pool}}$ across all groups, define

$$R^{\mathrm{pool}} = |\mathcal{R}^{\mathrm{pool}}|, \qquad V^{\mathrm{pool}} = |\mathcal{R}^{\mathrm{pool}} \cap \mathcal{I}_0|, \qquad \mathrm{FDR}^{\mathrm{pool}} = \mathbb{E}\left[\frac{V^{\mathrm{pool}}}{R^{\mathrm{pool}} \vee 1}\right].$$

Controlling overall FDR means $\mathrm{FDR}^{\mathrm{pool}} \le q$. In an assembling problem, the target is this pooled guarantee. The groupwise definition is still useful because the assumptions we impose on each family's e-values also imply groupwise FDR control if the analyst reports family-specific e-BH lists.

*Combining* means that several analyses speak about the same list $H_1, \dots, H_m$. For example, a genomics study may test the same $m$ genes using a model-based e-value, a permutation-based e-value, and an e-value that borrows pathway information. A clinical monitoring problem may test the same $m$ endpoints using evidence from a primary model and evidence from a robust sensitivity analysis. The final output is still one rejection list among the original $m$ hypotheses; the question is how much weight to give each evidence source.

*Assembling* means that different analyses cover different hypotheses, and the researcher wants one pooled list at the end. For example, three laboratories may screen disjoint gene panels, or several hospitals may monitor different sets of safety signals. Each laboratory or hospital can construct valid evidence for its own family, but the final scientific action may require ranking all discoveries together for follow-up. The bookkeeping must then account for the family sizes $m_\ell$ and for the fact that true nulls may be concentrated in the families that receive large weights.

**Combining procedures on the same hypotheses.** Suppose $L$ procedures all analyze the same $m$ hypotheses. Procedure $\ell$ produces a nonnegative evidence vector

$$E_1^{(\ell)}, \dots, E_m^{(\ell)}$$

and has already been proved aggregate-valid, meaning that its total expected true-null evidence is at most the number of hypotheses:

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i^{(\ell)}\right] \le m, \qquad \ell = 1, \dots, L.$$

This condition already gives a guarantee for each procedure separately: e-BH applied to $E_1^{(\ell)}, \dots, E_m^{(\ell)}$ controls FDR for procedure $\ell$'s rejection list. The useful question is whether we can combine the $L$ vectors first and then run e-BH once. The next proposition says yes: with suitable weights, the combined vector is still aggregate-valid, so e-BH on the combined e-values controls FDR for one final rejection list.

**Proposition 8.20: Combining aggregate-valid evidence**

Let

$$E_i = \sum_{\ell=1}^{L} w_{\ell i} E_i^{(\ell)},$$

with deterministic weights $w_{\ell i} \geq 0$. If

$$\sum_{\ell=1}^{L} \max_{1 \leq i \leq m} w_{\ell i} \leq 1,$$

then the combined vector satisfies

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \leq m.$$

Consequently, e-BH applied to $E_1, \ldots, E_m$ controls FDR at level $q$.

**Proof of Proposition 8.20**

Using linearity of expectation and nonnegative weights,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L} \sum_{i \in \mathcal{I}_0} w_{\ell i} \mathbb{E}[E_i^{(\ell)}] \\
&\leq \sum_{\ell=1}^{L} \left(\max_i w_{\ell i}\right) \sum_{i \in \mathcal{I}_0} \mathbb{E}[E_i^{(\ell)}] \\
&\leq \sum_{\ell=1}^{L} \left(\max_i w_{\ell i}\right) m \leq m.
\end{aligned}$$

The FDR conclusion follows from Theorem 8.15. □

The max-weight condition is the price of not knowing which hypotheses are true nulls. An average weight can look small over all hypotheses while placing too much weight on the true nulls. The maximum prevents this, so the combined e-values remain valid for a single e-BH run.

**Assembling disjoint hypothesis families.** Now suppose the hypotheses are split across disjoint families

$$\mathcal{G}_1, \ldots, \mathcal{G}_L, \qquad |\mathcal{G}_\ell| = m_\ell, \qquad \sum_{\ell=1}^{L} m_\ell = m.$$

Family $\ell$ produces evidence $E_i^{(\ell)}$ only for $i \in \mathcal{G}_\ell$, and assume that within that family

$$\mathbb{E}\left[\sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} E_i^{(\ell)}\right] \leq m_\ell.$$

This displayed bound has an immediate consequence: if family $\ell$ runs e-BH only on $\{E_i^{(\ell)} : i \in \mathcal{G}_\ell\}$, then Theorem 8.15 with $m = m_\ell$ gives $\mathrm{FDR}_\ell \leq q$. The assembling step builds one pooled list across all families. To do this, choose fixed nonnegative weights, form $E_i = w_{\ell i} E_i^{(\ell)}$ for $i \in \mathcal{G}_\ell$, and run e-BH once on all $m$ assembled e-values. The next proposition gives the condition that makes this pooled run control $\mathrm{FDR}^{\mathrm{pool}}$.

**Proposition 8.21: Assembling disjoint families**

For $i \in \mathcal{G}_\ell$, set

$$E_i = w_{\ell i} E_i^{(\ell)},$$

with deterministic weights $w_{\ell i} \geq 0$. If

$$\sum_{\ell=1}^{L} m_\ell \max_{i \in \mathcal{G}_\ell} w_{\ell i} \leq m,$$

then the pooled vector satisfies

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \leq m.$$

Thus e-BH controls $\mathrm{FDR}^{\mathrm{pool}}$ for the pooled rejection list.

**Proof of Proposition 8.21**

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L} \sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} w_{\ell i} \mathbb{E}[E_i^{(\ell)}] \\
&\leq \sum_{\ell=1}^{L} \left(\max_{i \in \mathcal{G}_\ell} w_{\ell i}\right) \sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} \mathbb{E}[E_i^{(\ell)}] \\
&\leq \sum_{\ell=1}^{L} m_\ell \max_{i \in \mathcal{G}_\ell} w_{\ell i} \leq m.
\end{aligned}$$

□

The proposition proves the assembling result: the one pooled e-BH run controls overall FDR for the pooled rejection list. The family-level aggregate bound imposed before the proposition has the separate consequence that family-specific e-BH runs would control $\mathrm{FDR}_\ell$. These are different rejection lists; the pooled guarantee does not require reporting the family-specific lists.

Figure 8.7 displays the distinction: combining aligns several evidence vectors coordinate by coordinate, while assembling concatenates separately constructed families before one pooled e-BH run.

**E-values as weights in p-value procedures.** A hybrid setting arises when both evidence scales are present for the same hypotheses: p-values $p_1, \ldots, p_m$ from confirmatory data, and e-values $E_1, \ldots, E_m$ from an earlier study, an external dataset, or prior structural knowledge. A natural combination divides each p-value by its e-value,

$$p_i^* = \min\left\{1, \frac{p_i}{E_i}\right\},$$

with $p_i^* = 1$ when $E_i = 0$, and runs BH at level $q$ on $p_1^*, \ldots, p_m^*$. If the p-value vector satisfies the PRDS condition of Chapter 6 and is independent of the e-value vector, this procedure controls FDR at level $(m_0/m)q$ [91]. Two features distinguish it from the weighted BH rules of Chapter 5. Classical weighted BH requires weights that average to one. Here, the null expectation bounds $\mathbb{E}[E_i] \leq 1$ replace that normalization requirement. The e-values for false nulls may have large means, and dependence within the e-value vector is unrestricted. Dividing PRDS p-values by independent random weights does not obviously preserve PRDS, so the classical weighted-BH

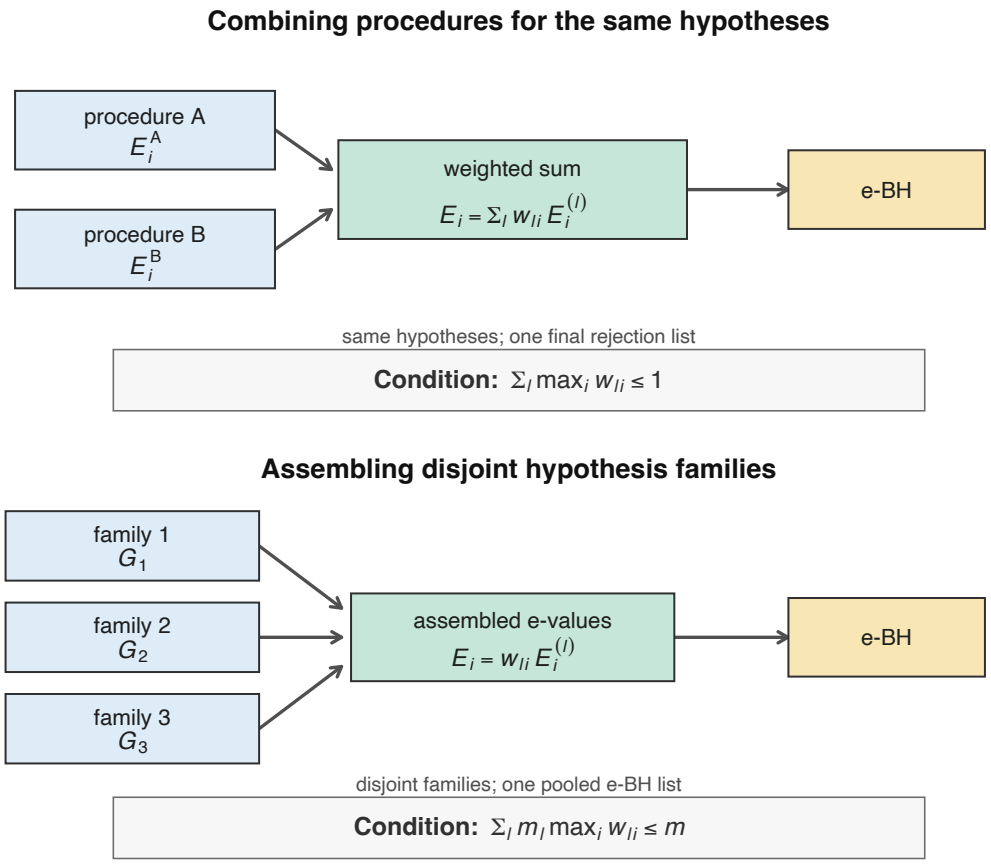



FIGURE 8.7. Combining evidence from several procedures on the same hypotheses and assembling evidence from disjoint hypothesis families. In assembling, the main output is one pooled e-BH list; family-level aggregate bounds also justify separate family lists when reported. The max-weight conditions protect against unknown true-null placement.

guarantee does not directly apply; the proof conditions on the e-values and controls the null contributions with $\mathbb{E}[E_i] \le 1$.

## 9. Assumptions in Plain Language

**Route: Core**

The e-value condition is a null expectation bound. For simple nulls, likelihood ratios give canonical examples. For composite nulls, the bound must hold under every distribution in the null model. A Bayes factor averaged over a null prior is not enough unless it also satisfies this requirement. A p-to-e calibrator must be fixed before seeing the p-value; choosing the conversion that looks best afterwards can invalidate it. Conversely, every valid e-value yields the p-value $\min\{1, 1/E\}$.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | For each e-value, verify nonnegativity and the derivation of its uniform null expectation bound. For a sequential product, use the analysis log to verify that every bet used only past data and that every factor has conditional null mean at most one. |
| **What these data cannot show** | Observed e-values cannot establish that the null law was correct, that the bets were predictable, or that no unrecorded looks or model changes occurred. |
| **Safer fallback** | If the only unsupported part is optional continuation, return to one prespecified analysis at a fixed sample size with a valid fixed-sample test. This does not repair a misspecified null model. |

Safe testing requires a process-level guarantee. A fixed-sample e-value can be used once at a prespecified time. An e-process can be monitored, stopped, and continued because its value at every stopping time remains an e-value. For the product construction in this chapter, each new factor must have conditional null expectation at most one given all the information already available. The choice of bet must use only that past information. Marginal validity of each factor alone does not justify multiplying factors that share data.

The e-BH procedure controls FDR under arbitrary dependence because its proof reduces FDP to aggregate null evidence before taking expectations. The required bound is $\mathbb{E}[\sum_{i\in\mathcal{I}_0} E_i] \le m$, under every distribution in the model. Individual e-values satisfy it, but the compound e-values of this chapter need only satisfy the bound as a vector. Dependence can still affect how valid evidence is constructed and how much power it has. The post-hoc guarantee also needs careful interpretation: for a level chosen from the data, it bounds the expectation of the ratio of FDP to that level, as in Corollary 8.16. It is not a conditional FDR guarantee given the chosen level.

To establish FDR control for a threshold rule, one must prove that its induced evidence vector satisfies $\mathbb{E}[\sum_{i\in\mathcal{I}_0} E_i] \le m$.

Combining and assembling require the fixed, nonnegative weights and max-weight bounds stated in their propositions. These bounds prevent overweighting the unknown true nulls. Combining addresses the same hypotheses across procedures; assembling produces one pooled list from disjoint families. The pooled FDR guarantee does not imply a separate FDR guarantee within each family's portion of that list. Finally, using e-values as weights in BH requires the stated independence between the p-value and e-value vectors and the PRDS condition on the p-values.

Before applying a result, specify the null model, the information available when each choice is made, and the rejection list or monitoring rule to which the guarantee will apply. Then check the expectation and weighting conditions used in that result's proof.

## 10. Bibliographic Notes

**Route: Core**

A unified book-length treatment of e-values, from which several of this chapter's refinements are drawn, is Ramdas and Wang [147]. E-values, calibration, and merging are developed by Vovk and Wang [188]; the calibrator concept goes back to Vovk [182], and the term e-power and its theory are developed in Chapter 3 of Ramdas and Wang [147]. Safe testing and optional

continuation are treated by Grünwald et al. [74]; the betting interpretation is emphasized by Shafer [160] and connected to the broader game-theoretic view by Ramdas et al. [151]. The numeraire, its unconditional existence and uniqueness, and the general theory of the reverse information projection are due to Larsson et al. [110], with information-theoretic precursors in Li [123] and Csiszár and Tusnády [45]. Universal inference is due to Wasserman et al. [192]; the power cost of splitting is studied by Dunn et al. [55]. The general definition of e-processes and their role in anytime-valid inference are developed by Howard et al. [88], Ramdas et al. [149], and Grünwald et al. [74]; the strict gap between e-processes and test supermartingales, illustrated by exchangeability testing, is established by Ramdas et al. [150]. Post-hoc validity and the equivalence between post-hoc valid testing and e-values are developed by Grünwald [73] and Koning [107]; the randomized Markov inequality is studied by Ramdas and Manole [146].

The e-BH procedure and its arbitrary-dependence FDR proof are due to Wang and Ramdas [191], including the observation that BY is e-BH on calibrated p-values. Compound e-values are named and systematically studied by Ignatiadis et al. [92], who prove the universality of e-BH; e-values as unnormalized weights in p-value procedures are analyzed by Ignatiadis et al. [91]; derandomization through averaged compound e-values is developed by Ren and Barber [152]. The super-uniformity lemma behind the PRDS remark is due to Blanchard and Roquain [27]. The threshold-to-e-BH viewpoint and the aggregation and assembling ideas are based on Li and Zhang [122] and Li and Zhang [121]. Randomized and closed variants of e-BH, which uniformly improve it in certain regimes, are surveyed in Chapter 9 of Ramdas and Wang [147]. Approximate and asymptotic e-values, including the universal t-test example, are developed in Chapter 10 of Ramdas and Wang [147] and in Section 6.2 of Wang and Ramdas [191].

Online FDR control begins with alpha-investing [64] and includes significance levels based on recent discovery (LORD) [98] and serial estimate of the alpha fraction that is futilely rationed on true null hypotheses (SAFFRON) [148]. The levels based on number of discoveries (LOND) rule has an e-value analogue (e-LOND), developed by Xu and Ramdas [194]. Confidence sequences in the modern nonasymptotic form are developed by Howard et al. [88]; betting confidence sequences for bounded means are developed by Waudby-Smith and Ramdas [193]. E-confidence intervals and false coverage rate (FCR) control through e-values, the e-BY procedure, are due to Xu et al. [195].

## 11. Exercises

**Route: Core**

### Basic.

**Exercise 8.22: Calibrator family**

Using Proposition 8.4, verify that $f_\kappa(p) = \kappa p^{\kappa-1}$ for $\kappa \in (0,1)$, $f(p) = p^{-1/2} - 1$, and $f(p) = -\log p$ are calibrators. Check nonnegativity and monotonicity as well as the integral condition. Compute the e-value assigned to $p = 0.01$ by the power calibrator $f_\kappa$ with $\kappa = 1/2$.

**Exercise 8.23: Expected log evidence in the Gaussian model**

Let $X_1, \dots, X_n$ be independent $N(\mu_1, 1)$ variables, with $\mu_1 > 0$. The Gaussian likelihood-ratio e-value in Example 8.6 is

$$E_n = \exp\left\{\mu_1 \sum_{j=1}^{n} X_j - \frac{n\mu_1^2}{2}\right\}.$$

Compute $\mathbb{E}_{\mu_1}[\log E_n]$, showing the expectation calculation.

## Intermediate.

**Exercise 8.24: Combining procedures**

Proposition 8.20 proves validity under a condition on maximum weights. Construct a counterexample showing that the weaker average-weight condition

$$\frac{1}{m}\sum_{i=1}^{m} w_{1i} + \frac{1}{m}\sum_{i=1}^{m} w_{2i} \le 1$$

does not suffice. Specify a null set $\mathcal{I}_0$, two nonnegative evidence vectors satisfying

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i^{(\ell)}\right] \le m, \qquad \ell = 1, 2.$$

Only this aggregate condition is required for each vector; its null coordinates need not individually have expectations at most one. Choose fixed nonnegative weights such that $E_i = w_{1i}E_i^{(1)} + w_{2i}E_i^{(2)}$ violates $\mathbb{E}[\sum_{i\in\mathcal{I}_0} E_i] \le m$. Explain how the weights concentrate evidence on the true nulls.

## Computational.

**Exercise 8.25: Optional stopping simulation**

Reproduce Figure A.1. Simulate 4,000 null sequences of 250 independent $N(0, 1)$ observations, with one observation per look. Write $S_t = \sum_{j=1}^{t} X_j$. At look $t$, compute the fixed-sample p-value $1 - \Phi(S_t/\sqrt{t})$ and the likelihood-ratio e-process of Example 8.6,

$$M_t = \exp\{\mu_1 S_t - t\mu_1^2/2\}, \qquad \mu_1 = 0.35.$$

For each $t$, estimate the probability that the p-value has fallen to 0.05 or below by look $t$, and the probability that $M_t$ has reached 20 or above by look $t$. Plot both curves and compare the e-process curve with the bound from Ville's inequality.

**Exercise 8.26: Dependence simulation**

Implement BH, BY, and e-BH, starting with the design of Figure 8.6: $m = 400$, 80 nonnulls, $n = 20$, $q = 0.10$, and 650 independent runs per setting. Generate unit-variance Gaussian statistics with equicorrelation $\rho \in \{0, 0.3, 0.6\}$, mean zero for nulls, and mean $0.65\sqrt{20}$ for nonnulls. Use $p_i = 1 - \Phi(Z_i)$ for BH and BY. For e-BH, use

$$E_i = \exp\{\lambda Z_i - \lambda^2/2\}, \qquad \lambda = 0.65\sqrt{20},$$

with $\lambda$ fixed before generating the data. Vary the correlation, the true nonnull mean, and the number of nonnulls, keeping $\lambda$ fixed throughout. Report empirical FDR and, when nonnulls are present, empirical average power as defined in Exercise 4.16, together with Monte Carlo SEs.

CHAPTER 9

# Conditional Randomization and Knockoff Filters

In a regression or supervised learning problem, a feature can look useful for two different reasons. It may contain information about the response that is not already in the other features, or it may be correlated with another feature that contains the information. A marginal test asks whether $X_j$ is associated with $Y$. The conditional question is more local: after all the other features $X_{-j}$ have been observed, does $X_j$ still add information about $Y$?

This chapter studies three tools for that conditional question. A CRT tests one feature at a time by redrawing $X_j$ from its conditional law given $X_{-j}$ and seeing whether the observed statistic looks unusually large. Fixed-X knockoffs build artificial comparison columns for a fixed linear design and target the coefficient null $\beta_j = 0$. Model-X knockoffs build artificial comparison features from a model for the feature distribution and target the conditional independence null $Y \perp\!\!\!\perp X_j \mid X_{-j}$. In plain language, knockoffs are engineered coordinatewise null competitors: they preserve the relevant feature geometry or distribution and are constructed without using the response, so a null original and its knockoff have symmetric roles. A real-feature win means that the original feature receives a larger importance score than its knockoff; a knockoff win means the reverse. Selection also requires passing the knockoff threshold. Knockoff wins estimate how many selected real features may be false discoveries. This formal role comes from the knockoff construction, not merely from looking null in the observed sample.

The price of this robustness to the response model is clear. CRTs and model-X knockoffs do not require a correct model for $Y \mid X$, but they do require knowledge, or accurate estimation, of the feature distribution. Fixed-X knockoffs make a different tradeoff: they use exact geometry for a linear fixed-design model and therefore do not need a model for the distribution of the rows of $X$.

## 1. Conditional Feature Nulls

**Route: Core**

Let $X = (X_1, \dots, X_p)$ denote the feature vector and $Y$ the response. For an i.i.d. sample of size $n$, $X_j$ will also denote the $n$-vector of observed values of feature $j$, and $X_{-j}$ the matrix of all remaining features. The model-X null for feature $j$ is

$$H_j : \quad Y \perp\!\!\!\perp X_j \mid X_{-j}.$$

This says that once the other features are known, replacing $X_j$ by another draw from its conditional distribution should not change the distribution of the response.

**Definition 9.1: Conditional feature null**

Feature $j$ is conditionally null if

$$\mathcal{L}(Y \mid X_j, X_{-j}) = \mathcal{L}(Y \mid X_{-j})$$

almost surely. Equivalently, $Y \perp\!\!\!\perp X_j \mid X_{-j}$.

In the linear model

$$Y = X\beta + \varepsilon, \qquad \varepsilon \perp X, \qquad \mathbb{E}[\varepsilon \mid X] = 0,$$

with a nondegenerate conditional distribution of $X_j \mid X_{-j}$, this conditional null corresponds to $\beta_j = 0$. The definition is more general, however. It allows nonlinear response models, interactions, and arbitrary prediction algorithms as long as the null is interpreted conditionally.

Conditional and marginal independence are different properties; neither implies the other in general. If two features are correlated, a null feature may have a visible marginal relationship with the response because it tracks another feature. That is the failure mode that CRTs and knockoffs are designed to avoid.

## 2. Conditional Randomization Tests

**Route: Core**

The CRT of Candès et al. [38] tests one conditional null at a time. Choose a statistic

$$T_j = T_j(X_j, X_{-j}, Y)$$

that becomes large when feature $j$ appears important. The statistic may be a regression coefficient, a lasso entry time, a random-forest importance score, or the drop in predictive accuracy after removing $X_j$. The validity of the CRT does not come from the statistic. It comes from comparing the observed statistic with statistics computed after resampling $X_j$ from the correct conditional law.

Figure 9.1 separates the two ingredients of the test: the observed feature supplies one score, while repeated draws from $\mathcal{L}(X_j \mid X_{-j})$ supply its conditional null reference distribution. The plus-one count in the last box is the finite-sample Monte Carlo p-value.

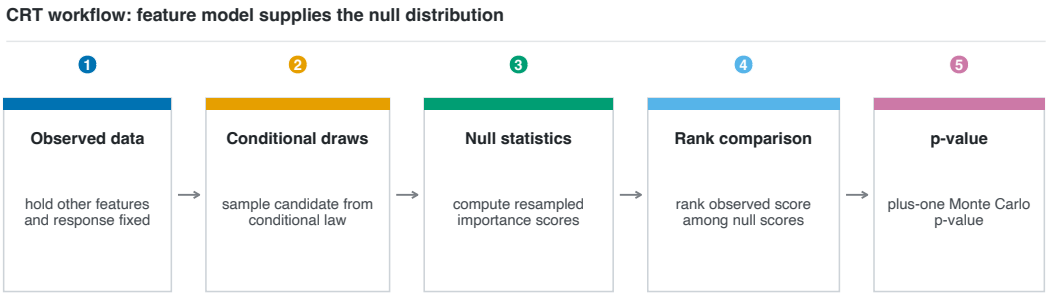



FIGURE 9.1. CRT. The conditional feature law, not a model for $Y \mid X$, supplies the null distribution. The final p-value uses the plus-one Monte Carlo correction.

Use the same statistic-building rule for the observed and resampled data. If that rule fits or tunes a model using $X_j$, repeat that fitting or tuning for every resample. A fit based only on $(X_{-j}, Y)$, or on independent training data, may instead be held fixed. For a randomized algorithm, use a common seed independent of the data, or independent identically distributed seeds for all $B+1$ evaluations; either choice treats the observed data and resamples symmetrically.

Given a fixed number $B$ of Monte Carlo draws, the CRT proceeds as follows.

1. Compute $T_j^{\text{obs}} = T_j(X_j, X_{-j}, Y)$.
2. For $b = 1, \dots, B,$ draw $X_j^{(b)} \sim \mathcal{L}(X_j \mid X_{-j})$.
3. Compute $T_j^{(b)} = T_j(X_j^{(b)}, X_{-j}, Y)$.
4. Return $p_j = \dfrac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_j^{(b)} \geq T_j^{\text{obs}}\}}{B+1}$.

The resamples are conditionally independent of one another and of the observed $X_j$, given $X_{-j}$, and are drawn without using $Y$. For i.i.d. rows, sample each entry from its feature conditional law given the other features in that row. The p-value is written for large values of $T_j$. For a two-sided question, one can use an absolute-value statistic and retain the same upper-tail comparison.

**Theorem 9.2: Finite-sample validity of the CRT**

Assume that the conditional law $\mathcal{L}(X_j \mid X_{-j})$ used to resample is correct, the resampling and statistic evaluations follow the symmetric construction above, and $H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$ holds. Conditional on $(X_{-j}, Y)$, the observed and resampled statistics $T_j^{\text{obs}}, T_j^{(1)}, \dots, T_j^{(B)}$ are exchangeable. Consequently the plus-one Monte Carlo p-value is super-uniform:

$$\mathbb{P}(p_j \leq \alpha) \leq \alpha, \qquad 0 \leq \alpha \leq 1.$$

**Proof of Theorem 9.2**

Under $H_j$, the conditional distribution of $X_j$ given $(X_{-j}, Y)$ coincides with $\mathcal{L}(X_j \mid X_{-j})$: the response carries no information about $X_j$ once $X_{-j}$ is fixed. The resampled draws $X_j^{(1)}, \dots, X_j^{(B)}$ come from the same conditional law by construction. Therefore the collection

$$X_j,\ X_j^{(1)}, \dots, X_j^{(B)}$$

is iid conditional on $(X_{-j}, Y)$, and the induced statistics $T_j^{\text{obs}}, T_j^{(1)}, \dots, T_j^{(B)}$ are exchangeable conditional on $(X_{-j}, Y)$. If there are no ties, order these $B+1$ values from largest to smallest and let $R_j$ be the rank of the observed statistic in that ordering. Exchangeability means that, conditional on $(X_{-j}, Y)$, the observed statistic is equally likely to occupy any of the $B+1$ positions. The plus-one p-value is exactly

$$p_j = \frac{R_j}{B+1},$$

because $R_j$ is the number of observed-or-resampled statistics at least as large as $T_j^{\text{obs}}$. Therefore

$$\mathbb{P}(p_j \leq \alpha \mid X_{-j}, Y) = \frac{\lfloor (B+1)\alpha \rfloor}{B+1} \leq \alpha.$$

If ties occur, counting all tied statistics in the upper tail can only increase the p-value relative to a random tie break, so the same inequality remains valid. Taking expectations over $(X_{-j}, Y)$ gives the unconditional bound. □

The plus-one correction is not cosmetic. Without it, the smallest possible Monte Carlo p-value would be zero, and exact finite-sample validity would be lost. The correction also makes clear how many null samples are needed if CRT p-values will be fed into a multiple testing

procedure. For example, if there are $p$ hypotheses and one hopes to use BH at level $q$, then the smallest p-value is $1/(B+1)$. A BH rejection at rank $k$ can occur only if $1/(B+1) \le qk/p$; a rejection at rank one requires $B+1 \ge p/q$. This grid condition concerns resolution, not validity of BH. The CRT theorem supplies marginally valid p-values, but applying BH also requires its independence or PRDS assumptions from Chapters 5–6. BY remains available under arbitrary dependence among the valid CRT p-values.

## 3. Why Marginal Resampling Fails

**Route: Core**

The CRT must resample from $X_j \mid X_{-j}$, not from the marginal law of $X_j$. The following example is the simplest way to see why.

**Example 9.3: A null feature with marginal association**

Let

$$(X_1, X_2) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \qquad Y = X_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

with $\varepsilon$ independent of $(X_1, X_2)$. Then

$$Y \perp\!\!\!\perp X_1 \mid X_2,$$

so feature 1 is conditionally null. Marginally,

$$\mathrm{Cov}(X_1, Y) = \mathrm{Cov}(X_1, X_2) = \rho.$$

Thus $X_1$ can be strongly correlated with the response even though it adds no information once $X_2$ is known.

Figure 9.2 makes the failure visible. Compare the middle and right panels: marginal resampling erases the $X_1$–$X_2$ dependence that generated the apparent association, whereas conditional resampling preserves that dependence while breaking only the information that is null under $H_1$.

In Example 9.3, a marginal randomization test would compare the observed association between $X_1$ and $Y$ with a null distribution obtained by drawing $X_1^* \sim N(0,1)$ independently of $X_2$. That null distribution is wrong: it describes a world in which $X_1^*$ no longer tracks $X_2$. The correct conditional draw is

$$X_1^* \mid X_2 = x_2 \sim N(\rho x_2,\ 1 - \rho^2),$$

which preserves the correlation structure that makes $X_1$ look important marginally.

After valid CRT p-values have been obtained, they can be used with the multiple testing procedures from earlier chapters. The important distinction is between individual p-value validity and joint dependence. The CRT theorem above says that, for each true null $j$, the p-value $p_j$ is super-uniform. It does not by itself say that the vector $(p_1, \ldots, p_p)$ is independent or PRDS. Thus ordinary BH may be used when the analyst has a separate reason to justify the BH dependence condition, such as independence across the null p-values or a direct PRDS argument for the joint CRT construction being used. In the absence of such a joint dependence argument, a dependence-robust procedure such as BY or another reshaping correction gives the conservative, generally valid route.

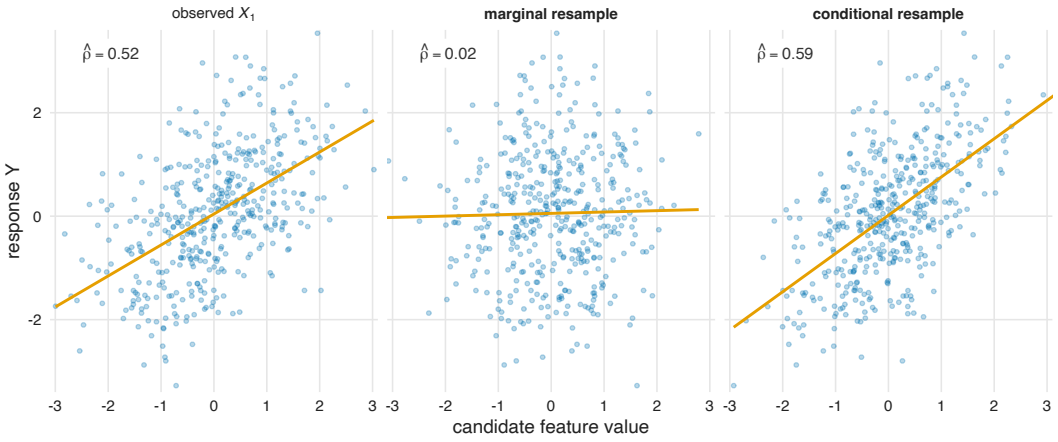



FIGURE 9.2. A conditionally null feature can be marginally correlated with the response. Each panel contains $n = 450$ observations, the data-generating correlation is $\rho = 0.65$, and the labels show empirical $\widehat{\rho}$ values in the plotted samples. Marginal resampling breaks the dependence between $X_1$ and $X_2$, whereas conditional resampling preserves it.

## 4. Fixed-X Knockoffs

**Route: Core**

We next move from testing one feature at a time to constructing synthetic knockoff competitors for all features simultaneously. The original knockoff construction of Barber and Candès [9] is for the fixed-design Gaussian linear model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

where $X \in \mathbb{R}^{n\times p}$ is treated as fixed. The lower-case $y$ emphasizes that the design $X$ is treated as fixed; the response remains random under the Gaussian linear model. Equivalently, inference conditions on $X$, not on the observed response vector. The null hypothesis is

$$H_j : \beta_j = 0.$$

Full column rank alone means $\operatorname{rank}(X) = p$, hence $n \geq p$. The explicit construction below also needs room to add $p$ orthogonal directions outside the span of $X$. When $X$ has full column rank, this extra geometric condition is $n - p \geq p$, or $n \geq 2p$. Variants are available when this condition is not met, but the core symmetry is easiest to see in this full-column-rank, $n \geq 2p$ case.

Let

$$\Sigma = X^\top X$$

and let $D = \operatorname{diag}(s)$ for a vector $s \in \mathbb{R}^p_+$. A fixed-X knockoff matrix $\widetilde{X} \in \mathbb{R}^{n\times p}$ is designed to satisfy

$$\widetilde{X}^\top \widetilde{X} = \Sigma, \qquad X^\top \widetilde{X} = \Sigma - D.$$

Equivalently,

$$\begin{pmatrix} X^\top X & X^\top \widetilde{X} \\ \widetilde{X}^\top X & \widetilde{X}^\top \widetilde{X} \end{pmatrix} = \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}.$$

Each knockoff column has the same inner products with all other original features as the original column does, except that its inner product with its own original feature is reduced by $s_j$. When

the columns of $X$ are normalized so that $\Sigma_{jj} = 1$, the correlation between $X_j$ and $\widetilde{X}_j$ is $1 - s_j$. Larger $s_j$ gives a more distinguishable knockoff competitor and usually improves power.

**Proposition 9.4: Fixed-X knockoff construction**

Suppose $X$ has full column rank, $n - \operatorname{rank}(X) \geq p$, and $D = \operatorname{diag}(s) \succeq 0$ satisfies

$$2\Sigma - D \succeq 0.$$

Let $U \in \mathbb{R}^{n \times p}$ have orthonormal columns satisfying $U^\top X = 0$, and choose $C \in \mathbb{R}^{p \times p}$ such that

$$C^\top C = 2D - D\Sigma^{-1}D.$$

Then

$$\widetilde{X} = X(I - \Sigma^{-1}D) + UC$$

satisfies

$$X^\top \widetilde{X} = \Sigma - D, \qquad \widetilde{X}^\top \widetilde{X} = \Sigma.$$

**Proof of Proposition 9.4**

Since $U^\top X = 0$,

$$X^\top \widetilde{X} = X^\top X(I - \Sigma^{-1}D) = \Sigma - D.$$

Next,

$$\begin{aligned} \widetilde{X}^\top \widetilde{X} &= (I - D\Sigma^{-1})\Sigma(I - \Sigma^{-1}D) + C^\top C \\ &= \Sigma - 2D + D\Sigma^{-1}D + 2D - D\Sigma^{-1}D \\ &= \Sigma. \end{aligned}$$

It remains to check that such a $C$ exists, i.e. that $2D - D\Sigma^{-1}D \succeq 0$. Consider the block Gram matrix

$$G = \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}.$$

On one hand, the congruence transformation $P = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} I & I \\ I & -I \end{smallmatrix}\right)$ brings $G$ to $P^\top G P = \operatorname{diag}(2\Sigma - D,\ D)$, so $G \succeq 0$ iff $2\Sigma - D \succeq 0$ and $D \succeq 0$. On the other hand, the Schur complement of the top-left block gives $G \succeq 0$ iff $\Sigma - (\Sigma - D)\Sigma^{-1}(\Sigma - D) = 2D - D\Sigma^{-1}D \succeq 0$. The first characterization says exactly that the assumed constraints $2\Sigma - D \succeq 0$ and $D \succeq 0$ make the target block Gram matrix positive semidefinite. The Schur-complement characterization then converts that same positive semidefiniteness into $2D - D\Sigma^{-1}D \succeq 0$, so a real matrix $C$ with $C^\top C = 2D - D\Sigma^{-1}D$ exists. □

Figure 9.3 displays the Gram structure in block form. An analogous block *covariance* matrix will reappear in §7 for Gaussian model-X knockoffs, with $\Sigma$ reinterpreted as a population covariance rather than the Gram matrix used here.

The vector $s$ is a power parameter. Assume from here that the columns of $X$ are normalized so that $\Sigma_{jj} = 1$; a simple equi-correlated choice is then

$$s_j = \min\{1, 2\lambda_{\min}(\Sigma)\}, \qquad j = 1, \ldots, p.$$

A more adaptive choice solves

$$\min_{0 \leq s_j \leq 1} \sum_{j=1}^{p} |1 - s_j| \qquad \text{subject to} \qquad 2\Sigma - \operatorname{diag}(s) \succeq 0.$$

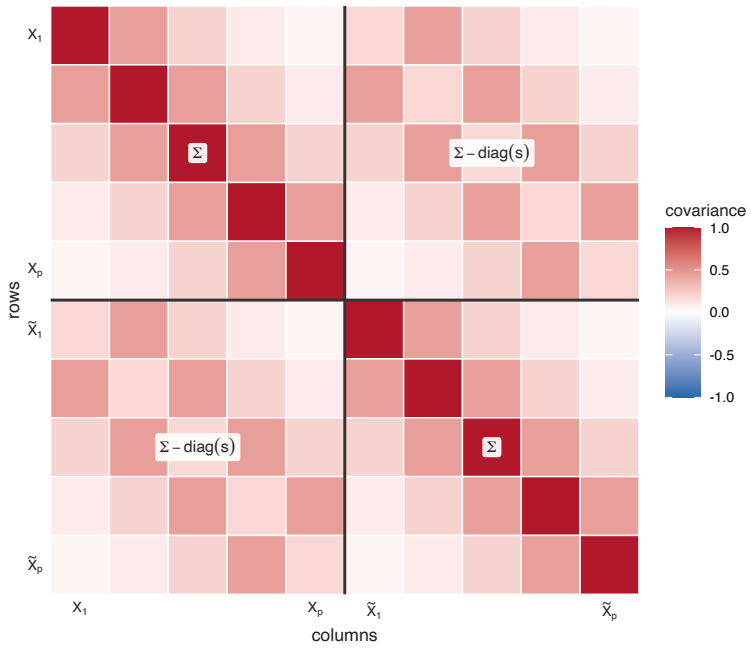



FIGURE 9.3. The knockoff block structure. Original and knockoff variables have matching within-block dependence; the diagonal gap $\mathrm{diag}(s)$ makes each knockoff a less-than-perfect copy of its original feature. In the fixed-X setting (§4) this is a Gram matrix; in the model-X setting (§7) it is a population covariance.

Both choices express the same principle: make $X_j$ and $\widetilde{X}_j$ as different as the feature correlation structure permits.

## 5. Knockoff Statistics and Thresholds

**Route: Core**

After constructing $\widetilde{X}$, fit a variable-selection method to the augmented design

$$[X, \widetilde{X}].$$

The output must be converted into a vector

$$W = (W_1, \ldots, W_p),$$

with one statistic for each original feature. Positive $W_j$ means that the original feature $X_j$ looks stronger than its knockoff $\widetilde{X}_j$; negative $W_j$ means that the knockoff looks stronger. The fixed-X vector $W$ must satisfy two structural properties.

**Definition 9.5: Sufficiency and antisymmetry**

A fixed-X knockoff statistic vector $W = (W_1, \ldots, W_p)$ is sufficient if it depends on the data only through

$$[X, \widetilde{X}]^\top [X, \widetilde{X}], \qquad [X, \widetilde{X}]^\top y.$$

It is antisymmetric if, for each coordinate $j$, swapping $X_j$ and $\widetilde{X}_j$ sends $W_j$ to $-W_j$ and leaves every $W_k$, $k \neq j$, unchanged. If several coordinates are swapped, exactly the corresponding signs of $W$ are flipped.

The standard example comes from the lasso path. Let $Z_j$ be the largest penalty value at which $X_j$ enters the lasso model fit on $[X, \widetilde{X}]$, and let $\widetilde{Z}_j$ be the analogous entry time for $\widetilde{X}_j$. Concretely, the lasso solves

$$\min_b \frac{1}{2}\|y - [X, \widetilde{X}]b\|_2^2 + \lambda\|b\|_1$$

as $\lambda$ decreases from a very large value. At very large $\lambda$, all coefficients are zero. A column with stronger predictive information tends to become nonzero earlier, meaning at a larger value of $\lambda$. Thus $Z_j$ and $\widetilde{Z}_j$ record which member of the pair $(X_j, \widetilde{X}_j)$ enters first, and how early it enters. Define the signed-max statistic

$$W_j = \max\{Z_j, \widetilde{Z}_j\} \operatorname{sign}(Z_j - \widetilde{Z}_j).$$

Large positive $W_j$ means the original feature beats its knockoff strongly. Large negative $W_j$ means the knockoff beats the original feature. For null features, the two events should be symmetric. The fitted path and any handling of nonunique solutions or ties must preserve the swap properties; a rule that gives priority to the original columns can break antisymmetry.

Let

$$\mathcal{W} = \{|W_j| : \ |W_j| > 0, \ j = 1, \ldots, p\}.$$

The knockoff+ threshold at target FDR level $q$ is

$$T_+ = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\},$$

with $T_+ = \infty$ if the set is empty. The selected variables are

$$\widehat{\mathcal{S}} = \{j : W_j \geq T_+\}.$$

A boundary observation: the numerator in the threshold inequality is at least 1, so for the ratio to fall below $q$ we need $\#\{j : W_j \geq t\} \geq 1/q$ at the chosen $t$. When the target level satisfies $q < 1/p$, *no* value of $t$ can satisfy the inequality, and knockoff+ returns $\widehat{\mathcal{S}} = \varnothing$ deterministically. This minimum discovery count comes from the plus-one correction. BH has no comparable deterministic lower bound on the number of discoveries: a single sufficiently small p-value can cross its first threshold. The version without the plus one,

$$T = \min\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\},$$

is often more powerful but controls a slightly modified FDR criterion. The plus-one numerator is what gives ordinary FDR control. This is the same sign-count idea as the mirror procedures in Section 11 of Chapter 5. Here the artificial comparison features supply the conditional fair-sign property required by that argument.

Figure 9.4 shows the diagnostic behind that correction. The null bars should balance across the positive and negative tails, so negative knockoff wins can proxy false positive original wins; signal appears as excess mass in the positive tail.

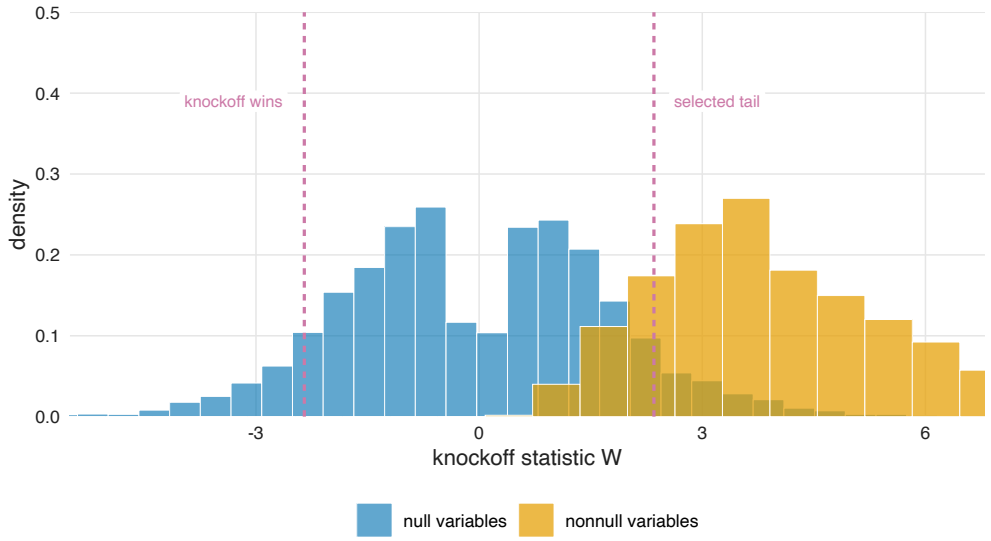



FIGURE 9.4. Knockoff signs. For true null variables, positive and negative $W_j$'s are symmetric; negative knockoff wins therefore estimate the number of false positive original wins. Nonnull variables should shift mass toward the positive tail. This synthetic display uses 6000 null and 900 nonnull statistics; the dashed lines at $\pm 2.35$ mark an illustrative threshold.

**Theorem 9.6: Fixed-X knockoff+ FDR control**

In the fixed-design Gaussian linear model, suppose $\widetilde{X}$ satisfies the fixed-X knockoff Gram equations and $W$ is sufficient and antisymmetric. Then the knockoff+ selection rule satisfies

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0,\ j \in \widehat{\mathcal{S}}\}}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq q.$$

**Proof of Theorem 9.6**

*Exchangeability.* Let $A = [X, \widetilde{X}]$, and let $A_{\mathrm{swap}(S)}$ be the augmented design obtained by interchanging $X_j$ and $\widetilde{X}_j$ for every $j \in S$. The fixed-X knockoff Gram equations imply

$$A_{\mathrm{swap}(S)}^\top A_{\mathrm{swap}(S)} = A^\top A$$

for every $S$. If $S$ contains only true nulls, then $A_{\mathrm{swap}(S)}^\top y$ and $A^\top y$ are Gaussian vectors with the same covariance by this identity. Their means also agree: when $j \in S$ is null, $\beta_j = 0$, and the knockoff equations give $\widetilde{X}_j^\top X_k = X_j^\top X_k$ for every $k \neq j$. Thus swapping null original/knockoff pairs is distribution-preserving.

*Sign symmetry.* Sufficiency says that $W$ is a function only of $A^\top A$ and $A^\top y$. Antisymmetry says that swapping pair $j$ flips $W_j$ and leaves the other $W_k$'s unchanged. Hence the law of $W$ is unchanged after multiplying $W_j$ by $-1$ for any chosen set of true-null coordinates. If $H_0 = \{j : \beta_j = 0\}$ and

$$\mathcal{C} = \sigma((|W_1|, \ldots, |W_p|), (W_j : j \notin H_0)),$$

then the signs $\operatorname{sign}(W_j)$, $j \in H_0$ with $|W_j| > 0$, are independent fair coins conditional on $\mathcal{C}$. This is the fixed-X sign-flip result of Barber and Candès [9, Lemma 1]; the preceding paragraph gives the mechanism behind the lemma in this setting.

*Stopped ratio.* For a threshold $t$, write

$$V^+(t) = \#\{j : \beta_j = 0,\ W_j \geq t\}, \qquad V^-(t) = \#\{j : \beta_j = 0,\ W_j \leq -t\}.$$

After conditioning on $\mathcal{C}$, the knockoff+ rule checks distinct candidate magnitudes from smallest to largest. Each failed check raises the threshold and removes all statistics at the current magnitude; the nonnull statistics and all magnitudes are already fixed. The reverse-filtration stopping-time calculation given below therefore gives

$$\mathbb{E}\left[\frac{V^+(T_+)}{1+V^-(T_+)} \,\middle|\, \mathcal{C}\right] \leq 1,$$

and hence the same inequality holds after averaging over $\mathcal{C}$.

*FDR bound.* If $T_+ = \infty$, the selected set is empty, so $\mathrm{FDP} = V^+(T_+) = V^-(T_+) = 0$. If $T_+ < \infty$, the observable knockoff+ threshold inequality and $V^-(T_+) \leq \#\{j : W_j \leq -T_+\}$ give

$$\begin{aligned}\mathrm{FDP} &= \frac{V^+(T_+)}{|\widehat{\mathcal{S}}| \vee 1} \\ &\leq \frac{1 + \#\{j : W_j \leq -T_+\}}{|\widehat{\mathcal{S}}| \vee 1} \cdot \frac{V^+(T_+)}{1+V^-(T_+)} \\ &\leq q \cdot \frac{V^+(T_+)}{1+V^-(T_+)}.\end{aligned}$$

The final inequality therefore holds in both cases. Taking expectations and applying the stopped sign-count calculation yields $\mathrm{FDR} \leq q$. □

Stopped sign-count calculation. We now verify the martingale input used in the proof. This calculation uses only conditional fair signs ordered by their magnitudes, so the model-X proof below can reuse it once model-X exchangeability has supplied the same sign property. Conditional on $\mathcal{C}$, list the nonzero null statistics in decreasing order of magnitude,

$$|W_{j_1}| \geq \cdots \geq |W_{j_m}| > 0,$$

breaking ties by feature index, independently of the signs. Write $\xi_r = \operatorname{sign}(W_{j_r})$. By sign symmetry, the $\xi_r$'s are iid fair coins. Define the active positive and negative counts and their ratio by

$$P_k = \#\{r \leq k : \xi_r = +1\}, \qquad Q_k = \#\{r \leq k : \xi_r = -1\}, \qquad M_k = \frac{P_k}{1+Q_k}, \qquad 0 \leq k \leq m.$$

For the proof, keep only the counts of active null signs and reveal individual signs as they are removed. This gives the reverse filtration

$$\mathcal{F}_k = \sigma(\mathcal{C},\, P_k,\, Q_k,\, \xi_{k+1}, \ldots, \xi_m), \qquad k = m, m-1, \ldots, 0.$$

As $k$ decreases, information increases: $\mathcal{F}_k \subseteq \mathcal{F}_{k-1}$. For $k \geq 1$, conditional on $\mathcal{F}_k$, all assignments of the $P_k$ positive and $Q_k$ negative signs to the $k$ active null indices are equally likely. Consequently, the sign removed when the active null set shrinks from $k$ to $k-1$ is positive with probability $P_k/k$ and negative with probability $Q_k/k$. Hence, when $Q_k > 0$,

$$\mathbb{E}(M_{k-1} \mid \mathcal{F}_k) = \frac{P_k}{k}\frac{P_k - 1}{1+Q_k} + \frac{Q_k}{k}\frac{P_k}{Q_k} = M_k.$$

When $Q_k = 0$, all active signs are positive and $M_{k-1} = k - 1 \le k = M_k$. Thus $(M_k, \mathcal{F}_k)$, read in the order $m, m-1, \dots, 0$, is a supermartingale.

*Moving between candidate thresholds, including ties.* The knockoff+ rule checks distinct positive magnitudes from smallest to largest. At a candidate threshold $t$, a statistic is active if $|W_j| \ge t$. If $k$ of the active statistics are null, then

$$\frac{V^+(t)}{1 + V^-(t)} = \frac{P_k}{1 + Q_k} = M_k.$$

After a failed check, remove all statistics with magnitude $t$ and move to the next candidate. Thus ties are kept together: *check the whole block before removing any of it.*

The one-sign calculation still applies to a whole block. For example, if a block contains two nulls, iterated conditional expectation gives

$$\mathbb{E}(M_{k-2} \mid \mathcal{F}_k) = \mathbb{E}[\mathbb{E}(M_{k-2} \mid \mathcal{F}_{k-1}) \mid \mathcal{F}_k] \le \mathbb{E}(M_{k-1} \mid \mathcal{F}_k) \le M_k.$$

The same argument works for any block size. Conditional on $\mathcal{C}$, all magnitudes and hence all block sizes are fixed. Removing only nonnull statistics leaves the null ratio unchanged. Thus the ratio remains a supermartingale when observed only at the actual candidate checks.

*Why the rule is allowed to stop.* At each check, the total positive count is $P_k$ plus the number of active positive nonnulls; the total negative count is $Q_k$ plus the number of active negative nonnulls. The nonnull contributions are already fixed by $\mathcal{C}$. Consequently, the information in $\mathcal{F}_k$ determines whether the knockoff+ inequality holds. It also records the removed signs, so all earlier checks can be reconstructed. The first passing check is therefore a stopping time for the candidate scan: the decision to stop uses only information available at that check. This information is used to justify the proof; the analyst does not need to know which features are null.

If no candidate passes, finish with the empty active set and ratio $M_0 = 0$, corresponding to $T_+ = \infty$. There are at most $p$ candidate checks and this one terminal step, and every ratio lies between zero and $m$. The finite optional-stopping theorem therefore says that the expected ratio at the chosen threshold cannot exceed its initial expectation:

$$\mathbb{E}\left[ \frac{V^+(T_+)}{1 + V^-(T_+)} \,\middle|\, \mathcal{C} \right] \le \mathbb{E}(M_m \mid \mathcal{C}).$$

If there are no nonzero null statistics, the ratio is always zero and the bound is immediate. This also covers the case in which all statistics are zero and there are no candidate thresholds.

*Bounding the initial expectation.* Initially, $Q_m \mid \mathcal{C} \sim \text{Binomial}(m, 1/2)$ and $P_m = m - Q_m$, so

$$M_m = \frac{m - Q_m}{1 + Q_m} = \frac{m+1}{1 + Q_m} - 1.$$

Using $\binom{m}{r}/(r+1) = \binom{m+1}{r+1}/(m+1)$, we obtain

$$\begin{aligned} \mathbb{E}(M_m \mid \mathcal{C}) &= \frac{m+1}{2^m} \sum_{r=0}^{m} \frac{\binom{m}{r}}{r+1} - 1 \\ &= \frac{1}{2^m} \sum_{r=0}^{m} \binom{m+1}{r+1} - 1 = 1 - 2^{-m} \le 1. \end{aligned}$$

Together with optional stopping, this proves the required stopped-ratio bound.

The proof says that null original wins and null knockoff wins are symmetric after conditioning on the magnitudes. At a fixed threshold, negative null wins therefore estimate positive null wins. Since $T_+$ is chosen after looking across thresholds, the estimate needs the plus-one correction; that correction is exactly what keeps the stopped estimate conservative.

## 6. Model-X Knockoffs

**Route: Core**

Fixed-X knockoffs are tied to the linear fixed-design model with target $\beta_j = 0$. Model-X knockoffs change both ingredients. The rows $X_i$ are treated as iid draws from an unknown population distribution, the response model is unrestricted, and the target is the conditional independence null $Y \perp\!\!\!\perp X_j \mid X_{-j}$ from §2 [38].

The two targets are inferentially different even when their algebra looks similar. Under the fixed-X Gaussian linear model, $E[y \mid X] = X\beta$, so $\beta_j = 0$ refers to the $j$th coefficient in that specified linear mean model. If the linear mean model is misspecified, the corresponding target may instead be the least-squares projection coefficient of the fixed mean vector onto the columns of $X$; that projection target must be stated explicitly. Under model-X, $Y \perp X_j \mid X_{-j}$ refers to a distributional statement that does not depend on linearity. They coincide in the special case of a correctly specified Gaussian linear model with random $X$, but in general one can hold without the other.

Notation also resets here. In §4, $\Sigma = X^\top X$ was the Gram matrix of a deterministic design; in the present section, $\Sigma = \mathrm{Cov}(X)$ is the population covariance of the random feature vector. The algebraic identities that follow are the same, but the statistical meaning is different.

**Definition 9.7: Model-X knockoff**

A random vector $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$ is a model-X knockoff copy of $X$ if it satisfies:

$$[X, \widetilde{X}]_{\mathrm{swap}(S)} \overset{d}{=} [X, \widetilde{X}] \qquad \text{for every } S \subseteq \{1, \ldots, p\},$$

where $\mathrm{swap}(S)$ exchanges $X_j$ and $\widetilde{X}_j$ for all $j \in S$, and

$$\widetilde{X} \perp\!\!\!\perp Y \mid X.$$

For a sample, the construction is applied row by row, or more generally to the joint feature array, without using the responses.

The first condition is pairwise exchangeability. It is stronger than matching marginal distributions. A row permutation or column shuffle generally does not work: it may give $\widetilde{X}_j$ the same marginal distribution as $X_j$, but it does not preserve the conditional relationship between $X_j$ and $X_{-j}$ for the same observational unit. Valid knockoffs are designed coordinatewise null competitors whose guarantee comes from pairwise exchangeability, not arbitrary fake features.

**Lemma 9.8: Null-swap property**

Let $\widetilde{X}$ be a model-X knockoff copy of $X$, and let $\mathcal{I}_0 = \{j : Y \perp\!\!\!\perp X_j \mid X_{-j}\}$. For every subset $S \subseteq \mathcal{I}_0$,

$$([X, \widetilde{X}], Y) \overset{d}{=} ([X, \widetilde{X}]_{\mathrm{swap}(S)}, Y).$$

**Proof of Lemma 9.8**

First consider one null coordinate $j$. Write $Z = [X, \widetilde{X}]$ and let $S_j Z$ swap its $j$th original/knockoff pair. The feature construction gives $S_j Z \overset{d}{=} Z$. Moreover, $\widetilde{X} \perp\!\!\!\perp Y \mid X$ and the

coordinatewise null imply that the conditional response law given $Z$ depends only on $X_{-j}$:

$$\mathcal{L}(Y \mid X, \widetilde{X}) = \mathcal{L}(Y \mid X) = \mathcal{L}(Y \mid X_{-j}).$$

A swap of pair $j$ leaves $X_{-j}$ unchanged. To make the resulting joint invariance explicit, let $h$ be any bounded measurable function of $Z$, let $A$ be a response event, and write $K_j(X_{-j}, A) = \mathbb{P}(Y \in A \mid X_{-j})$. Then

$$\begin{aligned}\mathbb{E}[h(S_jZ)\mathbf{1}\{Y \in A\}] &= \mathbb{E}[h(S_jZ)K_j(X_{-j}, A)] \\ &= \mathbb{E}[h(Z)K_j(X_{-j}, A)] \\ &= \mathbb{E}[h(Z)\mathbf{1}\{Y \in A\}].\end{aligned}$$

The middle equality uses feature swap invariance and the fact that the kernel's argument is unchanged by the swap. Hence $(S_jZ, Y)$ and $(Z, Y)$ have the same law.
Every null-coordinate swap therefore preserves this joint law. Applying these swaps successively proves the assertion for any subset $S \subseteq \mathcal{I}_0$, as in Candès et al. [38, Lemma 3.2]. □

Once the null-swap property holds, the same sign-flip logic used for fixed-X knockoffs applies. A model-X feature statistic may be produced by any algorithm, including random forests, neural networks, penalized regressions, or screening scores, as long as it has the right swap behavior. More explicitly, suppose the augmented data $(Y, X, \widetilde{X})$ are mapped to original and knockoff importance scores

$$(Z, \widetilde{Z}) = (Z_1, \dots, Z_p, \widetilde{Z}_1, \dots, \widetilde{Z}_p).$$

The score map is swap-equivariant if, for every coordinate set $S \subseteq \{1, \dots, p\}$, swapping $(X_j, \widetilde{X}_j)$ for $j \in S$ swaps the corresponding score pairs $(Z_j, \widetilde{Z}_j)$ and leaves all other score pairs in the same order. Given any antisymmetric comparison function $w_j(a, b) = -w_j(b, a)$, define

$$W_j = w_j(Z_j, \widetilde{Z}_j).$$

The simple choice $W_j = Z_j - \widetilde{Z}_j$ is enough for the argument; the signed-max statistic from §5 is a common lasso-path choice. Positive $W_j$ means the original feature beats its knockoff, and negative $W_j$ means the knockoff wins.

**Lemma 9.9: Model-X null sign symmetry**

Let $\mathcal{I}_0$ be the set of true conditional nulls. Under the conditions of Lemma 9.8, if the score map is swap-equivariant and $W_j = w_j(Z_j, \widetilde{Z}_j)$ is antisymmetric, then

$$(W_1, \dots, W_p) \overset{d}{=} (\varepsilon_1 W_1, \dots, \varepsilon_p W_p)$$

for every deterministic sign vector with $\varepsilon_j \in \{-1, 1\}$ on $\mathcal{I}_0$ and $\varepsilon_j = 1$ otherwise. Consequently, conditional on the indexed magnitudes $(|W_1|, \dots, |W_p|)$ and the nonnull statistics $(W_j : j \notin \mathcal{I}_0)$, the signs of the nonzero null statistics are independent fair signs.

**Proof of Lemma 9.9**

For any subset $S \subseteq \mathcal{I}_0$, Lemma 9.8 says that swapping the corresponding original and knockoff variables leaves the joint law of the augmented data and response unchanged. Swap-equivariance converts this data swap into the operation that interchanges $(Z_j, \widetilde{Z}_j)$ for $j \in S$. Antisymmetry then flips exactly the signs of $W_j$, $j \in S$, while preserving the

magnitudes and the nonnull statistics. Since every subset $S$ may be swapped, every null sign pattern has the same conditional probability given those quantities. □

For $t > 0$, define the selected positive-tail set and its size by

$$\mathcal{R}(t) = \{j : W_j \geq t\}, \qquad R(t) = |\mathcal{R}(t)|.$$

The null false discoveries at threshold $t$ are

$$V^+(t) = \#\{j \in \mathcal{I}_0 : W_j \geq t\},$$

while the negative null wins are

$$V^-(t) = \#\{j \in \mathcal{I}_0 : W_j \leq -t\}.$$

Because null signs are symmetric, negative wins estimate positive false wins. Use the same observable knockoff+ estimate and threshold as in Section 5. In the notation of this section,

$$\widehat{\mathrm{FDP}}(t) = \frac{1 + \#\{j : W_j \leq -t\}}{R(t) \vee 1}, \qquad \tau_q = T_+, \qquad \widehat{\mathcal{S}} = \{j : W_j \geq \tau_q\}.$$

As before, search the positive observed magnitudes and set the threshold to $\infty$ if none passes. The threshold rule is unchanged; model-X exchangeability now supplies the sign symmetry used in its proof.

**Theorem 9.10: Model-X knockoff+ FDR control**

Assume $\widetilde{X}$ is an exact model-X knockoff copy of $X$, generated without using $Y$. Let $W$ be constructed from a swap-equivariant score map and antisymmetric comparisons as above. Then the knockoff+ threshold controls FDR for the conditional nulls

$$H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$$

at level $q$.

**Proof of Theorem 9.10**

At the selected threshold,

$$\mathrm{FDP} = \frac{V^+(\tau_q)}{R(\tau_q) \vee 1}.$$

If $\tau_q = \infty$, the selected set is empty and $\mathrm{FDP} = V^+(\tau_q) = V^-(\tau_q) = 0$.
For $\tau_q < \infty$, the threshold inequality and $V^-(\tau_q) \leq \#\{j : W_j \leq -\tau_q\}$ give

$$\begin{aligned}
\mathrm{FDP} &= \frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \frac{1 + V^-(\tau_q)}{R(\tau_q) \vee 1} \\
&\leq \frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \frac{1 + \#\{j : W_j \leq -\tau_q\}}{R(\tau_q) \vee 1} \\
&\leq q \, \frac{V^+(\tau_q)}{1 + V^-(\tau_q)}.
\end{aligned}$$

The final inequality holds in the empty-selection case as well. It remains to show that the expectation of the final ratio is at most one. Condition on

$$\mathcal{C} = \sigma((|W_1|, \ldots, |W_p|), (W_j : j \notin \mathcal{I}_0)).$$

By Lemma 9.9, conditional on $\mathcal{C}$, the signs of the nonzero null statistics are independent fair signs. The threshold $\tau_q$ is found by checking distinct positive magnitudes from smallest to largest, removing the entire block at the current magnitude after each failed check. With the nonnull statistics and all magnitudes fixed, the first passing stage is the bounded stopping time verified in the stopped sign-count calculation following Theorem 9.6. That calculation applies verbatim and gives

$$\mathbb{E}\left[\frac{V^+(\tau_q)}{1+V^-(\tau_q)}\,\middle|\,\mathcal{C}\right] \le 1.$$

Averaging over $\mathcal{C}$ and using the pointwise FDP bound yields FDR $\le q$. □

**Remark 9.11: The threshold is not the filter**

The two-count ratio defining $\tau_q$ is only one component of the knockoff filter. Exchangeability of $(X, \widetilde{X})$, together with antisymmetry of the statistics, supplies the conditional coordinate-wise sign flips used in the proof: after the magnitudes and nonnull statistics are fixed, the nonzero null signs are independent fair signs. A generic collection of symmetric, Gaussian, exchangeable, or pairwise-uncorrelated scores need not have this property, and applying the same threshold to such scores can have FDR far above $q$. Section 2 of Chapter 6 gives finite PRDS and Gaussian examples; see also Zhang [201]. These are failures of a bare threshold, not counterexamples to a valid knockoff construction.

## 7. Gaussian Model-X Knockoffs

**Route: Core**

For the Gaussian model-X construction of Candès et al. [38], when

$$X \sim N(0, \Sigma), \qquad \Sigma \succ 0,$$

knockoffs are explicit. The two-coordinate case shows where the block covariance comes from. Let $X = (X_1, X_2) \sim N(0, \Sigma)$, with $\Sigma = (\sigma_{jk})_{j,k=1}^2$. Pairwise exchangeability under the full swap forces $(\widetilde{X}_1, \widetilde{X}_2)$ to have the same covariance matrix $\Sigma$. Exchangeability under the swap of only coordinate 2 forces

$$\mathrm{Cov}(X_1, \widetilde{X}_2) = \mathrm{Cov}(X_1, X_2) = \sigma_{12}, \qquad \mathrm{Cov}(\widetilde{X}_1, X_2) = \sigma_{12}.$$

Thus the only remaining free cross-covariances are the within-pair quantities $\mathrm{Cov}(X_j, \widetilde{X}_j)$. Writing them as $\sigma_{jj} - s_j$ gives

$$G_2 := \mathrm{Cov}(X_1, X_2, \widetilde{X}_1, \widetilde{X}_2) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{11}-s_1 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} & \sigma_{12} & \sigma_{22}-s_2 \\ \sigma_{11}-s_1 & \sigma_{12} & \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}-s_2 & \sigma_{12} & \sigma_{22} \end{pmatrix}.$$

Equivalently,

$$G_2 = \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}, \qquad D = \mathrm{diag}(s_1, s_2).$$

Ideally $s_j$ is large, because this makes $\widetilde{X}_j$ a more distinguishable competitor for $X_j$, but the displayed matrix must remain positive semidefinite.

For general $p$, choose $D = \mathrm{diag}(s)$ such that

$$D \succeq 0, \qquad 2\Sigma - D \succeq 0.$$

Define the joint Gaussian distribution

$$\begin{pmatrix} X \\ \widetilde{X} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}\right).$$

The block covariance is invariant under swapping any coordinate pair $(X_j, \widetilde{X}_j)$, and therefore the resulting $\widetilde{X}$ is a model-X knockoff copy when sampled without using $Y$, so that $\widetilde{X} \perp Y \mid X$ also holds.

Equivalently, one can sample conditionally:

$$\widetilde{X} \mid X = x \sim N\left((I - D\Sigma^{-1})x,\ 2D - D\Sigma^{-1}D\right),$$

where $x$ is treated as a column vector. The conditional covariance is positive semidefinite precisely under the same feasibility condition $2\Sigma - D \succeq 0$. As in fixed-X knockoffs, larger $s_j$ makes $\widetilde{X}_j$ less correlated with $X_j$, and hence usually improves the ability to distinguish original features from their knockoffs. When the features are standardized, the same semidefinite-programming (SDP) choice used in the fixed-X section,

$$\min_{0 \le s_j \le 1} \sum_{j=1}^{p} |1 - s_j| \qquad \text{subject to} \qquad 2\Sigma - \operatorname{diag}(s) \succeq 0,$$

tries to make each original/knockoff pair as different as the covariance structure allows.

## 8. Sequential Conditional Independent Pairs and Structured Feature Laws

**Route: Advanced**

Gaussian knockoffs are only one example. The sequential conditional independent pairs (SCIP) construction rests on a local characterization of pairwise exchangeability: $[X, \widetilde{X}]$ is invariant under all coordinate swaps if and only if, for every $j$, the pair $(X_j, \widetilde{X}_j)$ is exchangeable conditional on $(X_{-j}, \widetilde{X}_{-j})$. Response independence is then handled by constructing $\widetilde{X}$ without using $Y$. This characterization explains why independent features are easy. If $X_1, \dots, X_p$ are mutually independent, an independent copy $\widetilde{X} \overset{d}{=} X$ already has the required conditional exchangeability.

For a general known feature distribution, the SCIP construction of Candès et al. [38] builds knockoffs one coordinate at a time. At step $j$, it samples

$$\widetilde{X}_j \sim \mathcal{L}(X_j \mid X_{-j}, \widetilde{X}_1, \dots, \widetilde{X}_{j-1}),$$

where the conditional law is computed under the augmented joint distribution implied by the previous SCIP steps, not merely under the original feature law. The construction is recursive, but its invariant is simple: after step $j$, the partially augmented vector remains exchangeable under swaps of the completed coordinate pairs $(X_k, \widetilde{X}_k)$, $k \le j$.

Here is the finite-state proof. Let $P_{j-1}$ denote the joint mass of $(X_1, \dots, X_p, \widetilde{X}_1, \dots, \widetilde{X}_{j-1})$, and assume by induction that $P_{j-1}$ is symmetric under swaps of the completed pairs $(X_k, \widetilde{X}_k)$, $k < j$. The SCIP sampling rule gives

$$\mathbb{P}(\widetilde{X}_j = \widetilde{x}_j \mid X_{1:p} = x_{1:p}, \widetilde{X}_{1:j-1} = \widetilde{x}_{1:j-1}) = \frac{P_{j-1}(x_{-j}, \widetilde{x}_j, \widetilde{x}_{1:j-1})}{\sum_u P_{j-1}(x_{-j}, u, \widetilde{x}_{1:j-1})}.$$

These conditional formulas apply when the denominator is positive. On conditioning states of probability zero, the conditional distribution may be chosen arbitrarily. Therefore the joint mass after step $j$ is

$$P_j(x_{1:p}, \widetilde{x}_{1:j}) = \frac{P_{j-1}(x_{-j}, x_j, \widetilde{x}_{1:j-1}) P_{j-1}(x_{-j}, \widetilde{x}_j, \widetilde{x}_{1:j-1})}{\sum_u P_{j-1}(x_{-j}, u, \widetilde{x}_{1:j-1})}.$$

The displayed expression is symmetric in $x_j$ and $\widetilde{x}_j$, and the induction hypothesis preserves symmetry in the earlier completed pairs. Thus induction gives pairwise exchangeability for all completed pairs by the end of the procedure.

The usefulness of SCIP depends on whether these augmented conditional laws are tractable. Without structure, exact SCIP can be expensive. For example, in a dense Ising-type model on $\{-1,1\}^p$,

$$\mathbb{P}(X=x) \propto \exp\left\{\sum_{k<\ell}\omega_{k\ell}x_kx_\ell + \sum_{k=1}^{p} h_kx_k\right\},$$

the recursive masses $P_{j-1}$ in the SCIP denominator may require high-dimensional normalization after previous knockoffs have been introduced. The algorithm is exact in principle, but unstructured exact sampling can scale poorly unless additional graphical structure or Markov chain Monte Carlo (MCMC) machinery is exploited.

Markov models are a useful structured case because conditional distributions are local. Suppose $X_1,\dots,X_p$ form a first-order Markov chain on state space $\mathcal{X}$, with initial mass $\pi_1$, transition kernels $Q_j$, and joint mass

$$\pi_1(x_1)\prod_{j=2}^{p} Q_j(x_j \mid x_{j-1}).$$

For an interior coordinate $2 \le j \le p-1$, the ordinary full conditional is obtained by viewing this joint mass as a function of $x_j$ while holding $x_{-j}$ fixed:

$$\mathbb{P}(X_j = u \mid X_{-j} = x_{-j}) \propto Q_j(u \mid x_{j-1})Q_{j+1}(x_{j+1} \mid u).$$

At the boundaries the corresponding factors are $\pi_1(u)Q_2(x_2 \mid u)$ for $j=1$ and $Q_p(u \mid x_{p-1})$ for $j=p$.

In SCIP, the same locality appears, but each step also includes correction factors from previously generated knockoffs. For $p=4$, the first update is

$$\mathbb{P}(\widetilde{X}_1 = u \mid X_{2:4} = x_{2:4}) = \frac{\pi_1(u)Q_2(x_2 \mid u)}{N_1(x_2)}, \qquad N_1(v) = \sum_{r\in\mathcal{X}} \pi_1(r)Q_2(v \mid r).$$

The second update, conditioning on $(X_{-2}, \widetilde{X}_1)$, is

$$\mathbb{P}(\widetilde{X}_2 = u \mid X_{-2} = x_{-2}, \widetilde{X}_1 = \widetilde{x}_1) = \frac{Q_2(u \mid x_1)Q_3(x_3 \mid u)Q_2(u \mid \widetilde{x}_1)/N_1(u)}{N_2(x_3)},$$

where

$$N_2(v) = \sum_{r\in\mathcal{X}} Q_2(r \mid x_1)Q_3(v \mid r)\frac{Q_2(r \mid \widetilde{x}_1)}{N_1(r)}.$$

The later updates have the same form:

$$\mathbb{P}(\widetilde{X}_3 = u \mid X_{-3} = x_{-3}, \widetilde{X}_{1:2} = \widetilde{x}_{1:2}) = \frac{Q_3(u \mid x_2)Q_4(x_4 \mid u)Q_3(u \mid \widetilde{x}_2)/N_2(u)}{N_3(x_4)},$$

with

$$N_3(v) = \sum_{r\in\mathcal{X}} Q_3(r \mid x_2)Q_4(v \mid r)\frac{Q_3(r \mid \widetilde{x}_2)}{N_2(r)},$$

and

$$\mathbb{P}(\widetilde{X}_4 = u \mid X_{-4} = x_{-4}, \widetilde{X}_{1:3} = \widetilde{x}_{1:3}) = \frac{Q_4(u \mid x_3)Q_4(u \mid \widetilde{x}_3)/N_3(u)}{N_4},$$

where

$$N_4 = \sum_{r \in \mathcal{X}} Q_4(r \mid x_3) \frac{Q_4(r \mid \widetilde{x}_3)}{N_3(r)}.$$

The important point is locality: each update uses neighboring original coordinates, previous knockoffs, and recursive normalizing factors, not all coordinates in an unstructured way. This locality is what makes exact knockoff sampling feasible for some non-Gaussian structured features. Sesia et al. [159] carry this out for hidden Markov models in GWAS, where linkage disequilibrium (LD) between single-nucleotide polymorphisms (SNPs) is well approximated by a Markov chain on chromosomes.

## 9. Computation and Approximation

**Route: Advanced**

CRTs are conceptually direct but computationally expensive. Testing $p$ features with $B$ Monte Carlo resamples requires roughly $p(B+1)$ evaluations of the feature-importance statistic. If the statistic itself is a predictive model fitted by cross-validation (CV), this can be prohibitive. Several variants reduce this cost: Liu et al. [128] introduce a *distillation* CRT that screens out features and amortizes intermediate computations, often matching the power of the naive CRT with one or two orders of magnitude fewer statistic evaluations. Knockoffs pay a different cost: construct one augmented design and compute one set of feature statistics, but accept the need to build a valid knockoff generator.

Figure 9.5 compares those computational profiles in one controlled simulation. The panels report empirical FDR, empirical power, and the relative number of statistic evaluations.

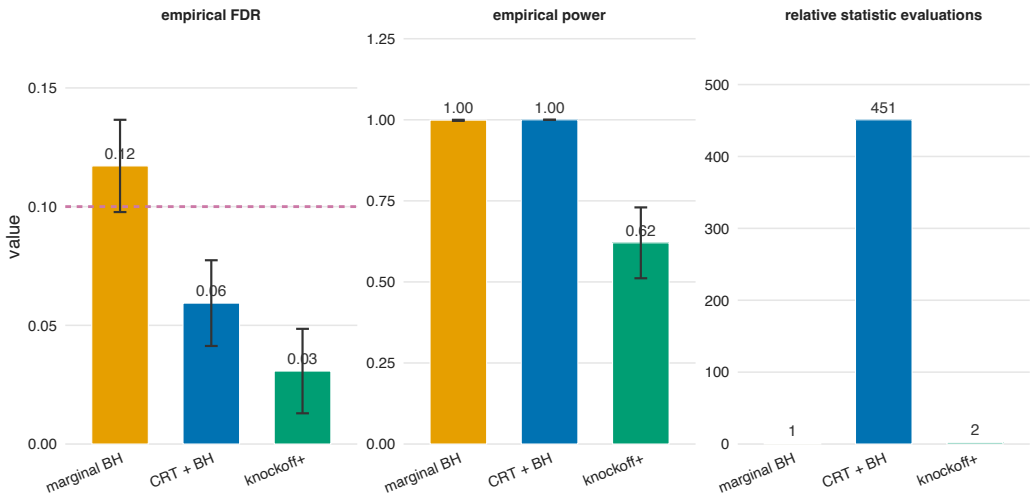



FIGURE 9.5. A Gaussian model-X simulation with $n = 220$, $p = 40$, 14 signals of size 1, AR(1) correlation $\rho = 0.5$, and FDR target $q = 0.10$. The empirical FDR and empirical power are averaged over 64 independent simulated data sets; their vertical bars are pointwise 95 percent normal simulation intervals. CRT uses $B = 450$ conditional resamples per feature, while knockoff+ uses one knockoff copy and a signed marginal-correlation statistic. The right panel shows relative statistic evaluations, not wall-clock time.

In practice, the feature distribution is rarely known exactly. Model-X knockoffs are then approximate. One may fit a Gaussian graphical model, a hidden Markov model (159), a copula model, a normalizing flow, or a deep generative model trained to match the swap-invariance condition (155), and then sample knockoffs from the fitted distribution. Their FDR control depends on how closely the fitted construction preserves swap exchangeability.

Generator-based methods make this approximation problem explicit. Given observations $X_1, \dots, X_n$, introduce external randomness $V_i$, for example $V_i \sim N(0, I_p)$, and set

$$\widetilde{X}_i = f_\theta(X_i, V_i).$$

The parameter $\theta$ is trained so that the augmented sample is difficult to distinguish from its coordinate-swapped versions. If $J$ is a discrepancy between distributions or empirical samples, a schematic objective is

$$\sum_{j=1}^{p} J\left([X, \widetilde{X}], [X, \widetilde{X}]_{\mathrm{swap}(\{j\})}\right).$$

Common discrepancies include maximum mean discrepancy (MMD), energy distance, Wasserstein distance, and Jensen–Shannon divergence. Optimizing this criterion can produce useful knockoff generators. The exact FDR theorem still requires exact pairwise exchangeability; an approximate construction needs a separate quantitative robustness bound.

Useful diagnostics include:

| | |
|---|---|
| swap discrepancies: | compare $[X, \widetilde{X}]$ with its coordinate-swapped versions, |
| two-sample distances: | MMD, energy distance, or Wasserstein distance, |
| adversarial checks: | train a classifier to distinguish swapped from unswapped pairs, |
| conditional checks: | inspect whether $\widetilde{X}_j$ preserves dependence on $X_{-j}$. |

These checks look for departures from swap exchangeability. If the learned construction does not approximately preserve the coordinatewise conditional sign-flip law for null statistics after their magnitudes and the nonnull statistics are fixed, negative knockoff wins need not estimate false discoveries reliably. Marginal or unconditional sign symmetry alone is insufficient. Barber et al. [11, Theorem 1] give such a bound for generators that are pairwise exchangeable with respect to specified approximate feature conditionals. Write $p_j(\cdot \mid x_{-j})$ and $\widehat{p}_j(\cdot \mid x_{-j})$ for the true and approximate conditional densities, with compatible supports. Here these functions are densities, not the CRT p-values. Their empirical discrepancy is

$$\widehat{\mathrm{KL}}_j = \sum_{i=1}^{n} \log \frac{p_j(X_{ij} \mid X_{i,-j})\, \widehat{p}_j(\widetilde{X}_{ij} \mid X_{i,-j})}{\widehat{p}_j(X_{ij} \mid X_{i,-j})\, p_j(\widetilde{X}_{ij} \mid X_{i,-j})}.$$

Under the generator and statistic conditions of that theorem, knockoff+ satisfies, for every $\epsilon \geq 0$,

$$\mathrm{FDR} \leq q\, e^{\epsilon} + \mathbb{P}\left(\max_{j \in \mathcal{I}_0} \widehat{\mathrm{KL}}_j > \epsilon\right).$$

Thus control near $q$ requires these discrepancies to be small simultaneously over null features with high probability. The discrepancy is a random sum over observations, not simply a population KL divergence between two feature models. Because it involves the unknown true conditionals, the earlier empirical diagnostics do not directly compute or certify this bound. Estimated conditionals trained on independent unlabeled data can be treated as fixed after conditioning on that training data; using the testing observations to fit the generator needs its own argument.

## 10. Multilayer Knockoffs

**Route: Research frontier**

The model-X knockoff filter also extends to structured families of variables. Katsevich and Sabatti [103] construct the *multilayer knockoff filter* (MKF) as a single procedure that simultaneously controls FDR at several resolutions through one joint threshold vector, not as a separate filter run independently at each layer. In genomics, for instance, one may want to report discoveries at the SNP, gene, and pathway levels. The output should be one set of selected variables, but that same set should induce reliable discoveries at each resolution. The exposition below is a simplified description meant to convey the mechanism; the formal algorithm and FDR theorem belong to Katsevich and Sabatti [103].

**Why another layer changes the target.** Let $R \subseteq \{1, \dots, p\}$ be a selected set of variables. A layer $\ell$ is a fixed partition of the variables into groups $\mathcal{G}_1^{(\ell)}, \dots, \mathcal{G}_{G_\ell}^{(\ell)}$. Write $\pi_\ell(j)$ for the layer-$\ell$ group containing variable $j$. The variable set $R$ induces the layer-$\ell$ group discoveries

$$\mathcal{R}_\ell(R) = \{\pi_\ell(j) : j \in R\}.$$

Thus a variable-level discovery automatically reports discoveries at coarser levels, such as genes, pathways, or brain regions.

This induced reporting is why variable-level FDR is not enough. In a two-layer example, let $V_{\text{ind}}(R)$ be the number of selected null variables and let $V_{\text{grp}}(R)$ be the number of selected null groups. The corresponding realized FDPs are

$$\text{FDP}_{\text{ind}}(R) = \frac{V_{\text{ind}}(R)}{|R| \vee 1}, \qquad \text{FDP}_{\text{grp}}(R) = \frac{V_{\text{grp}}(R)}{|\mathcal{R}_{\text{grp}}(R)| \vee 1}.$$

For an exact two-group calculation, take

$$\mathcal{G}_1 = \{1, 2, 3\}, \qquad \mathcal{G}_2 = \{4, 5, 6\},$$

and suppose variable (3) is the only nonnull variable. If the selected set is $R = \{1, 3, 4\}$, then variables (1) and (4) are false discoveries, so

$$\text{FDP}_{\text{ind}}(R) = \frac{2}{3}.$$

Both groups are reported, but only $\mathcal{G}_2$ is a null group because $\mathcal{G}_1$ contains the nonnull variable (3). Therefore

$$\mathcal{R}_{\text{grp}}(R) = \{\mathcal{G}_1, \mathcal{G}_2\}, \qquad \text{FDP}_{\text{grp}}(R) = \frac{1}{2}.$$

The two FDPs differ because their denominators count different discovery units. False variable discoveries often scatter across many null groups, while true variable discoveries may cluster inside a few non-null groups. In that common case $V_{\text{grp}}(R)$ can be close to $V_{\text{ind}}(R)$, while $|\mathcal{R}_{\text{grp}}(R)|$ can be much smaller than $|R|$. Ignoring the harmless $\vee 1$ terms when there are discoveries, this gives

$$\text{FDP}_{\text{grp}}(R) = \frac{V_{\text{grp}}(R)}{|\mathcal{R}_{\text{grp}}(R)|} \approx \frac{V_{\text{ind}}(R)}{|\mathcal{R}_{\text{grp}}(R)|} = \frac{|R|}{|\mathcal{R}_{\text{grp}}(R)|}\,\text{FDP}_{\text{ind}}(R).$$

The last equality uses $V_{\text{ind}}(R) = |R|\,\text{FDP}_{\text{ind}}(R)$, ignoring the $\vee 1$ term when $R \neq \varnothing$. Thus group FDP can be much larger than variable FDP when several selected variables fall in each selected group. The multilayer problem is therefore to return one variable set $R$ whose induced discoveries have controlled FDR at every reported layer.

**Group nulls and group statistics.** A layer-$\ell$ group $g$ is a *group null* if all variables in that group are null at the finest level:

$$g \in \mathcal{N}_\ell \quad \Longleftrightarrow \quad H_j : Y \perp\!\!\!\perp X_j \mid X_{-j} \text{ is true for every } j \in \mathcal{G}_g^{(\ell)}.$$

By Lemma 9.8, swapping all coordinates of such a null group preserves the joint data law: compose the single-coordinate null swaps. This is the symmetry needed for a group knockoff comparison. For example, a gene group is a null group if none of the SNPs assigned to that gene is conditionally associated with the response. Discovering such a group is a false group discovery; discovering a group with at least one non-null member is counted as a true group discovery at that layer.

The MKF requires layer-$\ell$ group statistics $W_1^{(\ell)}, \ldots, W_{G_\ell}^{(\ell)}$. These are group-level analogues of the usual knockoff statistics. For each layer, one constructs original and knockoff group importance scores $(Z_g^{(\ell)}, \widetilde{Z}_g^{(\ell)})$, then applies an antisymmetric comparison, for example

$$W_g^{(\ell)} = Z_g^{(\ell)} - \widetilde{Z}_g^{(\ell)}.$$

Large positive values mean the original group beats its knockoff; large negative values mean the knockoff group wins. The formal validity condition is a group sign-flip property: conditional on the magnitudes and the non-null group statistics, the signs of null group statistics behave as independent fair coins, after zero statistics are omitted. The aggregation must respect the group swap. For example, summing valid variable-level $W_j$'s within each group does change sign under a whole-group swap; summing their absolute values does not. Dedicated group importance scores can improve power by using the group structure during fitting. The theorem below applies when the layer-specific group knockoff construction and group importance statistic satisfy the flip-sign and compatibility conditions of Katsevich and Sabatti [103].

**Joint FDP estimate and joint thresholds.** Use positive observed group-statistic magnitudes as the threshold grid at each layer, together with $\infty$, so zero statistics are never selected. Given a threshold vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)$, MKF first keeps the variables whose groups pass every layer's threshold:

$$\widehat{R}(\boldsymbol{\tau}) = \{j : W_{\pi_\ell(j)}^{(\ell)} \geq \tau_\ell \text{ for every } \ell = 1, \ldots, L\}.$$

The induced layer-$\ell$ discoveries are

$$\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau}) \;=\; \mathcal{R}_\ell(\widehat{R}(\boldsymbol{\tau})).$$

Let $\alpha_1, \ldots, \alpha_L$ be the target FDR levels for the layers. For MKF($c$)+, the constant $c \geq 1$ scales the knockoff+ numerator: $c = 1$ is the uncorrected version, while $c = c_{\text{kn}}$, with $c_{\text{kn}} = 1.93$ in the theorem below, gives the formal worst-case guarantee. The layer-$\ell$ FDP estimate has the form

$$\widehat{\text{FDP}}^{(\ell)}(\boldsymbol{\tau}) = \frac{c\{1 + \#\{g : W_g^{(\ell)} \leq -\tau_\ell\}\}}{|\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau})| \vee 1}.$$

The numerator is layer-specific: negative group wins at layer $\ell$ are used as proxies for false positive group wins at that same layer. The denominator is joint: it counts only groups hit by variables that pass all layers. Thus negative wins are counted within each layer, but discoveries are counted only after all layer thresholds have been applied. This is the basic bookkeeping difference between MKF and running ordinary knockoff filters independently at each layer.

The procedure searches for the threshold vector $\boldsymbol{\tau}^*$ that is smallest in every coordinate among those satisfying $\widehat{\text{FDP}}^{(\ell)}(\boldsymbol{\tau}^*) \leq \alpha_\ell$ for every layer. Smaller thresholds are more liberal, so this is the most permissive threshold vector that satisfies all layer-wise FDP estimates. Katsevich and Sabatti [103] show that this threshold vector is well defined and can be found by an iterative search over the threshold grid. If no vector satisfies the inequalities, report no discoveries.

Otherwise the final output is $\widehat{R} = \widehat{R}(\boldsymbol{\tau}^*)$, and the layer-$\ell$ FDR is the expectation of the FDP of $\mathcal{R}_\ell(\widehat{R})$.

**Theorem 9.12: Multilayer knockoff FDR control, cited result [103]**

Assume the model-X exchangeability conditions of this chapter and suppose each layer-$\ell$ group statistic $W_g^{(\ell)}$ satisfies the formal flip-sign and compatibility conditions of Katsevich and Sabatti [103]. For the MKF($c$)+ procedure of that paper,

$$\mathrm{FDR}^{(\ell)}(\widehat{R}) \;\le\; \frac{c_{\mathrm{kn}}}{c}\,\alpha_\ell, \qquad c_{\mathrm{kn}} = 1.93, \qquad \text{for every layer } \ell.$$

In particular, running the corrected version with $c = c_{\mathrm{kn}}$, or running the uncorrected $c = 1$ version at levels $\alpha_\ell / c_{\mathrm{kn}}$, gives the formal guarantee $\mathrm{FDR}^{(\ell)} \le \alpha_\ell$.

Proof idea. Fix a layer $\ell$, and let

$$V_\ell^+(t) = \#\{g \in \mathcal{N}_\ell : W_g^{(\ell)} \ge t\}, \qquad V_\ell^-(t) = \#\{g \in \mathcal{N}_\ell : W_g^{(\ell)} \le -t\}.$$

At the selected threshold $\boldsymbol{\tau}^*$, the number of false layer-$\ell$ discoveries is at most $V_\ell^+(\tau_\ell^*)$, while the negative-tail count in $\widehat{\mathrm{FDP}}^{(\ell)}$ is at least $V_\ell^-(\tau_\ell^*)$. Since the selected vector is acceptable,

$$\mathrm{FDP}^{(\ell)}(\widehat{R}) \le \frac{\alpha_\ell}{c}\,\frac{V_\ell^+(\tau_\ell^*)}{1 + V_\ell^-(\tau_\ell^*)}.$$

The difficulty is that $\tau_\ell^*$ was chosen jointly using all layers, so it is not a one-layer stopping time. Katsevich and Sabatti's key decoupling step is to take the supremum over the layer-$\ell$ threshold:

$$\mathbb{E}\left[\frac{V_\ell^+(\tau_\ell^*)}{1 + V_\ell^-(\tau_\ell^*)}\right] \le \mathbb{E}\left[\sup_{t>0} \frac{V_\ell^+(t)}{1 + V_\ell^-(t)}\right].$$

This supremum bound decouples the layers: after we upper-bound by the worst case over $t$, the thresholds chosen by the other layers no longer appear. Now only the layer-$\ell$ null signs matter. Conditional on magnitudes and nonnull group statistics, order the nonzero null group statistics by decreasing magnitude; their signs are independent fair coins. The last display becomes the expected maximum, over a simple symmetric random walk, of the ratio "positive steps" divided by 1+ "negative steps." The random-walk bound in Katsevich and Sabatti [103] is at most $c_{\mathrm{kn}} = 1.93$, giving the theorem. Using $c = c_{\mathrm{kn}}$ converts this constant-factor bound into target-level FDR control.

Simulations in Katsevich and Sabatti [103] suggest that the uncorrected version $c = 1$ often controls FDR in practice. The constant $c_{\mathrm{kn}}$ is the worst-case correction required by the theorem, not a claim about typical behavior.

**Remark 9.13: Joint thresholds are essential**

The multilayer guarantee crucially does *not* come from running separate layer-wise filters and intersecting the resulting lifted rejection sets. A marginally valid layer-wise procedure need not remain calibrated for the post-intersection target: shrinking a rejection set can drop true discoveries while retaining false ones, inflating the FDP of the intersected output. The simultaneous threshold-choice step in the multilayer filter is what links the layers correctly; intersection on its own is a common fallacy. Exercise 9.22 gives an explicit counterexample.

When variables within a group are tightly correlated (LD blocks in genomics, for example), group-level discoveries are typically the scientifically meaningful target, and the multilayer filter's joint thresholding delivers them at a controlled per-layer FDR while still reporting variable-level findings inside each rejected group.

## 11. Assumptions in Plain Language

**Route: Core**

The procedures in this chapter target several related statistical nulls. The CRT and model-X knockoffs both target the conditional independence null $Y \perp\!\!\!\perp X_j \mid X_{-j}$ without assuming a parametric model for $Y \mid X$; their validity rests on knowing or accurately approximating the feature distribution. Fixed-X knockoffs target the linear-model coefficient null $\beta_j = 0$ inside a Gaussian fixed-design model; their validity rests on geometric properties of the design and on Gaussian errors, not on a feature distribution.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | For fixed-$X$ knockoffs, verify the two Gram-matrix identities exactly. For a fitted model-X generator, compare original–knockoff data with coordinate-swapped versions using the swap checks introduced above, and verify that the construction did not use $Y$. |
| **What these data cannot show** | These diagnostics cannot establish the exact conditional feature law or pairwise exchangeability. Passing marginal, covariance, or classifier checks is not a proof of finite-sample FDR control. If the feature law cannot be defended, do not attach the CRT or model-X finite-sample guarantee to the reported findings. |

The CRT is exact when the conditional law $\mathcal{L}(X_j \mid X_{-j})$ is correct and independent resamples are drawn conditionally on the observed $X_{-j}$, with symmetric statistic fitting and evaluation. A complicated statistic does not damage validity if used in the same way for every observed or resampled data set. A wrong resampling law can damage validity.

Fixed-X knockoffs are exact for the Gaussian linear fixed-design model when the knockoff matrix satisfies the Gram equations and the feature statistic is sufficient and antisymmetric. The construction is geometric and depends on the design $X$; it does not claim validity for arbitrary random-design conditional nulls.

Model-X knockoffs require a knockoff copy that is pairwise exchangeable with $X$ and generated without using $Y$. Row shuffling, marginal resampling, or poorly fitted generative models can break this exchangeability. For knockoffs fitted by graphical models, normalizing flows, or deep generators, Barber et al. [11] study how approximation affects FDR. Related generator-based constructions appear in Romano et al. [155].

Multilayer knockoffs inherit the assumptions of model-X knockoffs: a correct or accurately estimated feature law, and valid pairwise exchangeability. No separate within-layer or across-layer independence assumption is added, but the layer-specific group statistics must satisfy the formal flip-sign and compatibility conditions of Katsevich and Sabatti [103]. The formal worst-case statement includes the $c_{\mathrm{kn}} = 1.93$ correction constant in Theorem 9.12.

## 12. Bibliographic Notes

**Route: Core**

Fixed-X knockoffs and the original knockoff filter were introduced by Barber and Candès [9]. Model-X knockoffs, the CRT viewpoint, Gaussian model-X construction, and SCIP were developed by Candès et al. [38]. Robustness to approximate knockoffs is studied by Barber et al. [11]. The distinction between a bare mirror/knockoff+ threshold and a filter supported by conditional sign flips is examined under several dependent score laws by Zhang [201]. Metropolized knockoff sampling provides one route to exact or controlled sampling for complex distributions [17]. Hidden Markov model knockoffs for GWAS are due to Sesia et al. [159], and deep generative knockoffs are developed by Romano et al. [155]. Computationally efficient CRT variants, including the distillation CRT used to amortize statistic evaluations across features, are studied by Liu et al. [128]. A frontier extension to privacy-constrained inference is the differentially private model-X knockoff filter of Pournaderi and Xiang [145], which uses a Johnson–Lindenstrauss projection to privatize the knockoff matrix while preserving approximate exchangeability up to bounds determined by the projection dimension. Multilayer and group-aware knockoffs that exploit structured families are developed in Katsevich and Sabatti [103].

## 13. Exercises

**Route: Core**

### Basic.

**Exercise 9.14: Deriving the conditional resampling law**

In the Gaussian model of Example 9.3, assume $|\rho| < 1$ and derive the conditional law $X_1 \mid X_2 = x_2$ from the joint normal density. Identify its conditional mean and variance separately.

**Exercise 9.15: Pairwise exchangeability**

For a model-X knockoff copy with $p = 2$, write out explicitly the four pairwise-exchangeability identities corresponding to $S = \varnothing, \{1\}, \{2\}, \{1, 2\}$, using the coordinate order $(X_1, X_2, \widetilde{X}_1, \widetilde{X}_2)$.

### Intermediate.

**Exercise 9.16: Signed-max statistic**

Let $Z_j$ and $\widetilde{Z}_j$ be lasso entry times for the original and knockoff variables. Assume that the fitted path, including its handling of nonunique solutions, is equivariant under column swaps, as required in the chapter. Use $\operatorname{sign}(0) = 0$. Show that

$$W_j = \max\{Z_j, \widetilde{Z}_j\} \operatorname{sign}(Z_j - \widetilde{Z}_j)$$

is antisymmetric under swapping $X_j$ and $\widetilde{X}_j$.

**Exercise 9.17: Knockoff+ threshold**

For

$$W = (4.0,\ 2.2,\ -1.7,\ 0.8,\ -0.6,\ 0.4)$$

and target $q = 0.2$, compute the knockoff and knockoff+ thresholds by hand, using the candidate thresholds in $\{|W_j| : |W_j| > 0\}$. Identify the rejected coordinates under each rule, using the convention that the threshold is $\infty$ if no candidate value satisfies the threshold inequality.

**Exercise 9.18: Gaussian model-X conditionals**

Let $(X, \widetilde{X})$ be jointly Gaussian with mean zero, where $\Sigma \succ 0$, $D$ is diagonal, $D \succeq 0$, and $2\Sigma - D \succeq 0$. Its block covariance is

$$\begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix}.$$

Use the Gaussian conditioning formula to derive the conditional mean of $\widetilde{X} \mid X = x$. For the conditional covariance, use the Schur-complement identity in the proof of Proposition 9.4, and state the resulting conditional law.

## Computational.

**Exercise 9.19: CRT simulation**

Simulate the Gaussian example in Figure 9.2: $n = 450$, $\rho = 0.65$, $Y = X_2 + \varepsilon$, and $\varepsilon \sim N(0, 0.7^2)$. Use the absolute sample correlation $|\widehat{\rho}_{1Y}|$ as the statistic and test at level 0.05. With, say, $B = 999$ resamples inside each test and at least 500 outer repetitions, compare a marginal randomization test $X_1^* \sim N(0,1)$ with the CRT draw $X_1^* \mid X_2 = x_2 \sim N(\rho x_2, 1 - \rho^2)$. Estimate the rejection probability of both tests under the conditional null, and report Monte Carlo SEs for both estimates.

**Exercise 9.20: Gaussian knockoff implementation**

Write a function that generates Gaussian model-X knockoffs for $X \sim N(0, \Sigma)$. For a concrete check, take $p = 40$ and an AR(1) covariance $\Sigma_{jk} = 0.5^{|j-k|}$. Choose $D = \mathrm{diag}(s)$ with $s_j = \min\{1, 2\lambda_{\min}(\Sigma)\}$, and verify empirically that the sample covariance of $(X, \widetilde{X})$ is close to the target block covariance. Use an eigenvalue square root to sample the conditional Gaussian law, since its covariance can be singular.

Then simulate a linear model with $n = 220$, the first 14 coefficients equal to 1, all other coefficients zero, and independent standard normal errors. Use a swap-equivariant statistic, apply the knockoff+ threshold at $q = 0.1$, and use at least 200 independent repetitions. Report empirical FDR and average power, defining power in each repetition as the fraction of the 14 signals rejected, together with Monte Carlo SEs for both averages. State the statistic you use.

**Exercise 9.21: Fixed-X versus model-X targets under misspecification**

Generate independent mean-zero Gaussian design rows with a positive definite feature covariance, taking $n \geq 2p$, and choose two strongly correlated features $X_j$ and $X_k$. Generate $Y_i = \sin(2X_{ik}) + \varepsilon_i$, with independent $\varepsilon_i \sim N(0,1)$. For the realized full-rank design, set

$$\mu_i = \sin(2X_{ik}), \qquad \beta^{\mathrm{proj}}(X) = (X^\top X)^{-1} X^\top \mu.$$

Compute this finite-design projection and compare its $j$th coefficient with the conditional-independence null $Y \perp X_j \mid X_{-j}$. Distinguish this realized-design target from the population linear projection coefficient.

Apply (a) fixed-X knockoffs as a working linear-model analysis and (b) Gaussian model-X knockoffs, and compare their rejection sets. The nonlinear mean is generally outside the column space of $X$, so

the fixed-X FDR theorem does not justify the first analysis here. Finally, verify that the coefficient and conditional-independence nulls coincide under a correctly specified linear model with jointly Gaussian features of positive definite covariance and independent Gaussian noise.

**Advanced.**

**Exercise 9.22: Multilayer knockoff intersection – a fallacy**

This is a stylized rejection-set counterexample; constructing actual knockoff statistics with this exact behavior is not required. Consider the following probability model for two procedures, one at the variable level and one at the group level. The lifted group-level rejection set is the union of variables in the rejected groups. The task is to check whether intersecting this set with the variable-level rejection set preserves FDR control. In the model below, take four variables grouped into two layer-2 groups ($\{1,2\}$ and $\{3,4\}$), where variables 1 and 3 are true signals and variables 2 and 4 are nulls; both layer-2 groups are then non-null, so a layer-2 rejection of either group is a true group discovery. Fix $\alpha = 0.05$ and pick an event probability $\pi$ with $\alpha < \pi < 2\alpha$, e.g. $\pi = 0.08$. Let the following two events exhaust the possible outcomes of the variable-level and group-level procedures:

- Event $A$ (with probability $\pi$): at layer 1 the procedure rejects $\{1,4\}$ (one signal, one null), and at layer 2 the procedure rejects only group $\{3,4\}$ (a non-null group). The lifted layer-2 rejection set is $\{3,4\}$, and the intersection is $R = \{1,4\} \cap \{3,4\} = \{4\}$ – a null variable.
- Event $B$ (with probability $1-\pi$): no rejections at either layer.

Compute the layer-1 FDR of the variable-level procedure, the layer-2 FDR of the group-level procedure, and the variable-level FDR of the intersection, each as an expected FDP. Verify that the two separate procedures each control their layer's FDR at level $\alpha$, while the intersected output does not.

**Exercise 9.23: Row shuffling is not a knockoff**

Generate correlated Gaussian features, for example with $\Sigma_{jk} = 0.6^{|j-k|}$, and create a fake knockoff by applying an independent random row permutation to each feature column. Check pairwise exchangeability empirically by comparing $[X, \widetilde{X}]$ with $[X, \widetilde{X}]_{\text{swap}(\{j\})}$ for a fixed coordinate $j$, using a covariance comparison or a two-sample discrepancy. Describe which same-row dependence relationships are broken.

CHAPTER 10

# Conformal Prediction and Conformal P-Values

Prediction intervals are often reported as though the fitted model were correct. In modern applications the fitted model may be a random forest, a neural network, a lasso, a quantile regression engine, or an ensemble selected by CV. Even when the predictor is useful, it is rarely credible to treat the entire fitted model as known and correctly specified.

Conformal prediction separates two tasks. The first task is prediction: build an algorithm that produces small errors. The second task is calibration: convert the algorithm's errors on exchangeable calibration data into a prediction set with finite-sample coverage. The coverage statement is distribution-free and marginal over the next observation. Model quality still matters, but it matters through the size and shape of the prediction set, not through the validity of the rank argument.

Roadmap. The chapter has three reading routes.

- **Core route.** We develop the rank argument, construct prediction sets, adapt their widths to the prediction problem, and test for unusual observations.
- **Advanced route.** We study ways to reuse observations, combine prediction sets, and assess guarantees when the reporting level is chosen after seeing the data.
- **Appendix B route.** Appendix B examines coverage at a given covariate value (Section 1) and after selecting which cases to report (Section 3), data arriving over time (Section 5), changes in the data distribution (Section 6), and losses beyond miscoverage (Section 7).

The basic prerequisites are order statistics, conditional probability, and the distinction between a prediction interval for a future response and a confidence interval for a parameter. The multiple-testing sections use BH and PRDS from Chapters 5 and 6; the e-prediction and online-betting sections use Chapter 8.

## 1. The Target

**Route: Core**

Let

$$Z_i = (X_i, Y_i), \qquad i = 1, \dots, n+1,$$

where $X_i \in \mathcal{X}$ is a covariate and $Y_i \in \mathcal{Y}$ is a response. Write $D_n = (Z_1, \dots, Z_n)$ for the observed data and fix a miscoverage level $\alpha \in (0, 1)$. The goal is to construct a set

$$C_n(X_{n+1}) = C(D_n, \alpha, X_{n+1})$$

such that

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha.$$

The probability is over the training data and the test point. This is *marginal* coverage, not conditional coverage at each fixed covariate value.

A trivial procedure can satisfy the display by returning $\mathcal{Y}$ with probability $1 - \alpha$ and the empty set otherwise. Conformal prediction is interesting because it can wrap a useful prediction algorithm and produce sets that adapt to the actual difficulty of the prediction problem.

Recall from Example 8.8 in Chapter 8 that exchangeability means that relabeling the observations leaves their joint law unchanged. This is the core assumption for ordinary conformal prediction:

$$(Z_1, \dots, Z_{n+1}) \overset{d}{=} (Z_{\pi(1)}, \dots, Z_{\pi(n+1)})$$

for every permutation $\pi$. Independent identically distributed data imply exchangeability, but exchangeability is the property used by the proofs.

We first show how exchangeability makes an observation's rank useful for prediction. The constructions that follow preserve this symmetry in different ways. Each section states the resulting guarantee and the assumptions it requires.

## 2. The Rank Argument

**Route: Core**

The essential idea appears before covariates enter. Suppose

$$Y_1, \dots, Y_{n+1}$$

are exchangeable real-valued observations. Let

$$k = \lceil (1-\alpha)(n+1) \rceil$$

and let $Y_{(1)} \le \cdots \le Y_{(n)}$ be the order statistics of the first $n$ observations. Set

$$c_n = Y_{(k)}$$

when $k \le n$, with $c_n = \infty$ if $k = n+1$. Then

$$\mathbb{P}(Y_{n+1} \le c_n) \ge 1 - \alpha.$$

Figure 10.1 makes the finite-sample arithmetic explicit. With $n = 24$ and $\alpha = 0.2$, $(1-\alpha)(n+1) = 0.8 \times 25 = 20$, so $k = \lceil 20 \rceil = 20$ and the no-ties coverage is $k/(n+1) = 20/25 = 0.8$. The cutoff $c_n$ therefore uses rank 20 among the 24 calibration observations, with the future observation supplying the twenty-fifth possible rank.

The proof below uses a principle that will recur: after we forget the labels of an exchangeable sample, each observation is equally likely to occupy any position in the sample. Conditioning on a summary that is unchanged by relabeling preserves this symmetry. The next lemma states the principle formally; its proof can be deferred on a first reading. We assume the sample spaces in this chapter are standard Borel spaces, as are all finite-dimensional Euclidean and finite spaces used in the examples; this ensures that the conditional distributions below exist. Let $(Z_1, \dots, Z_m)$ be an exchangeable random vector taking values in a space $\mathcal{Z}$, and define its *empirical distribution*

$$\widehat{P}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Z_i},$$

where $\delta_z$ denotes the point mass at $z$. Call a function $f$ of the data *permutation-symmetric* if $f(z_1, \dots, z_m) = f(z_{\sigma(1)}, \dots, z_{\sigma(m)})$ for every permutation $\sigma$ of $\{1, \dots, m\}$; the empirical distribution, the order statistics of real-valued data, and any sample average are examples.

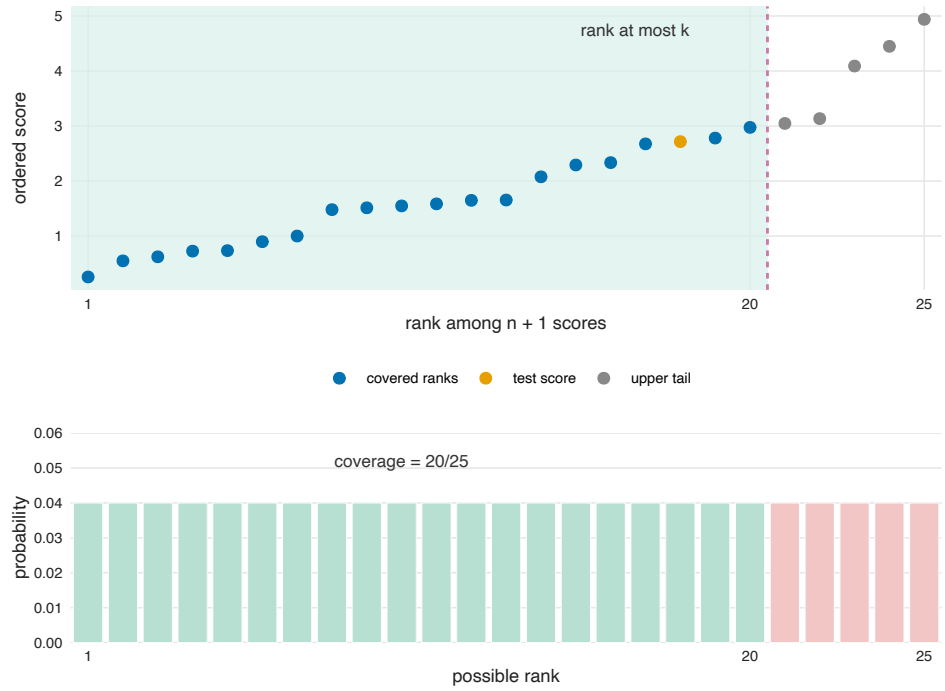



FIGURE 10.1. The conformal rank argument. Under exchangeability, the test score rank among the $n+1$ scores is uniform when scores are distinct, or when ties are broken by independent uniform random variables. The finite-sample correction uses $k = \lceil (1-\alpha)(n+1) \rceil$, not simply the empirical $1-\alpha$ quantile of the first $n$ scores. The cutoff is denoted $c_n$. The displayed example takes $n = 24$ and $\alpha = 0.2$, giving $k = 20$ and no-ties coverage exactly $20/25$.

**Lemma 10.1: Conditioning on a symmetric statistic**

Let $(Z_1, \ldots, Z_m)$ be exchangeable and let $f$ be a permutation-symmetric function of the data. Then the conditional law of $(Z_1, \ldots, Z_m)$ given $f(Z_1, \ldots, Z_m)$ is, almost surely, an exchangeable distribution. In particular, conditional on the empirical distribution $\widehat{P}_m$, each coordinate is a draw from $\widehat{P}_m$:

$$Z_i \mid \widehat{P}_m \;\sim\; \widehat{P}_m, \qquad i = 1, \ldots, m,$$

so that each $Z_i$ is uniformly distributed over the observed values when those values are distinct [5].

This statement describes each coordinate separately. It does not make the coordinates independent after conditioning on the empirical distribution.

**Route: Optional proof**

**Proof of Lemma 10.1**

Write $Z = (Z_1, \dots, Z_m)$ and, for a permutation $\sigma$, write $Z_\sigma = (Z_{\sigma(1)}, \dots, Z_{\sigma(m)})$. For the first claim, it suffices to show that, for every $\sigma$ and all measurable sets $A$ and $B$,

$$\mathbb{P}\{Z \in A,\ f(Z) \in B\} = \mathbb{P}\{Z_\sigma \in A,\ f(Z) \in B\},$$

since this display says that $Z$ and $Z_\sigma$ have the same joint law with $f(Z)$, and hence, by uniqueness of conditional distributions, the same conditional law given $f(Z)$ almost surely. Exchangeability gives $Z \overset{d}{=} Z_\sigma$, so

$$\mathbb{P}\{Z \in A,\ f(Z) \in B\} = \mathbb{P}\{Z_\sigma \in A,\ f(Z_\sigma) \in B\},$$

and permutation symmetry gives $f(Z_\sigma) = f(Z)$, so the right-hand side equals $\mathbb{P}\{Z_\sigma \in A,\ f(Z) \in B\}$, as required.

For the second claim, note that $\widehat{P}_m$ is permutation-symmetric, so the first claim implies that $\mathbb{P}(Z_i \in A \mid \widehat{P}_m)$ does not depend on $i$, almost surely. Averaging over the coordinates,

$$\mathbb{P}(Z_i \in A \mid \widehat{P}_m) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{P}(Z_j \in A \mid \widehat{P}_m) = \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{Z_j \in A\} \,\middle|\, \widehat{P}_m\right] = \widehat{P}_m(A),$$

which is the statement $Z_i \mid \widehat{P}_m \sim \widehat{P}_m$. □

This lemma is invoked repeatedly in the chapter: it underlies the rank argument in the proposition below, the exact-coverage treatment of ties, the PRDS proof for conformal p-values, the coverage-after-selection guarantee, and the analysis of online conformal p-values.

**Proposition 10.2: Finite-sample quantile correction**

If $Y_1, \dots, Y_{n+1}$ are exchangeable, then the quantile rule above satisfies

$$\mathbb{P}(Y_{n+1} \le c_n) \ge 1 - \alpha.$$

If there are no ties almost surely and $k \le n$, then

$$1 - \alpha \le \mathbb{P}(Y_{n+1} \le c_n) = \frac{k}{n+1} < 1 - \alpha + \frac{1}{n+1}.$$

**Proof of Proposition 10.2**

If $k = n + 1$, then $c_n = \infty$ and the claim is immediate. Suppose $k \le n$. Attach independent uniform tie breakers to the $n + 1$ observations and let $R_{n+1}$ be the resulting rank of $Y_{n+1}$. Exchangeability and symmetric tie breaking make $R_{n+1}$ uniform on $\{1, \dots, n + 1\}$, even when the $Y_i$'s tie.

With ties, the two events need not be equal; the correct containment is

$$\{R_{n+1} \le k\} \subseteq \{Y_{n+1} \le c_n\}.$$

Indeed, if the test observation occupies one of the first $k$ tie-broken positions, then the $k$th calibration value cannot be strictly smaller than the test value. Consequently,

$$\mathbb{P}(Y_{n+1} \le c_n) \ge \mathbb{P}(R_{n+1} \le k) = \frac{k}{n+1} \ge 1 - \alpha.$$

When there are no ties, the containment is an equality, so the coverage is exactly $k/(n+1)$, which also gives the stated upper bound. □

The same proof applies to any exchangeable scores. Conformal prediction is mostly a careful way of producing scores for the calibration points and the test point so that this rank proof remains valid.

## 3. Full Conformal Prediction

**Route: Core**

A score function assigns a numerical measure of disagreement or nonconformity [190]. For regression, a common choice is

$$S(x, y; \widehat{f}) = |y - \widehat{f}(x)|.$$

Smaller scores mean better conformity. To construct a full conformal prediction set, fix a test covariate $x$ and consider a possible response $y$. Add this candidate observation to the observed data to form the augmented sample

$$D_n^{x,y} = (Z_1, \dots, Z_n, (x, y)).$$

We hold $x$ fixed while examining different candidate responses $y$.

One common construction fits a predictor to this augmented sample. Denote that predictor by $\widehat{f}^{x,y}$: the superscript labels the candidate $(x, y)$ used for fitting and does not denote exponentiation. The expression $\widehat{f}^{x,y}(u)$ evaluates this fitted predictor at covariate value $u$. For absolute-residual scores, compute

$$R_i(x, y) = |Y_i - \widehat{f}^{x,y}(X_i)|, \qquad i = 1, \dots, n,$$

and

$$R_{n+1}(x, y) = |y - \widehat{f}^{x,y}(x)|.$$

The observed $X_i, Y_i$ stay fixed as $y$ varies, but the fitted predictor can change because the candidate is part of its training sample. Consequently, even the scores of the observed points can change. The arguments $(x, y)$ record this possible dependence; they do not assert that every score changes whenever either argument changes.

Example: predicting with the sample mean. Suppose the predictor ignores covariates and returns the mean of the augmented responses,

$$\widehat{\mu}(y) = \frac{\sum_{j=1}^{n} Y_j + y}{n+1}.$$

With observed responses $Y_1 = 0$ and $Y_2 = 2$, candidate $y = 0$ gives $\widehat{\mu}(0) = 2/3$ and absolute residuals $(2/3, 4/3)$ at the observed points. Candidate $y = 4$ gives $\widehat{\mu}(4) = 2$ and absolute residuals $(2, 0)$. The observed responses have not changed; their scores change through the fitted mean. This simple example ignores $x$.

Other score constructions are allowed. The general requirement is symmetry: relabeling the augmented observations must relabel their scores in the same way. Thus the fitting and scoring steps must treat the candidate observation by the same rule as the other observations. In the residual construction above, this holds when the fitted predictor is unchanged by reordering

its training observations. For any such symmetric construction, the conformal p-value for the candidate pair is

$$p(x,y) = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{R_i(x,y) \geq R_{n+1}(x,y)\}}{n+1},$$

where larger scores indicate worse conformity. The plus one counts the candidate itself, and the comparison includes ties. Retain the responses whose p-values exceed $\alpha$:

$$C_n(x) = \{y \in \mathcal{Y} : p(x,y) > \alpha\}.$$

At the true pair $(x,y) = (X_{n+1}, Y_{n+1})$, the augmented observations are the $n+1$ exchangeable data points, and their symmetrically computed scores are exchangeable as well. The rank argument makes $p(X_{n+1}, Y_{n+1})$ super-uniform, including when scores tie. Therefore

$$\mathbb{P}\{Y_{n+1} \notin C_n(X_{n+1})\} = \mathbb{P}\{p(X_{n+1}, Y_{n+1}) \leq \alpha\} \leq \alpha.$$

This probability averages over the observed data and the future pair. Holding $x$ fixed to describe the computation does not establish coverage conditional on every fixed $x$.

Full conformal is conceptually clean because all $n+1$ scores are produced symmetrically. It can also be expensive: in regression, the algorithm may need to be refit for many candidate values $y$. Split conformal gives up some statistical efficiency in exchange for a much simpler computation.

Why a one-fit shortcut fails. A tempting shortcut is to fit $\widehat{f}$ once on all $n$ training observations, calibrate with the in-sample residuals, and use the same fit at the test point. The resulting training and test residuals are generally not exchangeable because the training responses were used to construct $\widehat{f}$, whereas $Y_{n+1}$ was not. The failure can be complete. Suppose the $X_i$'s are i.i.d. from a continuous distribution, the $Y_i$'s are i.i.d. from a continuous distribution and independent of all the covariates, and $\widehat{f}$ is one-nearest-neighbor regression fitted on $Z_1, \ldots, Z_n$. At a covariate value $x$, this rule returns the response $Y_J$ attached to the closest training covariate $X_J$; any ties are broken by a fixed rule that depends only on the covariates. The training covariates are distinct almost surely, and at $x = X_i$ the closest training covariate is $X_i$ itself. Hence every in-sample residual is zero:

$$R_i^{\text{in}} = |Y_i - \widehat{f}(X_i)| = 0, \qquad i = 1, \ldots, n.$$

If $\alpha \geq 1/(n+1)$, then the required rank satisfies $\lceil (1-\alpha)(n+1) \rceil \leq n$. The selected order statistic is therefore one of the $n$ zero residuals, so the residual quantile is zero and the naive interval at $X_{n+1}$ is the singleton $\{\widehat{f}(X_{n+1})\}$. At the test covariate, let $J$ be the index selected by the nearest-neighbor rule. Then $\widehat{f}(X_{n+1}) = Y_J$. The index $J$ depends only on the covariates, while $Y_J$ and $Y_{n+1}$ are independent draws from a continuous distribution. Consequently,

$$\mathbb{P}\{Y_{n+1} = \widehat{f}(X_{n+1})\} = \mathbb{P}(Y_{n+1} = Y_J) = 0.$$

Thus the shortcut has zero coverage in this example. Split conformal avoids the problem by using one set of observations only to fit the predictor and a separate set for calibration. Conditional on the fitting set, the calibration scores and the test score are then exchangeable.

## 4. Split Conformal Prediction

**Route: Core**

Split conformal prediction, as treated for regression by Lei et al. [116], divides the observed data into a proper training set $D_1$ and a calibration set $D_2$, with

$$|D_1| = n_1, \qquad |D_2| = n_2, \qquad n_1 + n_2 = n.$$

Fit a predictor $\widehat{f}_{n_1}$ once, using only $D_1$, and keep it fixed as candidate responses are examined. For $i \in D_2$, compute calibration residuals

$$R_i = |Y_i - \widehat{f}_{n_1}(X_i)|.$$

These calibration scores remain fixed as the candidate $y$ varies; only the candidate score $|y - \widehat{f}_{n_1}(x)|$ changes. Let

$$k = \lceil (1-\alpha)(n_2+1) \rceil$$

and let

$$\widehat{c} = R_{(k)}$$

be the $k$th smallest calibration residual, with $\widehat{c} = \infty$ if $k = n_2 + 1$. The split conformal interval is

$$C_n(x) = [\widehat{f}_{n_1}(x) - \widehat{c},\ \widehat{f}_{n_1}(x) + \widehat{c}].$$

Figure 10.2 separates model fitting from calibration: the upper branch determines the center, while the lower branch determines the cutoff $\widehat{c}$. The finite-sample correction is the choice of rank $k$: the factor $n_2 + 1$ counts the calibration observations and one future observation. It does not change the residual values. For example, with $n_2 = 49$ and $\alpha = 0.1$, the calibration rank is $\lceil (1 - 0.1)(49 + 1) \rceil = \lceil 45 \rceil = 45$, giving no-ties coverage $45/50 = 0.9$.

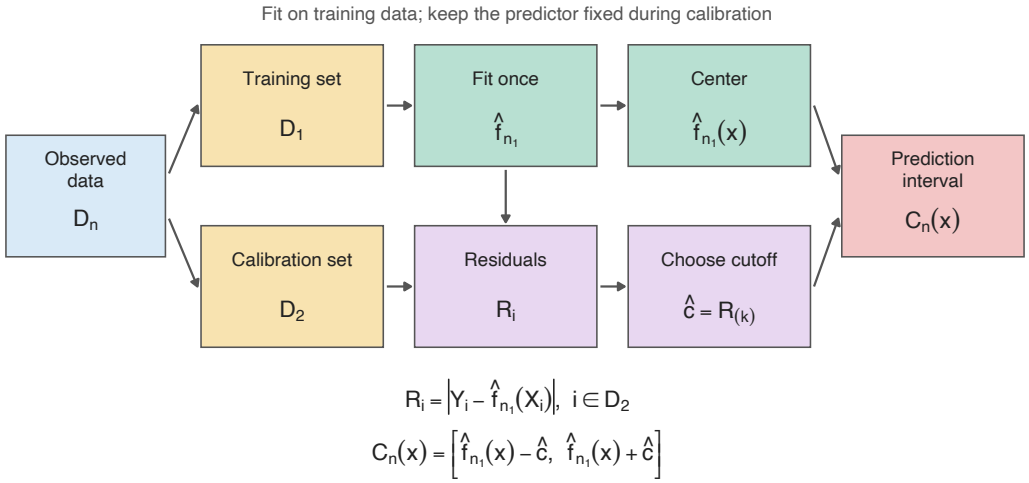



FIGURE 10.2. Split conformal prediction with absolute residuals. Split $D_n$ into a proper training set $D_1$ of size $n_1$ and a calibration set $D_2$ of size $n_2$. Fit $\widehat{f}_{n_1}$ on $D_1$; the same fitted predictor supplies the interval center and the residuals on $D_2$. "Choose cutoff" means take the $k$th smallest calibration residual, with $k = \lceil (1-\alpha)(n_2+1) \rceil$, setting $\widehat{c} = \infty$ if $k = n_2 + 1$. The interval combines this cutoff with the center at $x$. Coverage follows from exchangeability of the calibration and future test scores conditional on $D_1$.

**Theorem 10.3: Split conformal coverage**

Assume $Z_1, \ldots, Z_{n+1}$ are exchangeable and the split is fixed in advance or chosen independently of the data. Then the split conformal interval satisfies

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha.$$

If the calibration and test residuals have no ties almost surely, then the coverage is less than

$$1 - \alpha + \frac{1}{n_2 + 1}.$$

**Proof of Theorem 10.3**

Condition on the proper training set $D_1$. The fitted function $\widehat{f}_{n_1}$ is then fixed. The calibration observations $\{Z_i : i \in D_2\}$ and the test observation $Z_{n+1}$ remain exchangeable, so the scores

$$\{R_i : i \in D_2\}, \qquad R_{n+1} = |Y_{n+1} - \widehat{f}_{n_1}(X_{n+1})|$$

are exchangeable. The event $Y_{n+1} \in C_n(X_{n+1})$ is exactly $R_{n+1} \leq \widehat{c}$. Proposition 10.2 applied to these $n_2 + 1$ scores gives the conditional coverage statement. Averaging over $D_1$ gives marginal coverage. The no-ties upper bound follows from the same proposition. □

The proof also covers general nonconformity scores. Let

$$S(x, y) = S(x, y; \widehat{f}_{n_1})$$

be any score for which smaller values are more conforming. Compute

$$R_i = S(X_i, Y_i), \qquad i \in D_2,$$

and define

$$C_n(x) = \{y : S(x, y) \leq R_{(k)}\}.$$

The rank proof is unchanged. The art is to choose $S$ so that the resulting set is useful.

## 5. What conditional coverage means

**Route: Core**

To interpret a coverage statement, identify which data have been fixed and which observations are still random. The split-conformal proof establishes

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) \mid D_1\} \geq 1 - \alpha.$$

Here the predictor is fixed, while the calibration set $D_2$ and the test point are still random. Averaging over $D_1$ gives marginal coverage.

If we also fix $D_2$, the calibrated set is fixed as a function of the test covariate. Its coverage over a fresh test point can be below $1 - \alpha$ for some realized calibration samples. Under i.i.d. sampling, Appendix B describes how coverage varies across calibration samples. A different question fixes the test covariate $X_{n+1} = x$ and averages over the observed data and the response at that covariate. Ordinary conformal calibration does not generally give distribution-free coverage of at least $1 - \alpha$ at every such $x$. These statements condition on different information and should not be used interchangeably.

Section 1 of Appendix B gives the precise conditioning targets and the limits of distribution-free conditional coverage. It also examines weaker requirements, such as coverage within groups specified in advance.

## 6. Randomization and Ties

**Route: Core**

Without ties, split-conformal coverage lies on the grid $\{1/(n_2+1), \ldots, 1\}$; ties can make it more conservative. Auxiliary randomization can produce exact coverage.

Let $W$ have distribution function $F$, and let $U \sim \mathrm{Unif}(0,1)$ be independent of $W$. Write $F(w-) = \lim_{v \uparrow w} F(v) = \mathbb{P}(W < w)$ for the left limit. Thus $F(w) - F(w-) = \mathbb{P}(W = w)$ is the probability mass at $w$. Define the randomized distribution transform

$$F^*(W) = F(W-) + U\{F(W) - F(W-)\}.$$

Then

$$F^*(W) \sim \mathrm{Unif}(0,1).$$

When $F$ is continuous, $F^*(W) = F(W)$. At a value with positive probability mass, the uniform randomization spreads that mass across the jump.

For split conformal prediction, let $R_{n+1} = S(X_{n+1}, Y_{n+1})$, and let $F_{n_2+1}$ be the empirical distribution of the $n_2+1$ scores formed by the calibration scores and $R_{n+1}$. Draw a single $U \sim \mathrm{Unif}(0,1)$ independent of $D_1, D_2, (X_{n+1}, Y_{n+1})$, and use it once to randomize the test rank. To construct the prediction set, replace the unknown test score by $S(x,y)$ for each candidate response $y$, using the same draw $U$ for every candidate. This gives

$$C_n^*(x) = \Big\{ y : \frac{1}{n_2+1} \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) < S(x,y)\} + \frac{U}{n_2+1} \left( 1 + \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) = S(x,y)\} \right) \le 1 - \alpha \Big\}.$$

The membership statistic inside the display is a *smoothed rank*: the strict count places the test score below its tied block, and the single uniform draw $U$ spreads the tied block, which includes the test score itself, evenly across its width. The following lemma shows that this one draw makes the statistic exactly uniform, with no continuity assumption on the score distribution [5, 190].

**Lemma 10.4: Exact uniformity of the smoothed rank**

Let $S_1, \ldots, S_m$ be exchangeable real-valued random variables and let $U \sim \mathrm{Unif}(0,1)$ be independent of $(S_1, \ldots, S_m)$. Define

$$V = \frac{1}{m} \left[ \#\{i \le m-1 : S_i < S_m\} + U\big(1 + \#\{i \le m-1 : S_i = S_m\}\big) \right].$$

Then

$$\mathbb{P}(V \le \tau) = \tau \qquad \text{for every } \tau \in [0,1],$$

so $V \sim \mathrm{Unif}(0,1)$ exactly. Ties among the scores are allowed.

In words: a single uniform tie-break turns the normalized rank of the test score into an exact continuous uniform variable, not merely a variable supported on a discrete grid.

**Route: Optional proof**

**Proof of Lemma 10.4**

First dispose of $\tau = 0$. The bracketed term is at least $U$, and $U > 0$ almost surely, so $\mathbb{P}(V \le 0) = 0$. Fix $\tau \in (0, 1]$.
Let $S_{(1)} \le \cdots \le S_{(m)}$ be the order statistics of $S_1, \ldots, S_m$ and set

$$c_\tau = S_{(\lceil m\tau \rceil)}, \qquad K = \#\{i \le m : S_i < c_\tau\}, \qquad L = \#\{i \le m : S_i = c_\tau\}.$$

Since $c_\tau$ is one of the scores, $L \ge 1$. By the definition of the $\lceil m\tau \rceil$th order statistic, at least $\lceil m\tau \rceil$ scores are $\le c_\tau$ and at most $\lceil m\tau \rceil - 1$ scores are strictly below $c_\tau$; hence

$$K < m\tau \le K + L.$$

Next, locate $S_m$ relative to $c_\tau$. There are three cases.
If $S_m < c_\tau$, then every score $\le S_m$ is strictly below $c_\tau$, so

$$\#\{i \le m-1 : S_i < S_m\} + (1 + \#\{i \le m-1 : S_i = S_m\}) = \#\{i \le m : S_i \le S_m\} \le K.$$

Because $U \le 1$, the bracketed term in $V$ is at most this sum, so $V \le K/m < \tau$. Thus $V \le \tau$ holds with conditional probability one on this case.
If $S_m > c_\tau$, then all $K + L$ scores that are $\le c_\tau$ have index at most $m - 1$ and are strictly below $S_m$, so the strict count alone is at least $K + L \ge m\tau$, and the randomized term adds at least $U/m > 0$ almost surely. Thus $V > \tau$ almost surely on this case.
If $S_m = c_\tau$, then $\#\{i \le m-1 : S_i < S_m\} = K$ and $1 + \#\{i \le m-1 : S_i = S_m\} = L$, so $V = (K + UL)/m$ and

$$V \le \tau \iff U \le \frac{m\tau - K}{L}.$$

The quantile inequalities give $0 < m\tau - K \le L$, so the ratio on the right lies in $(0, 1]$, and the event has conditional probability exactly $(m\tau - K)/L$ because $U$ is uniform and independent of the scores.
Now condition on the multiset of score values, meaning the collection of values with repetitions retained and labels removed, using Lemma 10.1. Given the multiset, the quantities $c_\tau$, $K$, and $L$ are fixed, the score $S_m$ is equally likely to occupy each of the $m$ positions in the ordered list, and $U$ remains uniform and independent. Hence

$$\mathbb{P}(S_m < c_\tau \mid \text{multiset}) = \frac{K}{m}, \qquad \mathbb{P}(S_m = c_\tau \mid \text{multiset}) = \frac{L}{m},$$
$$\mathbb{P}(S_m > c_\tau \mid \text{multiset}) = \frac{m - K - L}{m}.$$

Combining the three cases,

$$\mathbb{P}(V \le \tau \mid \text{multiset}) = \frac{K}{m} \cdot 1 + \frac{L}{m} \cdot \frac{m\tau - K}{L} + \frac{m - K - L}{m} \cdot 0 = \frac{K}{m} + \frac{m\tau - K}{m} = \tau.$$

The right side does not depend on the multiset, so averaging over the multiset gives $\mathbb{P}(V \le \tau) = \tau$. □

Apply the lemma with $m = n_2 + 1$, taking the calibration scores $\{S(X_i, Y_i) : i \in D_2\}$ as $S_1, \ldots, S_{m-1}$ and the test score $S_m = S(X_{n+1}, Y_{n+1})$. Conditional on $D_1$, these $m$ scores are exchangeable, $U$ is independent of them, and the coverage event $\{Y_{n+1} \in C_n^*(X_{n+1})\}$ is exactly $\{V \le 1 - \alpha\}$. Therefore

$$\mathbb{P}\{Y_{n+1} \in C_n^*(X_{n+1}) \mid D_1\} = 1 - \alpha$$

exactly, with ties allowed. This exactness is useful with discrete scores, classification labels, and other settings where ties are common.

## 7. Conformal P-Values

**Route: Core**

Conformal scores can also be used for testing. In this section, $n$ denotes the number of reference observations; any data used to train the score form a separate sample.

A brief comment on sign conventions before we start. In the prediction-set sections above, the score $S(x, y) = |y - \widehat{f}(x)|$ is a *nonconformity* score: *smaller* values indicate better agreement with the fitted model. In the conformal p-value literature, it is also common to work with a *conformity* score, where *larger* values indicate better conformity. The two conventions differ only by a sign, and either is fine as long as the comparison in the p-value formula is reversed accordingly. We use the conformity convention in this section because the "small p-value means unusual" interpretation is then immediate.

Let the reference sample be

$$Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} P.$$

Fix a conformity score $S$, where larger values mean better conformity. For this fixed-score construction, a learned score must be trained on data independent of both the reference sample and the test points; in that case, condition on those training data throughout the argument so that $S$ is fixed. Fitting the score on the reference observations alone and treating a new point differently would recreate the in-sample calibration problem discussed above. For a new point $z$, define

$$p(z) = \frac{1 + \#\{1 \le i \le n : S(Z_i) \le S(z)\}}{n + 1}.$$

Small p-values indicate that $z$ conforms poorly relative to the reference sample. If $Z_{n+1} \sim P$ is independent of the reference sample and $S(Z)$ is continuous under $P$, then

$$p(Z_{n+1}) \in \left\{ \frac{1}{n+1}, \dots, 1 \right\}$$

is exactly uniform on this grid.

The conformal p-value is a randomization p-value. Recall the permutation test. Given a test statistic $T(z_1, \dots, z_{n+1})$, chosen so that large values are evidence against the null hypothesis that the $n + 1$ observations are exchangeable, the permutation-test p-value compares the observed statistic with its values under all relabelings of the data:

$$p_{\text{perm}} = \frac{1}{(n+1)!} \sum_{\sigma} \mathbf{1}\{T(Z_{\sigma(1)}, \dots, Z_{\sigma(n+1)}) \ge T(Z_1, \dots, Z_{n+1})\},$$

where the sum runs over all $(n + 1)!$ permutations $\sigma$ of $\{1, \dots, n + 1\}$. The identity permutation is one of the terms, so $p_{\text{perm}} \ge 1/(n + 1)!$, and under exchangeability $p_{\text{perm}}$ is super-uniform: $\mathbb{P}(p_{\text{perm}} \le t) \le t$ for every $t \in [0, 1]$ [5].

**Proposition 10.5: Conformal p-values are permutation-test p-values**

Set $Z_{n+1} = z$ and take the test statistic

$$T(z_1, \dots, z_{n+1}) = -S(z_{n+1}),$$

which is large when the observation placed in position $n + 1$ conforms poorly. Then

$$p_{\text{perm}} = p(z).$$

This numerical identity holds for every observed $z$. When the random test observation is exchangeable with the reference observations, the resulting p-value is super-uniform, with or without ties.

**Proof of Proposition 10.5**

For this statistic, the indicator inside the sum is $\mathbf{1}\{-S(Z_{\sigma(n+1)}) \geq -S(Z_{n+1})\} = \mathbf{1}\{S(Z_{\sigma(n+1)}) \leq S(Z_{n+1})\}$, which depends on $\sigma$ only through the index $i = \sigma(n+1)$ sent to position $n+1$. Each $i \in \{1, \ldots, n+1\}$ arises from exactly $n!$ permutations. Grouping the $(n+1)!$ permutations by $i$,

$$p_{\text{perm}} = \frac{1}{(n+1)!} \sum_{i=1}^{n+1} n!\, \mathbf{1}\{S(Z_i) \leq S(Z_{n+1})\} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{S(Z_i) \leq S(Z_{n+1})\}.$$

The term $i = n+1$ contributes 1, and the remaining $n$ terms count the reference scores that are at most $S(z)$. Hence $p_{\text{perm}} = \{1 + \#\{1 \leq i \leq n : S(Z_i) \leq S(z)\}\}/(n+1) = p(z)$. □

This identification connects conformal p-values to a recurring idea in the book: compare an observed statistic with values obtained by transformations or resampling that preserve the null distribution. The CRT of Chapter 9 has the same structure, with independently resampled copies of a covariate playing the role of the permuted data. The shared proof step is the super-uniform rank of the observed statistic within an exchangeable collection. The assumptions that create this collection differ: a CRT requires the correct conditional law of the resampled covariate, whereas the conformal construction here uses exchangeability of the reference and null test observations. Neither method requires the prediction model for the response to be correct [5, 190]. When the group is too large to enumerate, the permutation sum is approximated using independent uniformly sampled permutations; counting the identity permutation in both numerator and denominator, the same plus-one correction that appears in the conformal p-value, keeps the Monte Carlo p-value exactly super-uniform [144].

For the absolute-residual score

$$S(x, y) = -|y - \widehat{\mu}(x)|,$$

where $\widehat{\mu}$ is trained on data independent of the reference sample, write $R_i = |Y_i - \widehat{\mu}(X_i)|$ for the reference residuals. Then the inequality $S(Z_i) \leq S(z)$ is $-|Y_i - \widehat{\mu}(X_i)| \leq -|y - \widehat{\mu}(x)|$, which flips under the negation to $|y - \widehat{\mu}(x)| \leq |Y_i - \widehat{\mu}(X_i)|$. The p-value therefore becomes

$$p(x, y) = \frac{1 + \#\{1 \leq i \leq n : |y - \widehat{\mu}(x)| \leq |Y_i - \widehat{\mu}(X_i)|\}}{n+1}.$$

The split-conformal set is exactly the inversion of this p-value:

$$y \notin C_n(x) \quad \Longleftrightarrow \quad p(x, y) \leq \alpha.$$

To check the finite-grid convention, let $k = \lceil (1-\alpha)(n+1) \rceil$. For $k \leq n$, write $R_{(k)}$ for the $k$th smallest of the reference absolute residuals $R_1, \ldots, R_n$, counting repeated values separately. If the candidate absolute residual $|y - \widehat{\mu}(x)|$ exceeds $R_{(k)}$, at most $n - k$ reference residuals are at least as large, so $p(x, y) \leq (n + 1 - k)/(n+1) \leq \alpha$. If the candidate absolute residual is at most $R_{(k)}$, at least $n - k + 1$ reference residuals are at least as large, so $p(x, y) \geq (n + 2 - k)/(n+1) > \alpha$. These implications allow ties. When $k = n+1$, the cutoff is infinite and $p(x, y) \geq 1/(n+1) > \alpha$, so every candidate is included.

Now suppose we have $m$ new test points

$$Z_{n+1}, \dots, Z_{n+m},$$

which are mutually independent and independent of the reference sample. Define

$$p_j = p(Z_{n+j}), \qquad j = 1, \dots, m.$$

These p-values can be used for *outlier detection*: flagging observations that appear inconsistent with the reference distribution $P$. For test point $j$, the null hypothesis is $H_{0,j} : Z_{n+j} \sim P$; a small $p_j$ provides evidence against this hypothesis. A false discovery is a flagged point whose null hypothesis is true. FDR measures the expected fraction of such false calls among all flagged points, with this fraction defined as zero when no points are flagged.

The p-values are dependent because they share the same reference sample. Under continuous scores, however, they have a positive dependence structure that is compatible with BH; this PRDS property for conformal p-values is due to Bates et al. [18].

**Simultaneous coverage for many predictions.** Return here to the nonconformity scores used for split prediction, where smaller values indicate better agreement. Assume the calibration observations and the future observations are independent draws from the same distribution, independent of the proper training set $D_1$. Before using conformal p-values for FDR, consider a simpler question. Suppose the same split-conformal threshold $\widehat{c}$ is used to form prediction sets for $m$ independent future test points. Does sharing the calibration set make simultaneous coverage much easier? In general, no.

Condition on the proper training set $D_1$. Let $R'_1, \dots, R'_m$ be the future test scores, independent of the calibration scores, and let $F = F_{D_1}$ be their common distribution function. Here $n_2$ again denotes the number of calibration observations. Conditional on $\widehat{c}$ and $D_1$,

$$\mathbb{P}(R'_1 \le \widehat{c}, \dots, R'_m \le \widehat{c} \mid \widehat{c}, D_1) = F(\widehat{c})^m.$$

Therefore, conditional on $D_1$,

$$\mathbb{P}(R'_1 \le \widehat{c}, \dots, R'_m \le \widehat{c} \mid D_1) = \mathbb{E}[F(\widehat{c})^m \mid D_1] \ge \{\mathbb{E}[F(\widehat{c}) \mid D_1]\}^m \ge (1-\alpha)^m,$$

where the middle step is Jensen's inequality. The lower bound is asymptotically sharp. If $F$ is continuous and strictly increasing at its $(1-\alpha)$ quantile, then, for fixed $m$, the joint coverage of the $m$ prediction sets converges to $(1-\alpha)^m$ as $n_2 \to \infty$ [5]. The reason is that $\widehat{c}$ converges almost surely to the population $(1-\alpha)$ quantile of $F$, so $F(\widehat{c})^m \to (1-\alpha)^m$ almost surely, and dominated convergence (the integrand is bounded by one) gives $\mathbb{E}[F(\widehat{c})^m \mid D_1] \to (1-\alpha)^m$.

Choose $\alpha_{\mathrm{FWER}} \in (0,1)$ as the desired upper bound on the probability that at least one of the $m$ prediction sets misses its corresponding future response. This is a familywise error target; $\alpha$ remains the miscoverage level used to construct each individual set. Thus the familywise requirement is

$$\mathbb{P}\{\text{all } m \text{ prediction sets cover}\} \ge 1 - \alpha_{\mathrm{FWER}}$$

and is ensured by the choice

$$\alpha \le 1 - (1 - \alpha_{\mathrm{FWER}})^{1/m} \approx \frac{\alpha_{\mathrm{FWER}}}{m}.$$

The approximation applies when $\alpha_{\mathrm{FWER}}$ is small and recovers the familiar Bonferroni scale. The sharp limiting example shows why a substantially larger per-point level cannot give this guarantee uniformly over the stated sampling models. The lesson is the same as in multiple testing: if the goal is to avoid even one error, the price grows with $m$. If the goal is to control the fraction of false discoveries among many outlier calls, FDR and the PRDS theorem below are often the more natural tools [70, 185].

**Theorem 10.6: PRDS property of conformal p-values**

Let $S$ be a fixed conformity score and let $Z_1, \ldots, Z_n$ be independent draws from $P$. Assume $S(Z)$ is continuous for $Z \sim P$, and assume that $Z_{n+1}, \ldots, Z_{n+m}$ are mutually independent and independent of the reference sample. For test points whose null hypotheses $H_{0,j} : Z_{n+j} \sim P$ are true, the conformal p-values are PRDS on the true-null subset.

The same result applies to a score trained on independent data, by conditioning on those data and requiring the score distribution to be continuous almost surely under that conditioning. The appendix proof explains why the conditional argument also gives unconditional PRDS.

The monotone-rank proof of this theorem is given in Appendix B.

This theorem is the bridge back to Chapter 6. Although the p-values share a reference sample, each true-null p-value is super-uniform: its uniform grid law satisfies $\mathbb{P}(p_i \le t) = \lfloor (n+1)t \rfloor / (n+1) \le t$—and the theorem above establishes PRDS. Theorem 6.5 therefore applies directly: using BH at FDR target $q$ to flag outliers gives FDR $\le m_0 q/m \le q$, where $m_0$ is the number of true nulls among the $m$ test points. We reserve $\alpha$ in this chapter for prediction miscoverage and use $q$ for an FDR target. The grid distribution can also be smoothed. For one p-value, augmenting its rank with an independent draw $U \sim \mathrm{Unif}(0,1)$, exactly as in the randomized prediction set of Lemma 10.4, produces a *smoothed* conformal p-value whose null distribution is exactly $\mathrm{Unif}(0,1)$ rather than uniform on the grid $\{1/(n+1), \ldots, 1\}$, and it removes the discreteness slack $\lfloor (n+1)t \rfloor / (n+1) \le t$ quantified above. The price is an external source of randomness: two analysts holding the same data and the same score function need not report the same smoothed p-value, because the draw of $U$ is not part of the data. For several p-values, use fresh independent draws unless a joint randomization scheme has been analyzed; the marginal smoothing statement alone does not establish a new joint PRDS result. The discrete PRDS theorem above already suffices for BH.

## 8. Conformalized Quantile Regression

**Route: Core**

Split conformal intervals based on absolute residuals have constant half-width $\widehat{c}$. They are valid, but they do not adapt to heteroscedasticity. If the noise variance is small for some covariates and large for others, a constant-width interval can be too wide in easy regions and too narrow in hard regions. Conformalized quantile regression (CQR) uses fitted conditional quantiles to adapt the width to the covariates, followed by conformal calibration.

Quantile regression estimates conditional quantiles. For a conditional distribution

$$F(y \mid X = x) = \mathbb{P}(Y \le y \mid X = x),$$

the $\tau$th conditional quantile is

$$\xi_\tau(x) = \inf\{y : F(y \mid X = x) \ge \tau\}.$$

Because these are unknown population quantities rather than fitted estimates, the interval they define is called the oracle interval:

$$C^*(x) = [\xi_{\alpha/2}(x), \xi_{1-\alpha/2}(x)].$$

To verify its conditional coverage, write $F(\xi- \mid X = x) = \lim_{t \uparrow \xi} F(t \mid X = x)$ for the left limit of the conditional distribution function. The quantile definition gives

$$F(\xi_{\alpha/2}(x)- \mid X = x) \le \frac{\alpha}{2}, \qquad F(\xi_{1-\alpha/2}(x) \mid X = x) \ge 1 - \frac{\alpha}{2}.$$

The two probabilities of falling outside the interval are therefore

$$\mathbb{P}\{Y < \xi_{\alpha/2}(x) \mid X = x\} = F(\xi_{\alpha/2}(x)- \mid X = x) \le \frac{\alpha}{2},$$
$$\mathbb{P}\{Y > \xi_{1-\alpha/2}(x) \mid X = x\} = 1 - F(\xi_{1-\alpha/2}(x) \mid X = x) \le \frac{\alpha}{2}.$$

A union bound now gives

$$\begin{aligned}\mathbb{P}\{Y \notin C^*(x) \mid X = x\} &\le \mathbb{P}\{Y < \xi_{\alpha/2}(x) \mid X = x\} + \mathbb{P}\{Y > \xi_{1-\alpha/2}(x) \mid X = x\} \\ &\le \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.\end{aligned}$$

The oracle interval is unknown in practice, and estimated quantiles need not inherit this guarantee. CQR uses conformal calibration to recover finite-sample marginal validity for the estimated interval.

For a quantile level $\tau \in (0,1)$, let $b$ be a proposed value of the conditional quantile. The pinball loss measures its error at response $y$, with different penalties for underprediction and overprediction:

$$\rho_\tau(y,b) = \begin{cases}\tau(y-b), & y > b, \\ (1-\tau)(b-y), & y \le b.\end{cases}$$

Classical quantile regression estimates $\xi_\tau(x)$ by minimizing an average pinball loss, possibly with regularization.

CQR of Romano et al. [154] wraps estimated lower and upper quantiles with conformal calibration. Split the data into $D_1$ and $D_2$, with the split fixed in advance or chosen independently of the data. On $D_1$, fit

$$\widehat{\xi}_{\alpha/2}(x), \qquad \widehat{\xi}_{1-\alpha/2}(x).$$

We assume throughout that the fits are noncrossing, in the sense that $\widehat{\xi}_{\alpha/2}(x) \le \widehat{\xi}_{1-\alpha/2}(x)$ for every $x$; a quantile-regression fitter that does not guarantee this should be post-processed (for example, by sorting or by isotonic adjustment) before calibration. Define the score for a candidate pair $(x,y)$ by

$$A(x,y) = \max\{\widehat{\xi}_{\alpha/2}(x) - y,\ y - \widehat{\xi}_{1-\alpha/2}(x)\}.$$

The calibration scores are $A_i = A(X_i, Y_i)$, $i \in D_2$. The score is positive when $Y_i$ falls outside the fitted interval and negative when it lies inside with slack. Let

$$\widehat{c}_{\mathrm{CQR}} = \text{the } \lceil (1-\alpha)(n_2+1) \rceil \text{th smallest value of } \{A_i : i \in D_2\}.$$

As in ordinary split conformal prediction, set $\widehat{c}_{\mathrm{CQR}} = +\infty$ when $\lceil (1-\alpha)(n_2+1) \rceil = n_2 + 1$. The CQR prediction set is

$$\widetilde{C}_n(x) = \{y : A(x,y) \le \widehat{c}_{\mathrm{CQR}}\} = [\widehat{\xi}_{\alpha/2}(x) - \widehat{c}_{\mathrm{CQR}},\ \widehat{\xi}_{1-\alpha/2}(x) + \widehat{c}_{\mathrm{CQR}}],$$

with the convention that the interval is empty if the lower endpoint exceeds the upper endpoint. When the interval is nonempty, its calibrated width is the uncalibrated quantile width plus $2\widehat{c}_{\mathrm{CQR}}$:

$$\{\widehat{\xi}_{1-\alpha/2}(x) - \widehat{\xi}_{\alpha/2}(x)\} + 2\widehat{c}_{\mathrm{CQR}},$$

so calibration expands or contracts the fitted quantile envelope by $2\widehat{c}_{\mathrm{CQR}}$. Under the noncrossing condition above, the interval is nonempty automatically when $\widehat{c}_{\mathrm{CQR}} \ge 0$. Since CQR scores can be negative (a fitted interval already covers $Y_i$ with slack), $\widehat{c}_{\mathrm{CQR}}$ can be negative. A negative correction indicates that the fitted intervals are wider than needed on the calibration sample; very negative corrections should be interpreted cautiously because they may produce empty intervals for some $x$.

Figure 10.3 contrasts what is calibrated. Split conformal adds one constant cutoff to a fitted center, whereas CQR adds the single correction $\widehat{c}_{\mathrm{CQR}}$ to fitted lower and upper conditional-quantile curves. The plotted simulation uses $n_2 = 180$ and $\alpha = 0.1$, so the correction is the $\lceil 0.9(180+1) \rceil = \lceil 162.9 \rceil = 163$rd ordered calibration score.

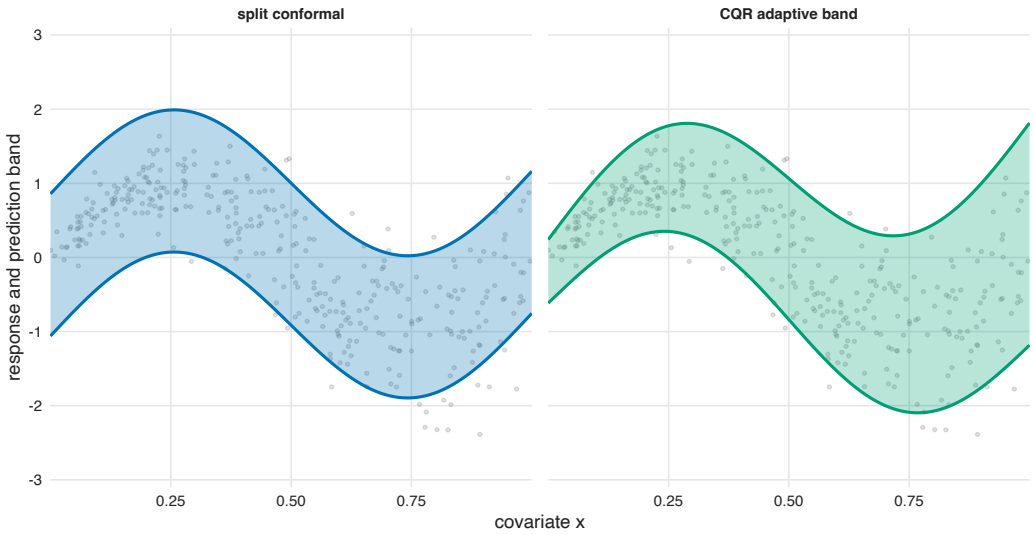



FIGURE 10.3. Split conformal and CQR adaptive bands in a heteroscedastic regression simulation with 240 training observations, 180 calibration observations, 350 independently plotted observations, and $\alpha = 0.1$. Split conformal uses a constant residual quantile; CQR calibrates a covariate-dependent lower and upper band with correction $\widehat{c}_{\mathrm{CQR}}$.

**Theorem 10.7: CQR coverage**

If $Z_1, \ldots, Z_{n+1}$ are exchangeable and the quantile models are fitted using only $D_1$, with the split fixed in advance or independent of the data, then

$$\mathbb{P}\{Y_{n+1} \in \widetilde{C}_n(X_{n+1})\} \geq 1 - \alpha.$$

If the calibration and test scores $A_i$ have no ties almost surely, then the coverage is less than $1 - \alpha + 1/(n_2 + 1)$.

**Proof of Theorem 10.7**

Conditional on $D_1$, the fitted quantile functions are fixed. The calibration scores $\{A_i : i \in D_2\}$ and the test score

$$A_{n+1} = \max\{\widehat{\xi}_{\alpha/2}(X_{n+1}) - Y_{n+1},\ Y_{n+1} - \widehat{\xi}_{1-\alpha/2}(X_{n+1})\}$$

are exchangeable. The event $Y_{n+1} \in \widetilde{C}_n(X_{n+1})$ is equivalent to $A_{n+1} \leq \widehat{c}_{\mathrm{CQR}}$. The rank argument gives the result. □

The conformal calibration is what gives finite-sample marginal validity. The quality of the quantile regression determines how short and locally adaptive the interval is.

## 9. Jackknife and Jackknife+

**Route: Advanced**

Sample splitting is computationally simple, but it may waste data. If $D_1$ is small, the predictor may be poor. If $D_2$ is small, the calibration quantile may be unstable. Leave-one-out methods try to use the data more efficiently.

For $i = 1, \dots, n$, let $\widehat{f}_{-i}$ be the predictor trained on the $n-1$ observations in $D_n$ other than $(X_i, Y_i)$. Define the leave-one-out residual

$$R_i^{\mathrm{LOO}} = |Y_i - \widehat{f}_{-i}(X_i)|.$$

Let $\widehat{c}_{\mathrm{LOO}}$ be the $\lceil (1-\alpha)(n+1) \rceil$-th smallest leave-one-out residual, with $\widehat{c}_{\mathrm{LOO}} = +\infty$ if this rank is $n+1$. This specifies the finite-sample quantile convention. The ordinary jackknife interval uses this cutoff and the predictor $\widehat{f}$ fitted to all of $D_n$:

$$[\widehat{f}(X_{n+1}) - \widehat{c}_{\mathrm{LOO}},\ \widehat{f}(X_{n+1}) + \widehat{c}_{\mathrm{LOO}}].$$

Here stability means that leaving out one observation changes the predictions only slightly. The interval can work well for stable algorithms, but it does not have a universal finite-sample coverage guarantee. The problem is asymmetry: the residuals are computed using leave-one-out fits, while the interval is centered at the full-data fit. For unstable algorithms, the mismatch can be severe.

Jackknife+ [12] changes the endpoints. For this two-sided construction take $0 < \alpha < 1/2$. At the test covariate $X_{n+1}$, form

$$L_i = \widehat{f}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}, \qquad U_i = \widehat{f}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}.$$

Let

$$L_{\mathrm{JK+}} = \text{the } \lfloor \alpha(n+1) \rfloor \text{th smallest value of } L_1, \dots, L_n,$$

and

$$U_{\mathrm{JK+}} = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{th smallest value of } U_1, \dots, U_n,$$

with two boundary conventions when the rank falls outside $\{1, \dots, n\}$: set $L_{\mathrm{JK+}} = -\infty$ when $\lfloor \alpha(n+1) \rfloor = 0$, and $U_{\mathrm{JK+}} = +\infty$ when $\lceil (1-\alpha)(n+1) \rceil = n+1$ (the latter occurs when $\alpha < 1/(n+1)$). The jackknife+ interval is

$$[L_{\mathrm{JK+}}, U_{\mathrm{JK+}}],$$

which is the whole line $\mathbb{R}$ when $\alpha < 1/(n+1)$. The two infinite endpoints occur together: both rank conditions are equivalent to $\alpha(n+1) < 1$. Otherwise both endpoints are finite.

Figure 10.4 shows the two endpoint collections before their final order statistics are taken. With $n = 28$ and $\alpha = 0.2$, the lower endpoint uses rank $\lfloor 0.2(28+1) \rfloor = \lfloor 5.8 \rfloor = 5$, while the upper endpoint uses rank $\lceil 0.8(28+1) \rceil = \lceil 23.2 \rceil = 24$. These become $L_{\mathrm{JK+}}$ and $U_{\mathrm{JK+}}$, respectively.

**Theorem 10.8: Jackknife+ coverage**

Under exchangeability and a fitting rule unchanged by reordering its training observations, the jackknife+ interval satisfies

$$\mathbb{P}\{Y_{n+1} \in [L_{\mathrm{JK+}}, U_{\mathrm{JK+}}]\} \geq 1 - 2\alpha.$$

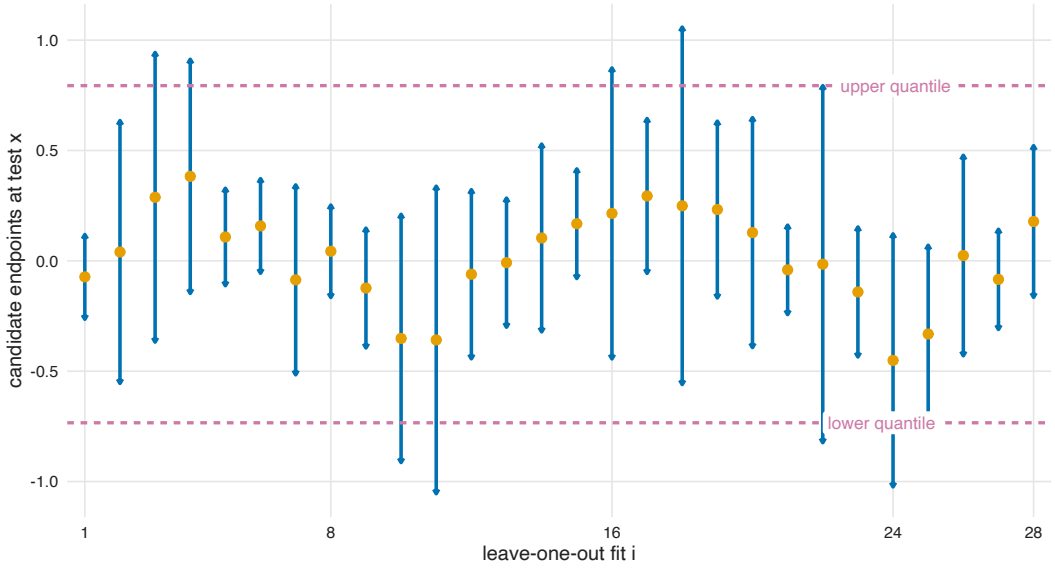



FIGURE 10.4. Jackknife+ endpoints. Each leave-one-out fit gives a lower and upper candidate endpoint at the test covariate. The final interval is formed from quantiles of these endpoint collections and is denoted $[L_{\mathrm{JK+}}, U_{\mathrm{JK+}}]$. The display uses $n = 28$ and $\alpha = 0.2$ so the endpoint quantiles are visibly interior.

Interpreting the coverage level. To obtain a guaranteed coverage level $1-\beta$ from this theorem, use $\alpha = \beta/2$ in the endpoint ranks. The proof compares prediction errors after leaving out pairs of observations and bounds how many observations can fail these comparisons. This counting argument, presented as a tournament in Appendix B, gives the factor two without requiring stability of the fitted predictor.

**Aggregating prediction sets.** Ensembles are common for point prediction, and the same question arises for prediction sets. Suppose $C_1(x), \ldots, C_K(x)$ are prediction sets, each with marginal coverage at least $1-\alpha$ for the same future pair $(X_{n+1}, Y_{n+1})$. No independence is assumed between the sets. Define the majority-vote set

$$C_{\mathrm{mv}}(x) = \left\{ y : \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{y \in C_k(x)\} > \frac{1}{2} \right\}.$$

Then

$$\mathbb{P}\{Y_{n+1} \in C_{\mathrm{mv}}(X_{n+1})\} \geq 1 - 2\alpha.$$

Indeed, if the majority-vote set misses $Y_{n+1}$, then at least half of the base sets miss it. Markov's inequality applied to the average miscoverage indicator gives the bound. The factor $2\alpha$ is the same kind of robust but conservative price that appears in jackknife+.

An additional calibration sample can be used to choose the voting threshold. For this step, first fix the base prediction rules independently of that sample. The calibration observations and the future observation must then be exchangeable, conditional on the data used to build the rules. Lowering the required vote fraction enlarges the set, giving a nested family: every smaller set is contained in every larger one. Under the remaining conditions of Theorem B.15 in Appendix B, its calibration rule gives a direct $1-\alpha$ marginal coverage guarantee. The majority-vote guarantee and calibration after aggregation are useful when several prediction sets are available but their dependence is hard to model [43, 68].

## 10. Conformal e-prediction

**Route: Research frontier**

A different generalization inverts e-value tests rather than p-value tests [69]. For a test covariate $x$ and candidate response $y \in \mathcal{Y}$, let $E_n(x, y) \geq 0$ be a statistic computed from the calibration data and the candidate pair $(x, y)$. The required validity condition is

$$\mathbb{E}\big[E_n(X_{n+1}, Y_{n+1})\big] \leq 1.$$

In words, the statistic evaluated at the true test pair must be an e-value. How to construct an informative statistic with this property is a separate problem; nonnegativity alone is not enough.

**Proposition 10.9: Fixed-level and post-hoc e-prediction**

Assume the displayed e-value condition and, for $a \in (0, 1)$, define

$$C_a(x) = \{y \in \mathcal{Y} : E_n(x, y) < 1/a\}.$$

Then every fixed $a$ gives marginal coverage

$$\mathbb{P}\{Y_{n+1} \in C_a(X_{n+1})\} \geq 1 - a.$$

Moreover, for every possibly data-dependent $\widehat{a} \in (0, 1)$,

$$\mathbb{E}\left[\frac{\mathbf{1}\{Y_{n+1} \notin C_{\widehat{a}}(X_{n+1})\}}{\widehat{a}}\right] \leq 1.$$

**Proof of Proposition 10.9**

For fixed $a$, the miscoverage event is $\{E_n(X_{n+1}, Y_{n+1}) \geq 1/a\}$, so Markov's inequality gives probability at most $a$. For the second statement, the pointwise inequality

$$\frac{\mathbf{1}\{E_n(X_{n+1}, Y_{n+1}) \geq 1/\widehat{a}\}}{\widehat{a}} \leq E_n(X_{n+1}, Y_{n+1})$$

holds: whenever the indicator is one, its defining inequality gives $1/\widehat{a} \leq E_n(X_{n+1}, Y_{n+1})$, and otherwise its left side is zero. Taking expectations proves the claim; this is exactly the post-hoc validity argument of Proposition 8.13. □

The second display is a budget-form guarantee, not ordinary marginal coverage at the random level $1 - \widehat{a}$. Thus a practitioner may choose $\widehat{a}$ to target a desired set size only if the resulting guarantee is reported in this budget form. Likewise, anytime validity requires a genuine e-process and Ville's inequality; a sequence obtained by simply recalculating e-values as calibration data arrive need not be an e-process. Gauthier et al. [69] give specific constructions for batch anytime-valid prediction and fixed-size sets with data-dependent coverage under additional assumptions. We do not reproduce those constructions here.

## 11. Assumptions in Plain Language

**Route: Core**

Exchangeability makes the rank argument work. Full conformal prediction preserves it by treating the augmented observations symmetrically during fitting and scoring. Split conformal prediction fits the predictor and chooses the score using the proper training set, then holds them fixed while calibrating. The fitted model may be misspecified; this can affect set size and how well the sets adapt to different covariates, while the rank-based coverage guarantee still holds under the stated assumptions.

Marginal coverage averages over the observed data and the next test point. Split conformal also gives coverage conditional on the proper training set, with the calibration data and test point still random. Neither statement provides coverage at every fixed covariate value or after fixing the realized calibration set. These stronger targets require separate analysis, as explained in Section 5 and Appendix B.

The treatment of ties is part of the procedure. Keeping ties in the tail count gives conservative p-values; the randomization described in this chapter makes ranks exactly uniform. Randomization does not repair a lack of exchangeability. Under exchangeability and a fitting rule unchanged by reordering observations, jackknife+ guarantees coverage of at least $1 - 2\alpha$ without a stability assumption. Majority voting has the same bound when every base set covers the same future response with probability at least $1 - \alpha$; the base sets need not be independent.

In the outlier-detection application of Section 7, each true-null test point must follow the reference law. With a fixed or independently trained score, a shared reference sample gives the dependence needed for BH under the independent-test and continuous-score assumptions of Theorem 10.6. Marginally valid p-values alone do not justify applying BH to an arbitrary dependent collection.

For e-prediction, the statistic at the true test pair must have expectation at most one. This condition gives coverage at each fixed reporting level. Choosing the level after seeing the data instead gives the expected weighted miscoverage bound in Proposition 10.9, which must be reported as such. Appendix B treats further questions about sequential observations, distribution changes, and losses beyond miscoverage.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | For split conformal, choose the predictor, score, and tuning choices using only the proper training data. For full conformal, check that fitting and scoring treat the candidate and observed points symmetrically. Plot calibration scores by collection time and by prespecified group, and record ties. |
| **What these data cannot show** | The calibration data cannot establish that future cases are exchangeable with the calibration cases. Marginal conformal coverage also does not imply coverage at every covariate value or within every subgroup. |
| **Sensitivity checks** | On a separate labeled evaluation set, report empirical coverage together with average interval length or set size, overall, over time, and in prespecified groups. |

| | |
|---|---|
| **Safer fallback** | If recent labeled deployment data are available, refit if needed and then reserve fresh observations for calibration. Without representative calibration data, do not extend the original coverage claim to the new setting. |

## 12. Bibliographic Notes

**Route: Core**

The modern conformal prediction framework is developed in Vovk et al. [190] and surveyed by Shafer and Vovk [161]. Distribution-free predictive inference for regression and split conformal methods are treated by Lei et al. [116]. CQR is due to Romano et al. [154]. Jackknife+ and related CV methods are developed by Barber et al. [12]. Conformal prediction under covariate shift through density-ratio reweighting is developed by Tibshirani et al. [178]; the broader treatment of conformal prediction beyond exchangeability is in Barber et al. [14]. The PRDS property of conformal p-values and BH validity with a shared calibration sample, used in our Theorem 10.6, are established in Bates et al. [18]. An accessible pedagogical introduction covering the breadth of the field is Angelopoulos and Bates [1]. The book-length theoretical treatment in Angelopoulos et al. [5] gives a broad account of conditional coverage, weighted variants, online conformal methods, and extensions beyond predictive coverage.

Training-conditional properties of split conformal prediction are studied by Vovk [184]; the failure of training-conditional coverage for full conformal prediction is established by Bian and Barber [26]. Group-wise calibration under the name Mondrian conformal prediction originates with Vovk et al. [189]. Distribution-free prediction bands and the limits of test-conditional coverage are developed in Lei and Wasserman [115] and Barber et al. [13]. Simultaneous or transductive conformal prediction for multiple test points is treated by Vovk [185] and Gazin et al. [70]. Selection-conditional conformal coverage and related selective conformal p-value methods are developed by Jin and Candès [100], Bao et al. [8], and Jin and Ren [101]. Aggregation of conformal prediction sets by majority vote is studied by Cherubin [43] and Gasparin and Ramdas [68].

CRC extends the conformal program from miscoverage to arbitrary monotone bounded losses; the core construction is from Angelopoulos et al. [4], and risk-controlling prediction sets in the dual form are from Bates et al. [16]. The learn-then-test (LTT) reduction, which uses multiple testing of calibration hypotheses to handle non-monotone losses, is from Angelopoulos et al. [2]. Conformal e-prediction and specific constructions for post-hoc, fixed-size, and batch anytime-valid prediction are developed by Gauthier et al. [69]; each construction requires the e-value or e-process validity conditions stated in that work.

Online full conformal prediction produces a conformal p-value for each incoming observation in a data stream; under exchangeability, with a symmetric score algorithm and no ties among the scores, the resulting sequence of p-values is mutually independent, which is the foundation of the online theory [183, 190]. The same independence supports testing exchangeability by betting: a test martingale built from the online conformal p-values accumulates evidence against exchangeability while remaining valid under optional stopping [151, 190]. Adaptive conformal inference updates the working miscoverage level online so that the realized coverage tracks the target even under distribution shift [71]. For outlier detection, conformal e-values combined with the e-BH procedure of Chapter 8 control FDR under arbitrary dependence among the test points, avoiding the PRDS analysis that BH requires [15, 114]; the positive-dependence analysis itself extends from independent to exchangeable test points in Marandon et al. [133]. A

classical universality result concerns prediction sets that are nested as the miscoverage level varies, unchanged by relabeling the observed sample, and marginally valid under every exchangeable law. A conformal predictor can give sets contained in those original sets at every level while retaining validity. This is an improvement or equality result, rather than a claim that every such procedure is itself exactly a conformal predictor; see Vovk et al. [190] and Shafer and Vovk [161, Section 4.4]. Two adjacent topics are deliberately left out of this chapter. The first is score optimality: which score functions give the shortest or best-shaped prediction sets when a working model is adopted, with conformal validity retained even when the model is wrong [5, 66, 83]. The second is probability calibration, whose target is calibrated predictive probabilities rather than covering sets; distribution-free calibration and the isotonic-regression-based Venn–Abers predictors are treated by Gupta et al. [77] and Vovk and Petej [186].

## 13. Exercises

**Route: Core**

### Intermediate.

**Exercise 10.10: Randomized distribution transform**

Let $W$ have CDF $F$, and define

$$F^*(W) = F(W-) + U\{F(W) - F(W-)\}, \qquad U \sim \mathrm{Unif}(0,1),$$

where $U$ is independent of $W$. Prove that $F^*(W) \sim \mathrm{Unif}(0,1)$ for an arbitrary CDF $F$. The finite-sample conformal application is already established in Lemma 10.4; here the task is to justify the general distributional statement, without assuming that $W$ has finite support.

### Computational.

**Exercise 10.11: Heteroscedastic simulation**

Use the design behind Figure 10.3:

$$X \sim \mathrm{Unif}(0,1), \qquad Y = \sin(2\pi X) + (0.15 + 0.75X)\varepsilon, \qquad \varepsilon \sim N(0,1),$$

with $X$ and $\varepsilon$ independent. Generate independent training, calibration, and test samples, using 240 training and 180 calibration observations and $\alpha = 0.1$. Fit the center by least squares and the 0.05 and 0.95 quantiles by linear quantile regression, all on the basis $\{1, \sin(2\pi x), \cos(2\pi x), x\}$. Order the two fitted quantile curves at each $x$ before conformal calibration.

Over at least 200 independent repetitions, each with at least 2,000 fresh test observations, compare the average interval length and empirical coverage of absolute-residual split conformal and CQR. Report Monte Carlo SEs based on the independent repetitions.

**Exercise 10.12: Jackknife+ endpoints**

For the dataset

$$(X_i, Y_i) \in \{(0, 1.0), (1, 1.8), (2, 3.2), (3, 3.9), (4, 5.1)\}, \qquad X_{n+1} = 2.5,$$

use leave-one-out least-squares simple linear regression with an intercept. Compute $R_i^{\mathrm{LOO}} = |Y_i - \widehat{f}_{-i}(X_i)|$, $\widehat{f}_{-i}(X_{n+1})$, $L_i = \widehat{f}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}$, and $U_i = \widehat{f}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}$ by hand or with a short script. For $\alpha = 0.2$, compute the jackknife+ interval from the order statistics of $\{L_i\}$ and $\{U_i\}$, reporting its final endpoints as $L_{\mathrm{JK+}}$ and $U_{\mathrm{JK+}}$.

**Exercise 10.13: Covariate shift**

Use the design behind Figure B.3:

$$X \sim N(\delta, 1), \qquad Y = \sin X + \left\{0.25 + \frac{0.75}{1 + e^{-X}}\right\}\varepsilon, \qquad \varepsilon \sim N(0, 1),$$

with $X$ and $\varepsilon$ independent. Fit a cubic polynomial by least squares on 300 training observations and calibrate absolute residuals on 300 independent observations, both drawn with $\delta = 0$. Use $\alpha = 0.1$. For test shifts $\delta \in \{0, 0.25, \ldots, 2.5\}$, compare ordinary split conformal with the weighted correction using the known ratio $w(x) = \exp(\delta x - \delta^2/2)$. Use the construction and assumptions in Section 6 of Appendix B, including its possible infinite endpoint.

At each shift, use at least 200 independent repetitions, regenerating training and calibration samples and at least 2,000 fresh test observations in each repetition. Report the average empirical coverage and its Monte Carlo SE for both methods, and the fraction of infinite weighted intervals. The theorem's coverage guarantee averages over calibration and test draws.

## Advanced.

**Exercise 10.14: Beyond marginal coverage**

Theorem 10.3 guarantees marginal coverage. Construct an example in which split conformal has valid marginal coverage but coverage well below $1 - \alpha$ conditional on $X_{n+1} \in A$, for a covariate set $A$ with $0 < \mathbb{P}(X \in A) < 1$. Specify the data distribution, predictor, calibration sample size, and $\alpha$, and calculate or estimate both the marginal coverage and the coverage on $A$ to substantiate your example.

**Exercise 10.15: Model selection breaks calibration**

Let $\widehat{f}_1$ and $\widehat{f}_2$ be two predictors pretrained on $D_1$, and for $j = 1, 2$ let $R_i^{(j)} = |Y_i - \widehat{f}_j(X_i)|$, $i \in D_2$, be their calibration scores. An analyst chooses the index $J$ with the smaller average calibration score $n_2^{-1} \sum_{i \in D_2} R_i^{(j)}$, breaking ties by a fixed rule, and then calibrates a split conformal interval for $\widehat{f}_J$ on the *same* set $D_2$.

Construct an explicit example or a small simulation in which this procedure has coverage below $1 - \alpha$. Specify the data distribution, the two predictors, the calibration sample size, and $\alpha$, and calculate or estimate the resulting coverage. Use the exchangeability requirement in the proof of Theorem 10.3 to explain the failure in your construction.

CHAPTER 11

# Debiased Lasso and High-Dimensional Confidence Intervals

The Lasso is one of the standard tools for prediction and variable screening when the number of features $p$ is comparable to or much larger than the sample size $n$. Its success is also the reason it is easy to misuse inferentially. The $\ell_1$ penalty shrinks estimates toward zero; that shrinkage is useful for prediction, but it creates bias that is not negligible on the $n^{-1/2}$ scale used by ordinary confidence intervals.

This chapter constructs confidence intervals and tests for the coefficient $\beta_j$ in a high-dimensional linear model. The coordinate $j$, or low-dimensional collection of coordinates, is specified before examining the data. The main device is to correct the Lasso by adding a nearly orthogonal score term. The correction is called debiasing, desparsifying, or one-step estimation in the literature. After developing the linear-model construction, an advanced section extends the same idea to other likelihoods, using logistic regression as a worked example. The supporting proofs are in Appendix C.

## 1. Model and Target

**Route: Core**

Table 11.1 compares three possible targets. Read each row from left to right: first identify what is random, then identify the coefficient being covered, and only then choose the variance formula and inferential interpretation.

TABLE 11.1. Targets that can be hidden behind the same linear algebra. The sampling unit and interpretation must be stated before fitting.

| Regime | Target | Interpretation |
|---|---|---|
| Fixed $X$, correct mean | $\beta$ in $E[y \mid X] = X\beta$ | A coefficient identified by the stated model restrictions, such as sparsity and suitable design conditions. |
| Random $X$, correct model | Population $\beta$ in $E[Y \mid X] = X^\top\beta$ | An identifiable conditional-mean coefficient under the sampling law for rows; a causal interpretation requires additional assumptions. |
| Misspecified mean | A least-squares projection onto a specified model | A unique coefficient requires full column rank (fixed design) or a nonsingular population second-moment matrix (random design); variance assumptions must also match the sampling model. |

When $p > n$, the unrestricted fixed-design minimization $\arg\min_b \|\mu - Xb\|_2^2$ does not identify a unique coefficient vector. One must specify restrictions that identify the coefficient of interest, or use a lower-dimensional, full-rank projection target as in Chapter 12.

We work with the linear model

$$y = X\beta + \varepsilon, \qquad X \in \mathbb{R}^{n\times p}, \quad y, \varepsilon \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^p.$$

The $i$th row of $X$ is $x_i^\top$, and the $j$th column is $X_j$. The empirical Gram matrix is

$$\widehat{\Sigma} = \frac{X^\top X}{n}.$$

For centered random rows with nonsingular covariance matrix $\Sigma$, the precision matrix is denoted by

$$\Gamma = \Sigma^{-1}.$$

Many papers use $\Theta$ for the precision matrix; in this book $\Gamma$ is the global notation, and $M$ will denote a generic empirical approximate inverse of $\widehat{\Sigma}$.

Let

$$\mathcal{S}_\beta = \{j : \beta_j \neq 0\}, \qquad s = |\mathcal{S}_\beta| = \|\beta\|_0.$$

The sparse high-dimensional regime allows $p$ to exceed $n$, but assumes that $s$ is small enough relative to $n$ and $\log p$. The target in this chapter is the original coefficient $\beta_j$, not a coefficient in a data-selected least-squares model. Post-selection targets are treated in Chapter 12.

## 2. Regularization and the Lasso

**Route: Core**

When $p > n$, least squares is not uniquely defined without further structure. Regularization imposes such structure by trading empirical fit against a complexity measure. Three common measures of complexity are

$$\|b\|_0 = \sum_{j=1}^p \mathbf{1}\{b_j \neq 0\}, \qquad \|b\|_1 = \sum_{j=1}^p |b_j|, \qquad \|b\|_2^2 = \sum_{j=1}^p b_j^2.$$

Constraints place upper bounds on these quantities. Figure 11.1 compares the resulting sets for two coefficients. The quantity $\|b\|_0$ counts nonzero coefficients; despite the notation, it is not a norm. Best subset selection uses the $\ell_0$ constraint and searches directly over sparse supports. It is statistically natural but computationally difficult. Ridge regression uses an $\ell_2$ penalty and is stable under collinearity, but it does not usually produce exact zeros. The Lasso [176] uses an $\ell_1$ penalty:

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n}\|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}. \tag{6}$$

The $\ell_1$ ball is convex and has corners, so the optimization problem is tractable and the solution is often sparse.

To see the shrinkage explicitly, consider an illustrative design with $X^\top X/n = I_p$: the columns are orthogonal and each has squared length $n$. This special case requires $p \leq n$, but it isolates the effect of the penalty. Put $z = X^\top y/n$. Completing the square gives

$$\frac{1}{2n}\|y - Xb\|_2^2 + \lambda\|b\|_1 = \text{constant} + \sum_{j=1}^p \left\{ \frac{1}{2}(b_j - z_j)^2 + \lambda |b_j| \right\},$$

where the constant does not depend on $b$. Each coordinate therefore solves a separate scalar minimization problem. Away from zero, setting the derivative to zero gives $b_j = z_j - \lambda \operatorname{sign}(b_j)$; zero minimizes the objective when $|z_j| \leq \lambda$. Together these cases give the soft-thresholding rule

$$\widehat{\beta}_j = \operatorname{sign}(z_j)(|z_j| - \lambda)_+, \qquad z_j = X_j^\top y/n.$$

Here $z_j$ is the least-squares coordinate. In this normalized orthogonal design, a coordinate beyond the threshold is still shifted toward zero by $\lambda$. Figure 11.2 shows how sparsity and shrinkage arrive together.

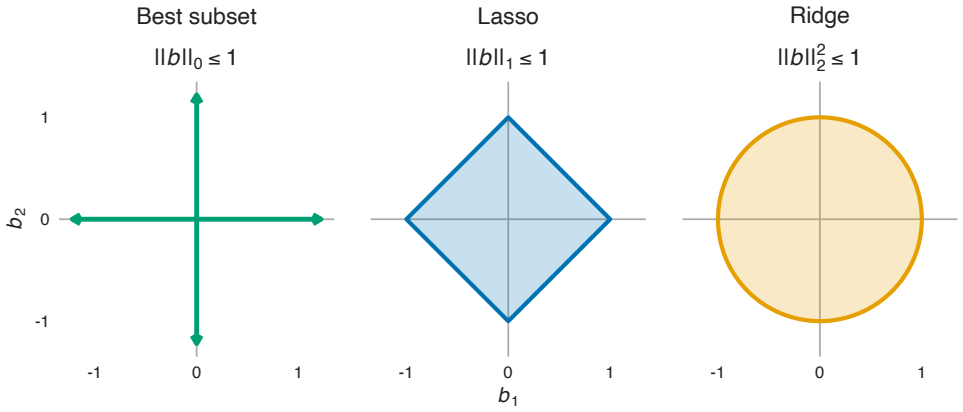



FIGURE 11.1. Three constraints on two coefficients. The set $\|b\|_0 \le 1$ permits at most one nonzero coefficient, so it consists of the two coordinate axes and extends beyond the plotted window. The $\ell_1$ constraint gives a diamond with corners on the axes, while the squared $\ell_2$ constraint gives a disk. These different shapes help explain why subset selection and the Lasso favor zero coefficients, whereas ridge regression usually retains them.

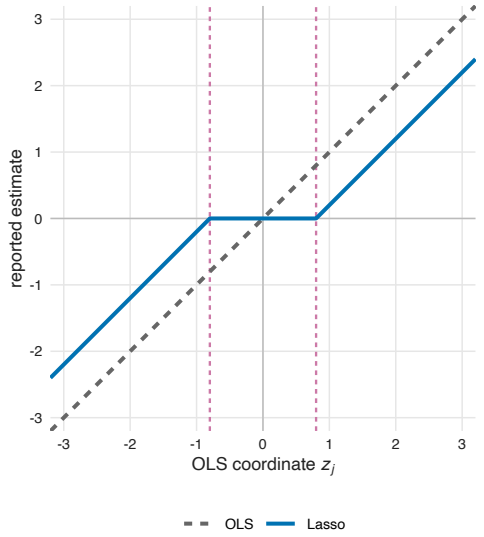



FIGURE 11.2. Lasso shrinkage when $X^\top X/n = I_p$. The dashed vertical lines mark $-\lambda$ and $\lambda$, with $\lambda = 0.8$. Soft thresholding sets the estimate to zero for $|z_j| \le \lambda$ and shifts larger coordinates toward zero by $\lambda$. This bias can remain visible on the scale used for ordinary confidence intervals.

To discuss the sampling distribution, suppose that $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$, with fixed $\sigma > 0$, and that $\lambda > 0$. The same shrinkage that helps prediction makes direct inference difficult. For a nonzero coefficient, it introduces bias; for a zero coefficient, thresholding creates a point mass at zero. In this example, ordinary centered normal approximations require $\sqrt{n}\lambda \to 0$, so the penalty must become negligible on the $n^{-1/2}$ scale. Such a small penalty generally cannot provide the regularization needed in high dimensions.

## 3. What Lasso Consistency Gives

**Route: Core**

The Lasso can still estimate a sparse vector well in prediction and $\ell_1$ norm. These rates are the starting point for debiasing.

**Definition 11.1: Compatibility condition**

For the true support $S$ with $|S| = s$, the empirical Gram matrix $\widehat{\Sigma} = X^\top X/n$ satisfies the compatibility condition with constant $\phi_0 > 0$ if, for every $u \in \mathbb{R}^p$ such that $\|u_{S^c}\|_1 \le 3\|u_S\|_1$,

$$\|u_S\|_1^2 \le \frac{s}{\phi_0^2}\, u^\top \widehat{\Sigma} u.$$

The condition says that sparse directions cannot be nearly invisible to the design. It is weaker than requiring all $p$ eigenvalues of the $p \times p$ matrix $\widehat{\Sigma}$ to be bounded away from zero. That stronger requirement is impossible when $p > n$, because $\operatorname{rank}(\widehat{\Sigma}) \le n < p$, so some eigenvalues are zero. The constant 3 in the cone restriction $\|u_{S^c}\|_1 \le 3\|u_S\|_1$ is not arbitrary: it is the cone that the Lasso error $h = \widehat{\beta} - \beta$ lives in whenever the tuning parameter satisfies the score event $\lambda \ge 2\|X^\top \varepsilon/n\|_\infty$, as the proof of Theorem 11.2 below makes precise. Other constants (e.g., $c \ge 1$ replacing 3) give an analogous restricted eigenvalue condition with different proof constants downstream.

**Theorem 11.2: A basic Lasso error bound**

Assume the compatibility condition in Definition 11.1. If the tuning parameter satisfies

$$\lambda \ge 2 \left\| \frac{X^\top \varepsilon}{n} \right\|_\infty,$$

then the Lasso estimator in (6) obeys

$$\frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2 + \lambda \|\widehat{\beta} - \beta\|_1 \le C\, \frac{s\lambda^2}{\phi_0^2}$$

for a universal constant $C$. In particular, if the displayed score bound holds with probability tending to one, $\lambda \asymp \sigma\sqrt{(\log p)/n}$, $\sigma$ is bounded, and $\phi_0$ is bounded away from zero, then

$$\|\widehat{\beta} - \beta\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right), \qquad \frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2 = O_p\left(\frac{s \log p}{n}\right).$$

**Proof of Theorem 11.2**

The optimality of $\widehat{\beta}$ implies the basic inequality obtained by comparing the objective at $\widehat{\beta}$ and at $\beta$. With $h = \widehat{\beta} - \beta$,

$$\frac{1}{2n}\|Xh\|_2^2 \le \frac{\varepsilon^\top Xh}{n} + \lambda(\|\beta\|_1 - \|\beta + h\|_1).$$

The score event bounds the stochastic term by

$$\left|\frac{\varepsilon^\top Xh}{n}\right| \le \left\|\frac{X^\top \varepsilon}{n}\right\|_\infty \|h\|_1 \le \frac{\lambda}{2}\|h\|_1.$$

Since $\beta_{S^c} = 0$, decomposability of the $\ell_1$ norm gives

$$\|\beta\|_1 - \|\beta + h\|_1 \le \|h_S\|_1 - \|h_{S^c}\|_1.$$

Combining these two displays yields

$$\frac{1}{2n}\|Xh\|_2^2 + \frac{\lambda}{2}\|h_{S^c}\|_1 \le \frac{3\lambda}{2}\|h_S\|_1.$$

Dropping the nonnegative prediction term shows $\|h_{S^c}\|_1 \le 3\|h_S\|_1$, so $h$ lies in the compatibility cone. The compatibility condition then gives

$$\|h_S\|_1 \le \frac{\sqrt{s}}{\phi_0}\frac{\|Xh\|_2}{\sqrt{n}}.$$

Let $P = \|Xh\|_2^2/n$. The previous two displays imply

$$\frac{P}{2} \le \frac{3\lambda}{2}\frac{\sqrt{s}}{\phi_0}\sqrt{P}.$$

If $P = 0$, the desired prediction bound already holds. Otherwise, divide by $\sqrt{P}/2$ to obtain $\sqrt{P} \le 3\lambda\sqrt{s}/\phi_0$, and hence $P \le 9s\lambda^2/\phi_0^2$. Finally,

$$\|h\|_1 \le 4\|h_S\|_1 \le \frac{4\sqrt{s}}{\phi_0}\sqrt{P} \le \frac{12s\lambda}{\phi_0^2}.$$

Multiplying the last display by $\lambda$ and adding the prediction bound gives

$$P + \lambda\|h\|_1 \le \frac{9s\lambda^2}{\phi_0^2} + \frac{12s\lambda^2}{\phi_0^2} = \frac{21s\lambda^2}{\phi_0^2}.$$

Thus $C = 21$ is one valid universal constant. On the score event, substituting $\lambda \asymp \sigma\sqrt{(\log p)/n}$ into the two separate bounds gives

$$\|h\|_1 \le \frac{12s\lambda}{\phi_0^2} = O\left(s\sigma\sqrt{\frac{\log p}{n}}\right), \qquad P \le \frac{9s\lambda^2}{\phi_0^2} = O\left(\frac{s\sigma^2 \log p}{n}\right).$$

Here $\phi_0$ is bounded away from zero. Since this event has probability tending to one and $\sigma$ is bounded, these bounds give the stated $O_p$ rates. □

Theorem 11.2 bounds the estimation error. Its prediction-error bound tends to zero under the usual condition $s \log p/n \to 0$. Its $\ell_1$-error bound is consistent under the stronger condition

$$s\sqrt{\frac{\log p}{n}} \to 0, \qquad \text{equivalently} \qquad \frac{s^2 \log p}{n} \to 0.$$

Debiased inference needs more than either consistency statement: after multiplication by $\sqrt{n}$, the bias remainder must still vanish.

## 4. Why the Lasso Limit Is Not Enough

**Route: Core**

The nonstandard behavior of the Lasso is already visible when $p$ is fixed. Suppose $p$ and $\beta$ are fixed, $X^\top X/n \xrightarrow{P} \Sigma_0$ with $\Sigma_0$ positive definite, and

$$\frac{X^\top \varepsilon}{\sqrt{n}} \xrightarrow{d} W, \qquad W \sim N(0, \sigma^2 \Sigma_0).$$

To find the limiting distribution, examine the objective near the true coefficient. Let $L_n(b)$ denote the objective in (6), with penalty $\lambda_n$, and write $b = \beta + u/\sqrt{n}$. Subtracting its value at $\beta$ and multiplying by $n$ gives

$$\begin{aligned} V_n(u) &= n\{L_n(\beta + u/\sqrt{n}) - L_n(\beta)\} \\ &= \frac{1}{2} u^\top \widehat{\Sigma} u - u^\top \frac{X^\top \varepsilon}{\sqrt{n}} + n\lambda_n \sum_{j=1}^{p} (|\beta_j + u_j/\sqrt{n}| - |\beta_j|). \end{aligned}$$

Its minimizer is exactly $\sqrt{n}(\widehat{\beta} - \beta)$: the change of variables changes the scale of the minimizer, while subtracting a constant and multiplying by a positive number do not change which coefficient fits best. For each fixed $u_j$,

$$\sqrt{n}(|\beta_j + u_j/\sqrt{n}| - |\beta_j|) \longrightarrow \begin{cases} u_j \operatorname{sign}(\beta_j), & \beta_j \neq 0, \\ |u_j|, & \beta_j = 0. \end{cases}$$

A nonzero coefficient retains its sign under a sufficiently small perturbation, making the penalty locally linear. At a zero coefficient, the absolute-value kink remains. Therefore, if $\sqrt{n}\,\lambda_n \to \lambda_0 \in [0, \infty)$, the limiting objective is

$$V(u) = \frac{1}{2} u^\top \Sigma_0 u - u^\top W + \lambda_0 \sum_{j=1}^{p} \left[ u_j \operatorname{sign}(\beta_j) \mathbf{1}\{\beta_j \neq 0\} + |u_j| \mathbf{1}\{\beta_j = 0\} \right].$$

Positive definiteness of $\Sigma_0$ makes this function strictly convex, so its minimizer is unique. The convex minimization argument of Knight and Fu [106] yields

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} U = \arg\min_{u \in \mathbb{R}^p} V(u).$$

The minimizer is random because each realization of $W$ produces a different objective and hence a different minimizing vector. The distribution of that vector is the limiting distribution of the scaled estimation error. When $\lambda_0 > 0$, it is generally not centered normal: the kink can produce a point mass, and the linear penalty can shift the center. When $\lambda_0 = 0$, the minimizer is $\Sigma_0^{-1} W$, recovering the usual centered normal limit. High-dimensional debiased methods instead construct a new estimator whose regularization bias is negligible on the inferential scale.

## 5. Debiasing by an Approximate Inverse

**Route: Core**

Let $M \in \mathbb{R}^{p\times p}$ be a matrix, computed from $X$, whose rows nearly invert the empirical Gram matrix. Write $m_j^\top$ for the $j$th row of $M$. The debiased Lasso estimator is

$$\widehat{b} = \widehat{\beta} + M\frac{X^\top(y - X\widehat{\beta})}{n}. \tag{7}$$

Substituting $y = X\beta + \varepsilon$ gives the exact decomposition

$$\widehat{b} - \beta = M\frac{X^\top\varepsilon}{n} + (I - M\widehat{\Sigma})(\widehat{\beta} - \beta). \tag{8}$$

For coordinate $j$,

$$\sqrt{n}(\widehat{b}_j - \beta_j) = \frac{m_j^\top X^\top\varepsilon}{\sqrt{n}} + \sqrt{n}\{e_j^\top - m_j^\top\widehat{\Sigma}\}(\widehat{\beta} - \beta), \tag{9}$$

where $e_j$ is the $j$th standard basis vector.

The first term in (9) is a score term. If $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$, then conditionally on $X$,

$$\frac{m_j^\top X^\top\varepsilon}{\sqrt{n}} \sim N\left(0, \sigma^2 m_j^\top\widehat{\Sigma}m_j\right).$$

For non-Gaussian errors, the same normal approximation follows from a Lindeberg-type CLT under standard moment conditions. The second term is the price of using an approximate inverse rather than the exact inverse of $\widehat{\Sigma}$, which typically does not exist when $p > n$.

Define the sample-space direction $d_j = Xm_j \in \mathbb{R}^n$. Then

$$\frac{d_j^\top X_k}{n} = m_j^\top\widehat{\Sigma}e_k,$$

so an approximate inverse makes this inner product close to one for the target column $k = j$, and close to zero for each other column. Figure 11.3 connects this requirement to residualizing a target column against a nuisance column. In high dimensions we control the individual inner products, not the projection onto the entire nuisance span, which may equal $\mathbb{R}^n$.

The remainder is controlled by Hölder's inequality:

$$\left|\sqrt{n}\{e_j^\top - m_j^\top\widehat{\Sigma}\}(\widehat{\beta} - \beta)\right| \le \sqrt{n}\,\|m_j^\top\widehat{\Sigma} - e_j^\top\|_\infty\|\widehat{\beta} - \beta\|_1. \tag{10}$$

If

$$\|m_j^\top\widehat{\Sigma} - e_j^\top\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

and Theorem 11.2 gives $\|\widehat{\beta} - \beta\|_1 = O_p(s\sqrt{(\log p)/n})$, then the scaled remainder is

$$O_p\left(\frac{s\log p}{\sqrt{n}}\right).$$

Thus a common sufficient condition for coordinatewise normality is

$$\frac{s\log p}{\sqrt{n}} \to 0. \tag{11}$$

This condition is stronger than the condition needed for prediction consistency, because inference must remove bias on the $n^{-1/2}$ scale.

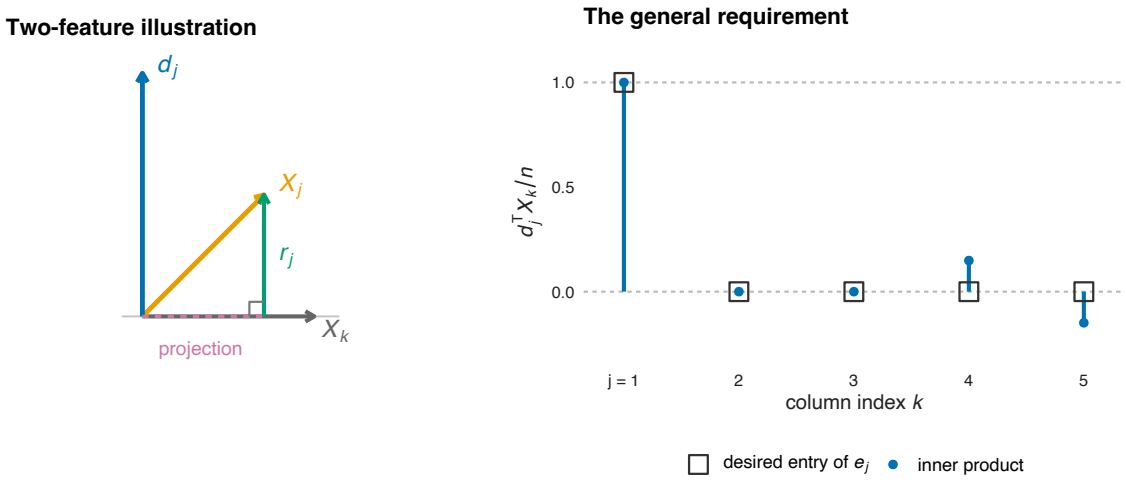



FIGURE 11.3. From residualization to approximate orthogonality. Left: in a two-feature illustration, subtracting the projection of $X_j$ onto $X_k$ gives a perpendicular residual $r_j$. Rescaling it as $d_j = nr_j/(r_j^\top X_j)$ gives $d_j^\top X_j/n = 1$. Right: an example with $n = 3$ and $p = 5$ shows inner products $d_j^\top X_k/n$ close to the entries of $e_j$, rather than exact orthogonality to the whole nuisance span. Small nuisance inner products limit sensitivity to errors in the other coefficients. Section 8 extends this idea beyond linear regression.

**Theorem 11.3: Coordinatewise debiased normality**

Fix a coordinate $j$ and a noise standard deviation $\sigma > 0$ that does not change with $n$. Suppose the Lasso satisfies $\|\widehat{\beta} - \beta\|_1 = O_p(s\sqrt{(\log p)/n})$, the row $m_j$ is $X$-measurable and satisfies $\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = O_p(\sqrt{(\log p)/n})$, and $s \log p/\sqrt{n} \to 0$. Suppose also that

$$0 < c \le m_j^\top \widehat{\Sigma} m_j \le C < \infty$$

with probability tending to one. Suppose also that, conditional on $X$, the noise satisfies one of the following conditions:

(i) $\varepsilon \sim N(0, \sigma^2 I_n)$;

(ii) the errors are independent with $E(\varepsilon_i \mid X) = 0$, $E(\varepsilon_i^2 \mid X) = \sigma^2$, and, for every $\eta > 0$,

$$\frac{\sum_{i=1}^n (x_i^\top m_j)^2 E[\varepsilon_i^2 \mathbf{1}\{|(x_i^\top m_j)\varepsilon_i| > \eta\sigma\|Xm_j\|_2\} \mid X]}{\sigma^2\|Xm_j\|_2^2} \xrightarrow{P} 0.$$

Then

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j)}{\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}} \xrightarrow{d} N(0,1).$$

Condition (ii) is the conditional Lindeberg condition. A convenient way to ensure it is to assume a uniform conditional $(2+\delta)$-moment bound for some $\delta > 0$, together with

$$\frac{\max_{1\le i\le n} |(Xm_j)_i|}{\|Xm_j\|_2} \to 0.$$

The ratio requires each observation's score weight to be small relative to the combined weights of all observations.

**Proof of Theorem 11.3**

Use the decomposition (9). The leading score term is conditionally Gaussian under Gaussian errors:

$$\frac{m_j^\top X^\top \varepsilon/\sqrt{n}}{\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}} \mid X \sim N(0,1).$$

Under condition (ii), the same standardized score is asymptotically normal by the conditional Lindeberg–Feller theorem. The influence-and-moment condition displayed above is sufficient because it prevents any single weighted error from dominating the sum. The remainder satisfies

$$\left|\sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta}-\beta)\right| \le \sqrt{n}\, O_p\left(\sqrt{\frac{\log p}{n}}\right) O_p\left(s\sqrt{\frac{\log p}{n}}\right) = O_p\left(\frac{s\log p}{\sqrt{n}}\right) = o_p(1).$$

Dividing by $\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}$, which is bounded away from zero and infinity with probability tending to one, keeps the remainder $o_p(1)$. Slutsky's theorem completes the argument. □

**Corollary 11.4: Heteroskedastic robust debiased interval**

Retain the rate and approximate-inverse assumptions of Theorem 11.3. Conditional on $X$, suppose the errors are independent, mean zero, and have variances $\sigma_i^2 = E(\varepsilon_i^2 \mid X)$, with the corresponding conditional Lindeberg condition. Define

$$v_j^2 = \frac{1}{n}\sum_{i=1}^n (x_i^\top m_j)^2 \sigma_i^2.$$

For residuals $\widehat{\varepsilon}_i = y_i - x_i^\top \widehat{\beta}$, define

$$\widehat{v}_j^2 = \frac{1}{n}\sum_{i=1}^n (x_i^\top m_j)^2 \widehat{\varepsilon}_i^2. \tag{12}$$

If $v_j$ is bounded away from zero and infinity and $\widehat{v}_j^2 - v_j^2 \xrightarrow{P} 0$, then

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j)}{\widehat{v}_j} \xrightarrow{d} N(0,1),$$

and a robust $(1-\alpha)$ interval is

$$\widehat{b}_j \pm z_{1-\alpha/2}\frac{\widehat{v}_j}{\sqrt{n}}.$$

**Proof of Corollary 11.4**

Conditional on $X$, the variance of the leading score is $v_j^2$. The conditional Lindeberg–Feller theorem gives its normal limit, the same remainder bound is $o_p(1)$, and sandwich consistency plus Slutsky's theorem gives the studentized statement. □

Here is one sufficient set of conditions for the variance consistency used in the corollary. Write $a_i = x_i^\top m_j$, and suppose the conditional fourth moments satisfy $\max_i E(\varepsilon_i^4 \mid X) \le C$ for a fixed constant $C$. Suppose also that

$$\frac{1}{n^2}\sum_{i=1}^n a_i^4 = o_p(1), \qquad B_n := \frac{1}{n}\sum_{i=1}^n a_i^2\{x_i^\top(\widehat\beta - \beta)\}^2 = o_p(1).$$

The first condition prevents a few observations from dominating the average of weighted squared errors. The second requires small prediction errors at observations carrying large score weights. To see why these conditions work, let $W_n = n^{-1}\sum_i a_i^2\varepsilon_i^2$. Conditional independence gives

$$E(W_n \mid X) = v_j^2, \qquad \mathrm{Var}(W_n \mid X) \le \frac{C}{n^2}\sum_{i=1}^n a_i^4 = o_p(1),$$

so $W_n - v_j^2 = o_p(1)$. Expanding the squared residuals and applying Cauchy–Schwarz then gives

$$|\widehat v_j^2 - W_n| \le 2\sqrt{W_n B_n} + B_n = o_p(1).$$

This argument allows the same observations to be used for fitting and for variance estimation. A convenient way to check the weighted prediction condition is

$$B_n \le \max_i a_i^2 \, \frac{\|X(\widehat\beta - \beta)\|_2^2}{n}.$$

Thus ordinary prediction consistency suffices when the score weights are uniformly bounded. Related high-dimensional results for robust SEs are given by Dezeure et al. [48, Section 2.2 and Theorem 2].

Under the homoskedastic assumptions of Theorem 11.3, a consistent estimate $\widehat\sigma \xrightarrow{P} \sigma$ instead gives the variance estimate

$$\widehat{\mathrm{Var}}_{\mathrm{hom}}(\widehat b_j) = \frac{\widehat\sigma^2}{n}\, m_j^\top \widehat\Sigma m_j. \tag{13}$$

Slutsky's theorem then gives the $1-\alpha$ confidence interval

$$\widehat b_j \pm z_{1-\alpha/2}\,\widehat\sigma\sqrt{\frac{m_j^\top\widehat\Sigma m_j}{n}}. \tag{14}$$

The scaled Lasso [172] estimates the regression coefficients and noise level together. A simple alternative is the residual mean square, $\widehat\sigma^2 = \|y - X\widehat\beta\|_2^2/n$. This is consistent if $r_n := \|X(\widehat\beta - \beta)\|_2/\sqrt{n} = o_p(1)$ and $\|\varepsilon\|_2^2/n \xrightarrow{P} \sigma^2$, because

$$\left|\widehat\sigma^2 - \frac{\|\varepsilon\|_2^2}{n}\right| \le r_n^2 + 2r_n\frac{\|\varepsilon\|_2}{\sqrt{n}} = o_p(1).$$

In words, the fitted mean must be accurate enough that replacing the errors by residuals has a negligible effect on their average square. Fitting too much noise can make a residual estimate too small, while leaving systematic signal in the residuals can make it too large.

Read Figure 11.4 from the top row to the bottom as the sparsity burden increases, and from the left column to the right as feature correlation increases. The main check is whether empirical

coverage stays near the nominal 95 percent line. Here the AR(1) design has $\Sigma_{jk} = \rho^{|j-k|}$, so $\rho$ directly controls how quickly feature correlations decay with coordinate distance.




FIGURE 11.4. A simulation for $n = 120$, $p = 70$, Gaussian AR(1) designs with correlations $\rho = 0.1$ or 0.55, target $\beta_1 = 0.55$, $\sigma = 1$, and 500 independent simulation runs per setting. All three methods use the known noise standard deviation. The naive interval is $\widehat{\beta}_1 \pm 1.96\sigma/\sqrt{n}$. Its center is affected by shrinkage, and its variance formula $\sigma^2/n$ does not adjust for correlated predictors. Both choices affect its coverage. Debiased score intervals are much closer to the nominal 95 percent level. The oracle support benchmark fits OLS using only the truly nonzero coordinates, whose identities are supplied to the procedure. These identities would be unknown in an ordinary analysis. Vertical bars are pointwise 95 percent Wald simulation intervals for empirical coverage.

Figure 11.5 compares intervals using the homoskedastic variance estimate in (13) and the residual sandwich estimate $\widehat{\mathrm{Var}}_{\mathrm{rob}}(\widehat{b}_j) = \widehat{v}_j^2/n$, where $\widehat{v}_j^2$ is defined in (12). The two methods have similar coverage in the constant-variance setting. In the second setting, observations with larger $|X_{i1}|$ have larger noise variance. A single estimate of $\sigma^2$ does not account for this variation: the variance of the leading score weights each observation's noise variance by $(x_i^\top m_j)^2$. The sandwich estimate retains these observation-specific contributions by using the individual squared residuals.

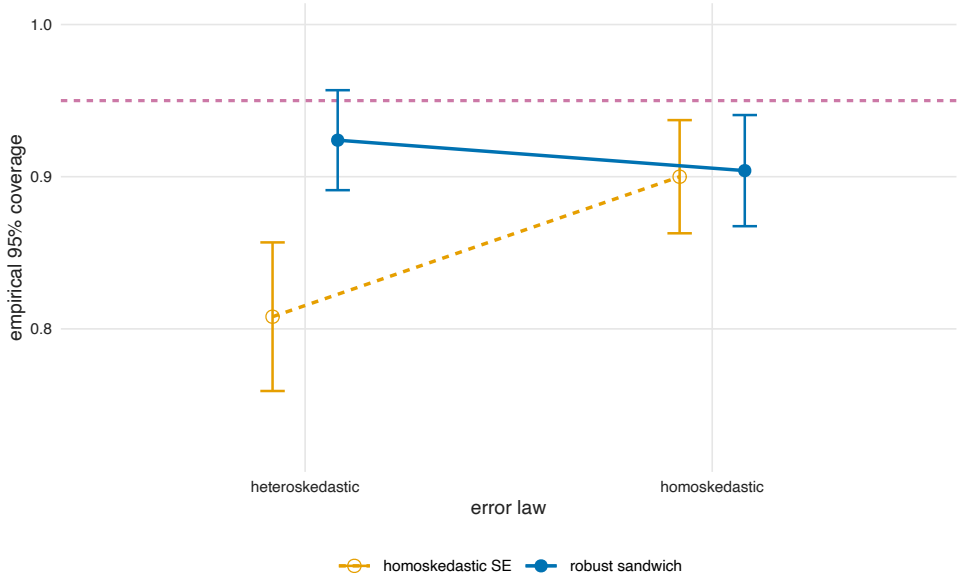



FIGURE 11.5. Coverage of nominal 95 percent debiased-Lasso intervals for $\beta_1 = 0.55$, using homoskedastic and robust SEs under constant and covariate-dependent noise variance. The simulation uses $n = 120$, $p = 70$, an AR(1) design with $\rho = 0.4$, five nonzero coefficients $(0.55, 0.45, 0.40, 0.35, 0.30)$, and 250 repetitions per error law. Constant noise has standard deviation 1; covariate-dependent noise has conditional standard deviation $0.45 + 0.8|X_{i1}|$. The homoskedastic intervals use $\widehat{\sigma}^2 = n^{-1}\sum_i \widehat{\varepsilon}_i^2$. Vertical bars are pointwise 95 percent Wald simulation intervals; the dashed horizontal line marks nominal 95 percent coverage.

## 6. Nodewise Lasso and Precision Approximation

**Route: Advanced**

It remains to construct rows $m_j$ that nearly invert $\widehat{\Sigma}$. A useful population identity comes from Gaussian graphical models. Let

$$X^\circ = (X_1^\circ, \dots, X_p^\circ)^\top \sim N(0, \Sigma), \qquad \Gamma = \Sigma^{-1},$$

where $X^\circ$ is an abstract population row used only to derive the precision identity below. The superscript distinguishes it from the observed design matrix. The population regression of $X_j^\circ$ on $X_{-j}^\circ$ has coefficient vector

$$\gamma_j^\star = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j},$$

and residual variance

$$\tau_j^2 = \Sigma_{jj} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}.$$

The block inverse formula gives

$$\Gamma_{jj} = \frac{1}{\tau_j^2}, \qquad \Gamma_{j,-j} = -\frac{(\gamma_j^\star)^\top}{\tau_j^2}. \tag{15}$$

Thus each row of the precision matrix can be recovered from a regression of one feature on the remaining features. When $p \gg n$, those regressions are again high-dimensional, so they are regularized by Lasso.

For each $j$, define the nodewise Lasso estimator

$$\widehat{\gamma}_j \in \arg\min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n}\|X_j - X_{-j}\gamma\|_2^2 + 2\lambda_j\|\gamma\|_1 \right\}. \tag{16}$$

The factor in front of the squared-loss term is a normalization choice; the common implementation in the `glmnet` package uses $1/(2n)$ instead of $1/n$. Multiplying the displayed objective by $1/2$ leaves the minimizer unchanged and gives the usual solver form with penalty coefficient $\lambda_j$. Calibrating $\lambda_j$ therefore requires matching the entire objective convention of the solver, not just comparing the raw penalty coefficient in isolation. Let $\widehat{c}_j \in \mathbb{R}^p$ be the vector with

$$(\widehat{c}_j)_j = 1, \qquad (\widehat{c}_j)_k = -\widehat{\gamma}_{j,k} \quad (k \neq j),$$

where $\widehat{\gamma}_{j,k}$ is the coefficient assigned to column $k$ in the regression excluding column $j$. Set

$$\widehat{\tau}_j^2 = \frac{1}{n}\|X_j - X_{-j}\widehat{\gamma}_j\|_2^2 + \lambda_j\|\widehat{\gamma}_j\|_1.$$

This combination is the one given by the Karush–Kuhn–Tucker (KKT) identity for the nodewise program: $X_j^\top(X_j - X_{-j}\widehat{\gamma}_j)/n = \widehat{\tau}_j^2$, so the diagonal condition $(\widehat{\Gamma}\widehat{\Sigma})_{jj} = 1$ holds exactly. To see the identity, write $\widehat{r}_j = X_j - X_{-j}\widehat{\gamma}_j$. The KKT conditions for (16) give

$$\frac{X_{-j}^\top \widehat{r}_j}{n} = \lambda_j \widehat{\kappa}_j, \qquad \widehat{\kappa}_j \in \partial\|\widehat{\gamma}_j\|_1,$$

Here $\partial\|\widehat{\gamma}_j\|_1$ denotes the set of subgradients of the $\ell_1$ norm. For each $k \neq j$, the component $\widehat{\kappa}_{j,k}$ equals $\operatorname{sign}(\widehat{\gamma}_{j,k})$ when $\widehat{\gamma}_{j,k} \neq 0$, and lies in $[-1,1]$ when $\widehat{\gamma}_{j,k} = 0$. Multiplying each component by its coefficient therefore gives $\widehat{\gamma}_j^\top \widehat{\kappa}_j = \|\widehat{\gamma}_j\|_1$. Since $X_j = \widehat{r}_j + X_{-j}\widehat{\gamma}_j$,

$$\frac{X_j^\top \widehat{r}_j}{n} = \frac{\|\widehat{r}_j\|_2^2}{n} + \widehat{\gamma}_j^\top \frac{X_{-j}^\top \widehat{r}_j}{n} = \frac{\|\widehat{r}_j\|_2^2}{n} + \lambda_j\|\widehat{\gamma}_j\|_1 = \widehat{\tau}_j^2.$$

The nodewise precision-row estimator is

$$\widehat{\Gamma}_j^\top = \frac{\widehat{c}_j^\top}{\widehat{\tau}_j^2}. \tag{17}$$

Stacking these rows gives $\widehat{\Gamma}$, which can be used as $M$ in (7). Under sparsity of the precision rows and regularity of the design, nodewise Lasso satisfies

$$\|\widehat{\Gamma}\widehat{\Sigma} - I\|_{\max} := \max_{j,k} |(\widehat{\Gamma}\widehat{\Sigma} - I)_{jk}| = O_p\left(\sqrt{\frac{\log p}{n}}\right),$$

which is the approximation needed in (10); see Meinshausen and Bühlmann [136], van de Geer et al. [180], and Dezeure et al. [47].

To see how nodewise regression helps inference, consider the two-feature model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, with target $\beta_1$. The nodewise residual is $\widehat{r}_1 = X_1 - X_2\widehat{\gamma}_1$. Provided $\widehat{\tau}_1^2 > 0$, set $d_1 = Xm_1 = \widehat{r}_1/\widehat{\tau}_1^2$. The KKT identity above gives $d_1^\top X_1/n = 1$. Therefore, substituting the model into the debiased estimator gives

$$\widehat{b}_1 = \widehat{\beta}_1 + \frac{d_1^\top(y - X\widehat{\beta})}{n}, \qquad \widehat{b}_1 - \beta_1 = \frac{d_1^\top \varepsilon}{n} + \frac{d_1^\top X_2}{n}(\beta_2 - \widehat{\beta}_2).$$

The normalization cancels the initial estimation error in $\beta_1$. The first term on the right is noise. The second is the error in estimating the other coefficient, $\beta_2$, multiplied by $d_1^\top X_2/n$. Making this inner product small reduces how strongly that error affects the corrected estimate of $\beta_1$.

Nodewise Lasso controls this multiplier through its KKT conditions:

$$\left|\frac{d_1^\top X_2}{n}\right| \le \frac{\lambda_1}{\widehat{\tau}_1^2}.$$

Thus the remaining error is small when this bound and the error in $\widehat{\beta}_2$ have a small product. With more features, one sums these contributions over all other coefficients. The next section develops that decomposition and explains how the choice of direction also affects the noise variance.

## 7. The Optimal-Projection View

**Route: Advanced**

The same construction can be understood without starting from a precision matrix. Fix a target coordinate $j$ and rewrite the model as

$$y - X_{-j}\beta_{-j} = X_j\beta_j + \varepsilon.$$

The left side is not observed because $\beta_{-j}$ is unknown. Replacing it by the Lasso gives

$$\widehat{\eta}_j = y - X_{-j}\widehat{\beta}_{-j} = X_j\beta_j + \varepsilon + X_{-j}(\beta_{-j} - \widehat{\beta}_{-j}).$$

Consider a sample-space direction $d_j \in \mathbb{R}^n$, now normalized to satisfy $d_j^\top X_j = n$. Define

$$\widetilde{\beta}_j = \frac{d_j^\top \widehat{\eta}_j}{n}.$$

Then

$$\sqrt{n}(\widetilde{\beta}_j - \beta_j) = \frac{d_j^\top \varepsilon}{\sqrt{n}} + \frac{d_j^\top X_{-j}(\beta_{-j} - \widehat{\beta}_{-j})}{\sqrt{n}}. \tag{18}$$

The first term has conditional variance $\sigma^2\|d_j\|_2^2/n$. The second term is bounded by

$$\sqrt{n}\left\|\frac{d_j^\top X_{-j}}{n}\right\|_\infty \|\widehat{\beta}_{-j} - \beta_{-j}\|_1.$$

Thus $d_j$ should have moderate Euclidean norm and small correlations with the nuisance columns $X_{-j}$, while keeping unit sensitivity to $X_j$. The nodewise residual direction is one way to produce such a $d_j$.

Javanmard and Montanari [97] formulated this tradeoff row by row. Their program lives in parameter space rather than sample space: parametrize the sample-space direction as $d_j = Xm_j$ for a vector $m_j \in \mathbb{R}^p$. Then

$$d_j^\top X/n = m_j^\top X^\top X/n = m_j^\top \widehat{\Sigma},$$

so requiring $d_j^\top X_j/n \approx 1$ and small $d_j^\top X_k/n$ for $k \neq j$ is exactly the requirement $m_j^\top \widehat{\Sigma} \approx e_j^\top$. In these coordinates the Javanmard–Montanari row program reads

$$\widehat{m}_j \in \arg\min_{m\in\mathbb{R}^p} m^\top \widehat{\Sigma} m \quad \text{subject to} \quad \|m^\top \widehat{\Sigma} - e_j^\top\|_\infty \le \eta, \tag{19}$$

with $\eta \asymp \sqrt{(\log p)/n}$. The program is trivially feasible when $\eta \ge 1$, because $m = 0$ then satisfies the constraint $\|e_j\|_\infty \le \eta$. That feasibility is not statistically useful: the standard rates come from the much smaller regime $\eta \asymp \sqrt{(\log p)/n}$, which is feasible under sparse-precision designs. Setting $\eta$ too small can leave an empty feasible set. For example, when $\eta = 0$, feasibility requires $e_j$ to belong to the column space of $\widehat{\Sigma}$, which need not hold when $p > n$. The constraint controls the bias remainder, and the objective controls the asymptotic variance $\sigma^2 m_j^\top \widehat{\Sigma} m_j$, which

equals $\sigma^2\|d_j\|_2^2/n$ under $d_j = Xm_j$. This makes the statistical tradeoff explicit: more exact orthogonalization reduces bias but may require a high-variance direction.

## 8. Beyond Linear Regression: Decorrelated Scores

**Route: Advanced**

The projection argument suggests a broader question. Can we construct a score for one coefficient that is insensitive to errors in estimating the other coefficients, even when the model is not linear? For a binary outcome, for example, we may want a confidence interval for one logistic-regression coefficient while adjusting for many other predictors. The construction of Ning and Liu [141] answers this question by removing the contribution of the other parameter scores. We first describe the construction for a correctly specified likelihood; general losses and misspecification are discussed at the end of the section and in Appendix C.

**A target score and its nuisance parameters.** Suppose $U_1, \dots, U_n$ are independent observations with a common density $f(u;\beta)$, where $\beta \in \mathbb{R}^p$. Partition the parameter as $\beta = (\theta, \gamma^\top)^\top$: $\theta \in \mathbb{R}$ is the coefficient of interest, and $\gamma \in \mathbb{R}^{p-1}$ contains the other coefficients. These are called *nuisance parameters* because they must be accounted for even though they are not the target. Write $\beta^\star = (\theta^\star, (\gamma^\star)^\top)^\top$ for the true value. All population expectations in this section use the true data law, which is held fixed when we differentiate the score. For regression, we may use the conditional response density given the covariates; expectations then average over independent response–covariate pairs. The average negative log-likelihood is

$$\ell_n(\theta,\gamma) = -\frac{1}{n}\sum_{i=1}^{n} \log f(U_i;\theta,\gamma).$$

A score measures how the fit changes when a parameter changes. Here we use derivatives of the *negative* log-likelihood, so our scores have the opposite sign to the usual log-likelihood scores. In particular, $\partial_\theta \ell_n$ is a scalar, whereas $\partial_\gamma \ell_n$ is the vector of nuisance derivatives.

Simply substituting an estimate $\widehat{\gamma}$ into the target score can leave a large error. To see why, hold $\theta = \theta^\star$ fixed and expand the scalar target score $\gamma \mapsto \partial_\theta \ell_n(\theta^\star, \gamma)$ around the true nuisance parameter $\gamma^\star$:

$$\begin{aligned}\partial_\theta \ell_n(\theta^\star, \widehat{\gamma}) = \partial_\theta \ell_n(\theta^\star, \gamma^\star) \\ + \partial^2_{\theta\gamma} \ell_n(\theta^\star, \gamma^\star)(\widehat{\gamma} - \gamma^\star) + R_n.\end{aligned}$$

Here $R_n$ is the higher-order Taylor remainder. The first term on the right is the score we would compute if the nuisance parameter were known. The next term shows how estimating it changes the score: the row of mixed second derivatives multiplies the nuisance estimation error. Even when the score at the truth has the usual $n^{-1/2}$ fluctuations, this added term need not be negligible on that scale.

We therefore modify the target score so that its expected derivative with respect to the nuisance parameters vanishes at the truth. The remaining empirical error and Taylor remainder will still need to be controlled. Let $I = E\{\nabla^2 \ell_n(\beta^\star)\}$ be the Fisher information per observation, with blocks indexed by $\theta$ and $\gamma$. The Hessian $\nabla^2 \ell_n$ is the matrix of second derivatives, describing local curvature. Assume $I_{\gamma\gamma}$ is invertible and set

$$w^\star = I_{\gamma\gamma}^{-1} I_{\gamma\theta}.$$

This population inverse defines the desired direction; its empirical version need not exist when $p > n$. Define the *decorrelated score*

$$S_n(\theta,\gamma;w) = \partial_\theta \ell_n(\theta,\gamma) - w^\top \partial_\gamma \ell_n(\theta,\gamma). \tag{20}$$

With $w = w^\star$ held fixed, its expected derivative in the nuisance directions, evaluated at the truth, is

$$E\{\partial_{\gamma^\top} S_n(\beta^\star; w^\star)\} = I_{\theta\gamma} - w^{\star\top} I_{\gamma\gamma} = 0.$$

Here $S_n(\beta; w)$ abbreviates $S_n(\theta, \gamma; w)$. Thus small changes in the nuisance parameters have no first-order effect on the population score. This property is called *Neyman orthogonality*. It reduces sensitivity to estimation error; it does not make the error disappear.

Under the usual likelihood regularity conditions, the covariance of the per-observation score vector equals its expected Hessian. The same cancellation therefore makes the decorrelated score uncorrelated with the nuisance scores at the truth, which explains the name "decorrelated." It does not assert independence. The remaining information about $\theta$ is

$$J = I_{\theta\theta} - w^{\star\top} I_{\gamma\theta}.$$

This is the *partial information*: it measures the information left after accounting for the nuisance parameters. For the regular likelihood considered here, $\mathrm{Var}\{\sqrt{n} S_n(\beta^\star; w^\star)\} = J$.

**Estimating the direction and using the score.** Start with a full penalized fit

$$\widehat{\beta} = (\widehat{\theta}, \widehat{\gamma}^\top)^\top \in \arg\min_\beta \{\ell_n(\beta) + \lambda \|\beta\|_1\}, \qquad \widehat{H} = \nabla^2 \ell_n(\widehat{\beta}).$$

The penalty $\lambda$ controls the initial fit. A second tuning parameter $\lambda_w \geq 0$ controls how closely the estimated direction removes nuisance sensitivity. One choice is

$$\widehat{w} \in \arg\min_w \|w\|_1 \quad \text{subject to} \quad \|\widehat{H}_{\gamma\theta} - \widehat{H}_{\gamma\gamma} w\|_\infty \leq \lambda_w. \tag{21}$$

The constraint bounds every remaining nuisance derivative; minimizing the $\ell_1$ norm favors a sparse direction. This is a Dantzig-type program, meaning that it imposes a bound on the estimating-equation error rather than inverting the full empirical matrix. Its statistical accuracy requires conditions on the direction $w^\star$, as well as on $\beta^\star$.

With $\widehat{\gamma}$ and $\widehat{w}$ fixed, define

$$\widehat{S}_n(\theta) = S_n(\theta, \widehat{\gamma}; \widehat{w}).$$

The estimated partial information is the derivative of this function at the initial estimate:

$$\widehat{J} = \widehat{H}_{\theta\theta} - \widehat{w}^\top \widehat{H}_{\gamma\theta}. \tag{22}$$

For a prespecified null $H_0 : \theta^\star = \theta_0$, the score statistic is $T_n = \sqrt{n}\, \widehat{S}_n(\theta_0)/\sqrt{\widehat{J}}$. The nuisance estimate comes from the full fit; this construction does not require fitting a new model under each proposed null.

For estimation, take one Newton step toward a zero of the fitted score:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{S}_n(\widehat{\theta})}{\widehat{J}}. \tag{23}$$

This update solves the linear equation $\widehat{S}_n(\widehat{\theta}) + \widehat{J}(\theta - \widehat{\theta}) = 0$. Our negative-log-likelihood convention fixes the signs of both the score and its derivative. Testing evaluates the score at $\theta_0$; the one-step correction evaluates it at $\widehat{\theta}$ to construct an interval center. Both formulas require $\widehat{J} > 0$. A nonpositive value does not support the normal calibration below.

**Proposition 11.5: Inference from a decorrelated score**

Let $S_n^\star = S_n(\beta^\star; w^\star)$ denote the score that would be available if the true nuisance parameters and direction were known. Assume $J$ is bounded above and away from zero, $\widehat{J}/J \to_p 1$, and

$$\frac{\sqrt{n}\, S_n^\star}{\sqrt{J}} \Rightarrow N(0,1).$$

For testing, suppose

$$\sqrt{n}\{\widehat{S}_n(\theta^\star) - S_n^\star\} = o_p(1).$$

Then under $H_0$, $T_n \Rightarrow N(0,1)$, giving the two-sided asymptotic p-value $2\{1 - \Phi(|T_n|)\}$. For the one-step interval, suppose instead that

$$\sqrt{n}\{\widehat{S}_n(\widehat{\theta}) - S_n^\star - \widehat{J}(\widehat{\theta} - \theta^\star)\} = o_p(1).$$

Then $\sqrt{n\widehat{J}}(\widetilde{\theta} - \theta^\star) \Rightarrow N(0,1)$, and

$$\widetilde{\theta} \ \pm \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{J}}}$$

has limiting coverage $1 - \alpha$.

The two $o_p(1)$ assumptions in Proposition 11.5 control different score approximation errors. For testing, the fitted score at $\theta^\star$ must approximate $S_n^\star$. For the one-step interval, the fitted score at $\widehat{\theta}$ must approximate $S_n^\star + \widehat{J}(\widehat{\theta} - \theta^\star)$. In each case, the difference must be negligible on the $n^{-1/2}$ scale. Orthogonality alone does not imply these remainder bounds. Appendix C proves the proposition and gives explicit conditions that imply the two expansions. The proof separates two familiar products: the error allowed by the constraint in (21) times the nuisance $\ell_1$ error, and the direction's $\ell_1$ error times the largest absolute nuisance-score component. A nonlinear loss adds a Taylor remainder. Each term must vanish after multiplication by $\sqrt{n}$, just as in (10). Neither an accurate prediction fit nor sparsity alone verifies all these requirements.

**Worked example: one logistic-regression coefficient.** Suppose $Y_i \in \{0,1\}$, and let $x_i$ be the observed covariate vector from an i.i.d. sample of response–covariate pairs. Fix a coordinate $j$, write $\theta = \beta_j$, and collect the other coefficients in $\gamma = \beta_{-j}$. The model specifies

$$\eta_i(\theta,\gamma) = x_{ij}\theta + x_{i,-j}^\top \gamma, \qquad \pi_i(\theta,\gamma) = \frac{\exp\{\eta_i(\theta,\gamma)\}}{1 + \exp\{\eta_i(\theta,\gamma)\}},$$

where $\pi_i$ is the probability that $Y_i = 1$ at covariate $x_i$. The average negative log-likelihood is

$$\ell_n(\theta,\gamma) = \frac{1}{n}\sum_{i=1}^n [\log\{1 + \exp(\eta_i)\} - Y_i\eta_i].$$

In the $(\theta,\gamma)$ coordinate order, put $q_i = (x_{ij}, x_{i,-j}^\top)^\top$. Differentiation gives $\nabla\ell_n = n^{-1}\sum_i(\pi_i - Y_i)q_i$ and $\nabla^2\ell_n = n^{-1}\sum_i \pi_i(1-\pi_i)q_i q_i^\top$. Thus the ordinary Gram matrix is replaced by a weighted Gram matrix.

Fit the penalized logistic model once, let $\widehat{\pi}_i = \pi_i(\widehat{\theta},\widehat{\gamma})$, and set $\widehat{a}_i = \widehat{\pi}_i(1-\widehat{\pi}_i)$. The constraint for $\widehat{w}$ becomes

$$\left\| \frac{1}{n}\sum_{i=1}^n \widehat{a}_i x_{i,-j}(x_{ij} - x_{i,-j}^\top w) \right\|_\infty \le \lambda_w.$$

It asks the residual of the target covariate to have small weighted correlations with all the other covariates. These weights arise from likelihood curvature; they are not additional observation

weights chosen by the analyst. The fitted score at a proposed coefficient $\theta$ is

$$\widehat{S}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\pi_i(\theta,\widehat{\gamma}) - Y_i\}(x_{ij} - x_{i,-j}^\top \widehat{w}). \tag{24}$$

The first factor is the fitted probability minus the outcome, the negative of the usual outcome residual; the second is a covariate residual. At $\theta = \theta_0$, the outcome probabilities use the proposed null value, while the weights used to estimate $\widehat{w}$ came from the full initial fit. This distinction matters when implementing the test.

For both the test and the one-step update, use

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i x_{ij}(x_{ij} - x_{i,-j}^\top \widehat{w}).$$

At the population level, exact weighted orthogonality makes partial information equal to the expected weighted squared covariate residual. Their empirical versions need not be identical, because the fitted orthogonality is only approximate. The displayed cross-product is the curvature used in (23). The asymptotic guarantees still require accurate estimation of both $\gamma^\star$ and $w^\star$, enough remaining information about the target, and control of the nonlinear remainder. Section 4.2 of Ning and Liu [141] verifies these requirements under stated logistic-model conditions.

**Connection to debiased Lasso and the scope of the result.** For $\ell_n(\beta) = \|y - X\beta\|_2^2/(2n)$, the Hessian is $X^\top X/n$, so weighted residualization reduces to ordinary residualization. If $\widehat{w}$ is chosen by the nodewise Lasso, the one-step correction becomes exactly the corresponding coordinate of (7); Exercise 11.7 asks for the algebra. This statement concerns the estimator: with unscaled squared loss and error variance $\sigma^2$, the score variance is $\sigma^2 J$, so the SE must also include $\sigma$.

The information-based SE above uses correct likelihood specification. With a general loss, curvature and score variance can differ; with a misspecified likelihood, the target also changes to a minimizer of population risk. Appendix C explains the resulting sandwich variance and its additional requirements. Finally, the proposition concerns a prespecified scalar target. Uniform inference over a growing family and inference after selecting a coefficient require further arguments, as the next section explains.

## 9. Many Coordinates and Multiple Testing

**Route: Advanced**

We now return to the linear model and its notation. For a set of pre-specified coordinates $J$, the debiased estimator produces z-statistics

$$Z_j^{\mathrm{db}} = \frac{\sqrt{n}\,\widehat{b}_j}{\widehat{\sigma}\sqrt{m_j^\top \widehat{\Sigma} m_j}}, \qquad j \in J,$$

for testing $H_{0j} : \beta_j = 0$. Two-sided p-values can be formed as

$$p_j = 2\{1 - \Phi(|Z_j^{\mathrm{db}}|)\}.$$

If only a few coordinates are tested, coordinatewise normality may be enough. If many coordinates are tested, the approximation must be uniform over the tested set, and the dependence among the debiased scores matters for FWER or FDR guarantees.

For an index set $J$ whose size grows with $n$, a precise uniform Gaussian approximation requires tail conditions, anti-concentration, covariance estimation control, and logarithmic factors that depend on the theorem being used. As a useful scaling heuristic, one needs two

strengthenings of Theorem 11.3. First, the remainder bound (10) must be controlled uniformly: $\max_{j\in J}\|m_j^\top\widehat{\Sigma}-e_j^\top\|_\infty = O_p(\sqrt{(\log p)/n})$, and the sparsity-product condition is pushed toward $s(\log p)\sqrt{\log|J|}/\sqrt{n}\to 0$, up to theorem-specific log factors. Second, the leading score vector $(m_j^\top X^\top\varepsilon/\sqrt{n})_{j\in J}$ must satisfy a high-dimensional Gaussian approximation, valid simultaneously over $J$; standard results in this direction extend the high-dimensional CLT of Chernozhukov et al. [42] to debiased Lasso statistics.

Once uniform approximation holds, Bonferroni adjustments require only the marginal validity of the p-values, paid for by a $\sqrt{\log|J|}$ factor in the critical value. BH-type FDR analyses require stronger control of the joint behavior or PRDS-style dependence assumptions, as in Chapter 6.

**Bootstrap-assisted simultaneous inference.** Bonferroni can be conservative because it ignores the dependence among the debiased z-statistics, which can be substantial when the design has correlated columns. Zhang and Cheng [203] develop a bootstrap-assisted procedure that accounts for this dependence. We describe a residual-based version of this approach. Its starting point is the leading random term

$$S_j=\frac{1}{\sqrt{n}}\sum_{i=1}^n(m_j^\top x_i)\varepsilon_i,\qquad \sqrt{n}\,(\widehat{b}_j-\beta_j)=S_j+o_p(1),\quad j\in J,$$

with the remainder controlled uniformly over $J$ as discussed above. The errors $\varepsilon_i$ are unobserved because the true regression coefficients are unknown. The outcome Lasso fit supplies residuals $\widehat{\varepsilon}_i=y_i-x_i^\top\widehat{\beta}$. When the fitted means are sufficiently accurate, these residuals provide information about the noise variation that drives the estimation errors.

For each bootstrap draw, hold the design, fitted directions $m_j$, and residuals fixed. Generate new independent multipliers $\xi_1,\dots,\xi_n\sim N(0,1)$, independently of the data, and form

$$S_j^*=\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\,(m_j^\top x_i)\,\widehat{\varepsilon}_i,\qquad j\in J.$$

The residuals supply an estimated noise contribution for each observation; the multipliers generate new random fluctuations of those contributions. Use the same multiplier $\xi_i$ for observation $i$ in every coordinate $j$. This shared multiplier retains the estimated dependence between coordinates. In particular,

$$\mathrm{Cov}^*(S_j^*,S_k^*)=\frac{1}{n}\sum_{i=1}^n(m_j^\top x_i)(m_k^\top x_i)\widehat{\varepsilon}_i^2,$$

where $\mathrm{Cov}^*$ denotes covariance over the multipliers, with the observed data held fixed. This is a residual estimate of the covariance of the leading scores; when $j=k$, it is the sandwich variance estimate in (12). Repeated draws reuse the same residuals and directions, so the Lasso need not be refitted.

Let $c^*_{1-\alpha}$ be the conditional $(1-\alpha)$-quantile of $\max_{j\in J}|S_j^*|$, approximated by repeated multiplier draws. The resulting simultaneous intervals are

$$\widehat{b}_j\ \pm\ \frac{c^*_{1-\alpha}}{\sqrt{n}},\qquad j\in J.$$

Under the conditions below, these cover all the target coefficients with asymptotic probability $1-\alpha$. Under constant noise variance, one may instead use $\widehat{\sigma}\xi_i$ in place of $\widehat{\varepsilon}_i\xi_i$, as in the homoskedastic construction of Zhang and Cheng [203].

The simultaneous guarantee requires the uniform remainder bounds discussed above, suitable tail and variance bounds for the scores, and uniform accuracy of the estimated score covariance matrix. For the residual-based version, the residuals must approximate this whole covariance matrix accurately enough, not just its diagonal entries. The allowed growth of $\log|J|$ must satisfy

the corresponding high-dimensional bootstrap theorem; coordinatewise asymptotic normality alone is insufficient. The simultaneous intervals can be narrower than Bonferroni intervals when the bootstrap captures substantial dependence among the coordinates. The resulting intervals also support max-statistic tests of composite hypotheses such as $H_0 : \beta_j = 0$ for all $j \in J$ and step-down procedures for FWER control on the full coordinate set.

It is important not to reinterpret these p-values as post-selection p-values for the variables chosen by the Lasso. If the question is "among the variables selected by an algorithm, which selected coefficients are significant in the selected model?", the target and conditioning event are different; Chapter 12 treats that problem.

## 10. Assumptions in Plain Language

**Route: Core**

The assumptions behind debiased Lasso are structural. Sparsity of $\beta$ keeps the Lasso estimation error small enough that a bias correction can remove it. Compatibility or restricted-eigenvalue conditions ensure that sparse signals are visible through the design. Sparsity of the relevant precision rows, or a successful projection program, makes approximate orthogonalization feasible. Noise assumptions justify the normal approximation and variance formula.

The strongest-looking condition,

$$s \log p / \sqrt{n} \to 0,$$

has a clear interpretation: after multiplying by $\sqrt{n}$, the product of the Lasso $\ell_1$ error and the approximate-inverse error must vanish. This makes the remaining regularization bias negligible relative to the SE.

For the extension in Section 8, both the nuisance coefficients and the score projection must be estimated accurately. The remaining curvature must be large enough to identify the target, and the nonlinear Taylor error must be small enough for the normal approximation. The SE also depends on whether likelihood curvature equals score variance. These requirements replace any claim that the linear-model sparsity condition alone guarantees inference for every loss.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | For each reported coefficient, compute $\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty$ and inspect the fitted score contributions for observations that dominate the sum. If unequal error variances are plausible, compare the constant-variance and sandwich SEs. |
| **What these data cannot show** | These checks cannot confirm that only a few regression coefficients are nonzero, or that the relevant rows of the inverse covariance matrix have few nonzero entries. They also cannot guarantee that the bias left after correction is small enough for valid intervals, or that $\beta_j$ answers the scientific question. |
| **Sensitivity checks** | Choose a few tuning settings in advance for the Lasso fit and the approximate inverse. Repeat the analysis with these settings, and report how much the confidence intervals and conclusions change. |

## 11. Failure Modes and Diagnostics

**Route: Core**

Debiased Lasso can fail for several reasons.

- If $\beta$ is too dense, the Lasso $\ell_1$ error is too large and the bias remainder in (10) is not negligible.
- If the design is highly collinear, the approximate inverse can have large variance or fail to satisfy the required $\ell_\infty$ approximation.
- If the precision rows are dense, nodewise Lasso may not estimate $\Gamma$ accurately even when $\beta$ itself is sparse.
- If the noise variance is poorly estimated, the interval width in (14) is wrong even when the center is asymptotically normal.
- If the linear model is misspecified, the target may become a projection coefficient rather than a structural coefficient; the variance formula may also need heteroscedasticity-robust modification.

In practice, one should examine the stability of intervals over reasonable tuning choices, the size of estimated SEs, the empirical correlations $m_j^\top \widehat{\Sigma} - e_j^\top$, and the plausibility of sparsity for both $\beta$ and the relevant precision rows.

**Tuning and diagnostics.** Choose a few penalty values around those used for prediction, and examine how the estimates and SEs change. Figure 11.6 illustrates four quantities for the first coefficient in a model with more features than observations. For a target coordinate $j$, write $a_{ij} = x_i^\top m_j$ for observation $i$'s score weight and $\widehat{\varepsilon}_i = y_i - x_i^\top \widehat{\beta}$ for its fitted residual.

(1) *Approximate-inverse error* measures how closely the score direction has the desired inner products with the design columns:

$$A_j = \|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = \max_k \left| \frac{d_j^\top X_k}{n} - \mathbf{1}\{k = j\} \right|, \qquad d_j = X m_j.$$

The desired value is one for the target column and zero for every other column. Thus $A_j$ is the largest discrepancy from these values. It is the factor multiplying the Lasso error in (10).

(2) *Normalized observation weights* show how the score direction is distributed across observations:

$$w_{ij} = \frac{|a_{ij}|}{\{\sum_{\ell=1}^n a_{\ell j}^2\}^{1/2}} = \frac{|x_i^\top m_j|}{\|X m_j\|_2}.$$

Here $a_{ij} = x_i^\top m_j$ is the $i$th component of the direction $d_j = X m_j$. The normalized weights $w_{ij}$ have squares that sum to one. Under independent errors with constant variance, $w_{ij}^2$ is observation $i$'s share of the variance of the leading noise term. The panel plots one weight per observation, making it easy to see whether a few observations receive much larger weights than the others.

(3) *Homoskedastic and robust SEs* use, respectively, a common noise variance and observation-specific squared residuals:

$$\widehat{\mathrm{se}}_{\mathrm{hom},j} = \frac{\widehat{\sigma}}{n} \sqrt{\sum_{i=1}^n a_{ij}^2}, \qquad \widehat{\mathrm{se}}_{\mathrm{rob},j} = \frac{1}{n} \sqrt{\sum_{i=1}^n a_{ij}^2 \widehat{\varepsilon}_i^2} = \frac{\widehat{v}_j}{\sqrt{n}}.$$

In this figure, $\widehat{\sigma}^2 = n^{-1} \sum_i \widehat{\varepsilon}_i^2$. The two curves show how the estimated uncertainty changes with both the variance formula and the nodewise penalty.

(4) *Sensitivity to the nodewise penalty* is shown by plotting the debiased estimate itself as $\lambda_j$ varies. Write $m_j(\lambda_j)$ for the direction obtained from nodewise Lasso at penalty $\lambda_j$, including the scaling in (17), and define its observation weights by $a_{ij}(\lambda_j) = x_i^\top m_j(\lambda_j)$. Then

$$\widehat{b}_j(\lambda_j) = \widehat{\beta}_j + \frac{1}{n}\sum_{i=1}^{n} a_{ij}(\lambda_j)\widehat{\varepsilon}_i.$$

The outcome Lasso fit and its residuals are held fixed; changing $\lambda_j$ changes $m_j(\lambda_j)$, and hence the score weights and correction. Compare the movement of the estimate with the SEs to assess how much the interval endpoints change.

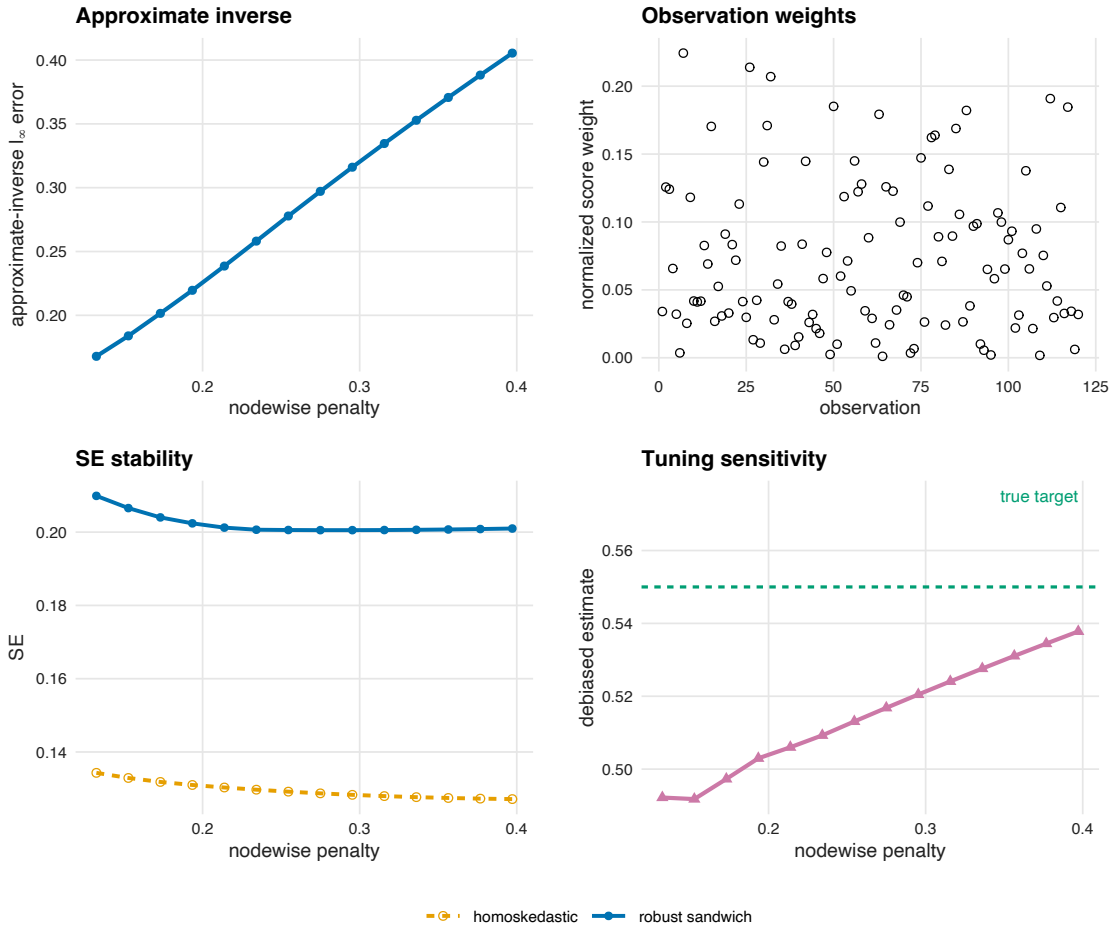



FIGURE 11.6. Four diagnostics for $\beta_1$ from one simulated data set with $n = 120$, $p = 180$, AR(1) correlation $\rho = 0.45$, eight nonzero coefficients, and covariate-dependent noise. The panels show $A_1$ (upper left), the weights $w_{i1}$ (upper right), the two SEs (lower left), and $\widehat{b}_1(\lambda_1)$ (lower right), as defined in the text. The outcome Lasso fit is held fixed. The weights use $\lambda_1 \approx 0.234$; the other panels vary $\lambda_1$. The horizontal line in the lower-right panel marks the true coefficient $\beta_1 = 0.55$.

## 12. Limits of Adaptivity

**Route: Research frontier**

Suppose we want a confidence interval for $\beta_1$. The other coefficients $\beta_2, \dots, \beta_p$ are called *nuisance coefficients*: they affect the response, although $\beta_1$ is our target. If we knew in advance which of them could be nonzero, we could regress on those columns together with $X_1$. For a fixed number of columns whose normalized Gram matrix has eigenvalues bounded away from zero and infinity, OLS gives interval length of order $n^{-1/2}$. This *known-support* benchmark assumes that the identities of the relevant variables are supplied.

In a high-dimensional problem, we may instead know only that at most $s$ coefficients are nonzero. To make the comparison precise, consider the Gaussian model studied by Cai and Guo [34]:

$$x_i \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma), \qquad \varepsilon \sim N_n(0, \sigma^2 I_n), \qquad \varepsilon \text{ independent of } X.$$

For fixed constants $K > 1$ and $B > 0$, define

$$\Theta(s) = \left\{ (\beta, \Sigma, \sigma) : \|\beta\|_0 \le s, \quad K^{-1} I_p \preceq \Sigma \preceq K I_p, \quad 0 < \sigma \le B \right\}.$$

The matrix bounds mean that every eigenvalue of $\Sigma$ lies between $K^{-1}$ and $K$. Both $\Sigma$ and $\sigma$ are unknown. This class does not require sparse rows of $\Sigma^{-1}$. The positions and values of the nonzero coefficients are also unknown. For example, among $p = 10{,}000$ features, $\Theta(3)$ allows any three coefficients to be nonzero, while $\Theta(20)$ also allows models with four through twenty nonzero coefficients. Thus $\Theta(3) \subset \Theta(20)$.

An interval $C_j$ is called *honest over* $\Theta(s)$ if it has coverage at least $1 - \alpha$ for every model in that class:

$$\inf_{\vartheta \in \Theta(s)} \mathbb{P}_\vartheta\{\beta_j \in C_j\} \ge 1 - \alpha.$$

Here $\vartheta$ includes the regression coefficients and the design and noise parameters. At the 95 percent level, the same interval procedure must cover at least 95 percent of the time under each allowed model. The target coordinate $j$ remains fixed. This is an exact coverage requirement at each sample size. For asymptotic results, the corresponding requirement is

$$\liminf_{n \to \infty} \inf_{\vartheta \in \Theta(s)} \mathbb{P}_\vartheta\{\beta_j \in C_j\} \ge 1 - \alpha,$$

where $p$ and $s$ may grow with $n$. It allows an approximation error that vanishes uniformly over the class as the sample size increases. We use this asymptotic version when discussing attainable rates below.

*Adaptivity* asks whether the data can determine how short the interval should be. Could one procedure retain coverage over $\Theta(20)$, yet give intervals as short as those available when the smaller bound of three nonzero coefficients is supplied in advance? For $0 < \alpha < 1/2$, the results of Cai and Guo [34] apply in the sparsity range

$$1 \le s \le S \le c \min\{p^\kappa, n/\log p\}, \qquad 0 < \kappa < 1/2,$$

with fixed $\kappa$ and a sufficiently small fixed $c > 0$. Their Theorem 1 and equation (3.14) give the smallest achievable worst-case expected length over $\Theta(s)$, up to constant factors:

$$r(s) = \frac{1}{\sqrt{n}} + \frac{s \log p}{n}.$$

Here *worst-case expected length* means $\sup_{\vartheta \in \Theta(s)} E_\vartheta |C_j|$, the largest average interval length among models in the class.

Theorem 2 of Cai and Guo [34] shows what changes when coverage is required over the larger class $\Theta(S)$. For any interval satisfying that requirement, the worst-case expected length even on the smaller class obeys

$$\sup_{\vartheta \in \Theta(s)} E_\vartheta |C_j| \gtrsim \frac{1}{\sqrt{n}} + \frac{S \log p}{n} = r(S).$$

The bound concerns the largest expected length on $\Theta(s)$; it does not assert the same lower bound at every model in that class. When $r(s) = o\{r(S)\}$, one procedure therefore cannot both cover the larger class and achieve the smaller class's shorter rate throughout $\Theta(s)$. Small additional coefficients can be difficult to distinguish from zeros while still affecting the bias correction. If $s$ and $S$ differ only by a fixed factor, their rates can have the same order. When $S \log p \lesssim \sqrt{n}$, both rates have order $n^{-1/2}$, and adaptation at that rate is possible.

## 13. Bibliographic Notes

**Route: Core**

The Lasso was introduced by Tibshirani [176]. Fixed-dimensional asymptotics for Lasso-type estimators were developed by Knight and Fu [106]. The standard textbook reference for Lasso theory, compatibility/restricted-eigenvalue conditions, and high-dimensional regression foundations is Bühlmann and van de Geer [33]. Debiased and desparsified Lasso methods for high-dimensional confidence intervals and tests were developed in Zhang and Zhang [199], Javanmard and Montanari [97], and van de Geer et al. [180]. The nodewise Lasso construction is closely tied to Gaussian graphical model estimation by Meinshausen and Bühlmann [136]. Dezeure et al. [47] provide a broad account of high-dimensional inference methods and software, while Cai and Guo [34] give minimax and adaptivity limits for high-dimensional confidence intervals. Uniform Gaussian approximations of debiased Lasso statistics over many coordinates extend the maximum-of-sums theory of Chernozhukov et al. [42]. The bootstrap-assisted simultaneous-inference procedure introduced in Section 9 is from Zhang and Cheng [203], who prove its validity for an index set whose size may grow exponentially with $n$ and demonstrate substantial efficiency gains over Bonferroni when the design has correlated columns.

Ning and Liu [141] develop a general decorrelated-score framework for penalized high-dimensional models. Their construction connects tests of a low-dimensional target with one-step confidence intervals and extends beyond the Lasso penalty and least-squares loss. Section 8 uses the likelihood version and their logistic example; Appendix C gives an explicit remainder argument and explains the distinction between curvature and variance. Their paper also treats additional models, efficiency, and uniform guarantees under corresponding assumptions.

A practically important direction not pursued in detail here is the *streaming* version of debiased Lasso. In modern high-dimensional generalized linear models (GLMs) with online or batch-arriving observations, the storage and computational demands of forming the full $p \times p$ precision-matrix surrogate become prohibitive. Han et al. [79] combine stochastic gradient updates with one-pass debiasing and establish asymptotic normality under their model and sparsity assumptions. Their algorithm uses $O(p)$ memory for inference on a prespecified coordinate. An alternative is to seek confidence sequences (Chapter 8) for the same target. Such a construction needs bias and variance control over all monitored times; the fixed-sample debiasing theorem does not itself justify optional stopping.

## 14. Exercises

**Route: Core**

### Intermediate.

**Exercise 11.6: Precision rows from regressions**

Let $X^\circ \sim N(0, \Sigma)$ and $\Gamma = \Sigma^{-1}$. Use the block inverse formula to prove (15). Interpret the zeros of $\Gamma_{j,-j}$ in terms of conditional linear regression.

**Exercise 11.7: Decorrelated scores recover linear debiasing**

Use the loss $\ell_n(\beta) = \|y - X\beta\|_2^2/(2n)$, fix a coordinate $j$, and write $\theta = \beta_j$, $\gamma = \beta_{-j}$. Take $(\widehat{\theta}, \widehat{\gamma}) = (\widehat{\beta}_j, \widehat{\beta}_{-j})$ from the Lasso in (6). For an estimated direction $\widehat{w}$, put $\widehat{r}_j = X_j - X_{-j}\widehat{w}$. Derive the fitted decorrelated score and its derivative $\widehat{J}$. Assuming $\widehat{r}_j^\top X_j \neq 0$, show that the one-step estimator is

$$\widetilde{\theta} = \widehat{\beta}_j + \frac{\widehat{r}_j^\top (y - X\widehat{\beta})}{\widehat{r}_j^\top X_j}.$$

Now choose $\widehat{w} = \widehat{\gamma}_j$ from the nodewise Lasso in (16). Use its KKT identity to show that this equals the $j$th coordinate of (7) with $M = \widehat{\Gamma}$.

### Computational.

**Exercise 11.8: Small debiasing simulation**

Use $n = 150$, $p = 80$, and a known noise standard deviation $\sigma = 1$. For each of the four combinations $s \in \{5, 12\}$ and $\rho \in \{0.1, 0.55\}$, generate 500 independent data sets. Draw the design rows independently from $N_p(0, \Sigma)$, where $\Sigma_{kl} = \rho^{|k-l|}$, then center each column and scale it to have squared length $n$. Generate errors $\varepsilon \sim N_n(0, I_n)$ independently of $X$, and set $y = X\beta + \varepsilon$, with $\beta_1 = 0.55$, $\beta_2 = \cdots = \beta_s = 0.45$, and all remaining coefficients zero. Use the penalties

$$\lambda = 0.95\sqrt{\frac{2\log p}{n}}, \qquad \lambda_1 = 0.65\sqrt{\frac{2\log p}{n}}$$

in (6) and (16), respectively. These are the signal, correlation, and penalty choices used in Figure 11.4, with a different sample size and number of predictors.

Let $\widehat{r}_1 = X_1 - X_{-1}\widehat{\gamma}_1$ be the nodewise residual and compute

$$\widehat{b}_1 = \widehat{\beta}_1 + \frac{\widehat{r}_1^\top (y - X\widehat{\beta})}{\widehat{r}_1^\top X_1}.$$

For the oracle fit, use the true support $S = \{1, \ldots, s\}$ and put $\widehat{\beta}_S^{\mathrm{or}} = (X_S^\top X_S)^{-1} X_S^\top y$. Compare the following nominal 95 percent intervals for $\beta_1$:

$$I_{\mathrm{naive}} = \widehat{\beta}_1 \ \pm \ z_{0.975}\frac{\sigma}{\sqrt{n}},$$

$$I_{\mathrm{db}} = \widehat{b}_1 \ \pm \ z_{0.975}\frac{\sigma\|\widehat{r}_1\|_2}{|\widehat{r}_1^\top X_1|},$$

$$I_{\mathrm{or}} = \widehat{\beta}_1^{\mathrm{or}} \ \pm \ z_{0.975}\sigma\sqrt{\{(X_S^\top X_S)^{-1}\}_{11}}.$$

For each setting and method, report the fraction of intervals that cover $\beta_1$, its Monte Carlo SE, and the average interval length. For empirical coverage $\widehat{c}$, use $\{\widehat{c}(1-\widehat{c})/500\}^{1/2}$ as the Monte Carlo SE.

### Exercise 11.9: Nodewise Lasso diagnostics

Use one data set from Exercise 11.8 with $s = 5$ and $\rho = 0.55$, and identify its replicate and random seed. Compute the nodewise rows for $j \in \{1, 2, 10\}$, using $\lambda_j = 0.65\sqrt{2 \log p/n}$. For each coordinate, report the approximate-inverse error and the variance factor

$$\|\widehat{\Gamma}_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \quad \text{and} \quad \widehat{\Gamma}_j^\top \widehat{\Sigma} \widehat{\Gamma}_j.$$

Also report $\sigma^2 \widehat{\Gamma}_j^\top \widehat{\Sigma} \widehat{\Gamma}_j / n$, the conditional variance of the estimator's leading noise term.

### Exercise 11.10: Javanmard–Montanari row program

Generate a small design with $p > n$, so that $\widehat{\Sigma}$ is singular. Choose a target coordinate $j$ before generating the data. Solve the Javanmard–Montanari program (19) for a decreasing grid of tolerances $\eta$, including $\eta = 1$ and $\eta = 0$. Report the sample size, number of predictors, design distribution, column scaling, target coordinate, and tolerance grid. For each feasible solution, tabulate the variance factor $m^\top \widehat{\Sigma} m$ and approximation error $\|m^\top \widehat{\Sigma} - e_j^\top\|_\infty$. Record which tolerances are infeasible and distinguish them from any numerical failures reported by the solver.

CHAPTER 12

# Selective Inference and False Coverage Rate

Classical confidence intervals are usually derived for questions fixed before the data are inspected. Modern analyses often reverse the order. The data are collected, interesting variables or clusters are found, and then intervals or p-values are reported for the same discoveries. This changes both the reference distribution and sometimes the target parameter itself.

Exploration often suggests scientific questions. Once we select which intervals to report using the same data, however, their coverage can fall below the advertised level. Sorić summarized this point in the language of "interesting" confidence intervals: coverage over all intervals does not transfer automatically to coverage over the subset selected for attention [167].

This chapter covers two broad responses. The first controls an average error rate over selected intervals, the FCR. The second conditions on the selection event and changes the reference law accordingly. Post-selection inference (POSI) gives a third, deliberately broad, option: it protects against arbitrary model selection by using simultaneous intervals over all possible selected-model targets; the full construction appears in Section 6.

## 1. A Selected-Interval Problem

**Route: Core**

Let $\theta_1, \dots, \theta_m$ be parameters and let $C_i$ be a marginal $1-\alpha$ confidence interval for $\theta_i$:

$$\mathbb{P}_\theta(\theta_i \in C_i) \geq 1-\alpha \qquad \text{for each fixed } i.$$

Let $\widehat{\mathcal{S}} \subseteq \{1, \dots, m\}$ be a data-dependent selected set. The selected intervals are

$$\{C_i : i \in \widehat{\mathcal{S}}\}.$$

The marginal statement above does not imply

$$\mathbb{P}_\theta(\theta_i \in C_i \mid i \in \widehat{\mathcal{S}}) \geq 1-\alpha.$$

Selection typically favors large noisy estimates, and large noisy estimates are exactly the estimates whose naive intervals are most likely to have missed their targets.

Figure 12.1 shows the phenomenon in the Gaussian sequence model. Draw

$$\theta_i \overset{\text{i.i.d.}}{\sim} N(0, 0.2^2), \qquad Z_i \mid \theta_i \sim N(\theta_i, 1), \qquad i = 1, \dots, 20,$$

and form 90 percent marginal intervals $Z_i \pm z_{0.95}$. If we report only intervals excluding zero, the selected intervals are biased toward extreme observations.

The same phenomenon appears in regression. If a model is chosen by Akaike information criterion (AIC), Bayesian information criterion (BIC), stepwise search, looking at residual plots, or using the same data to screen variables, the usual $t$-statistic from the final model is no longer distributed as though the model had been fixed in advance. The distortion can be severe when many candidate variables were available.

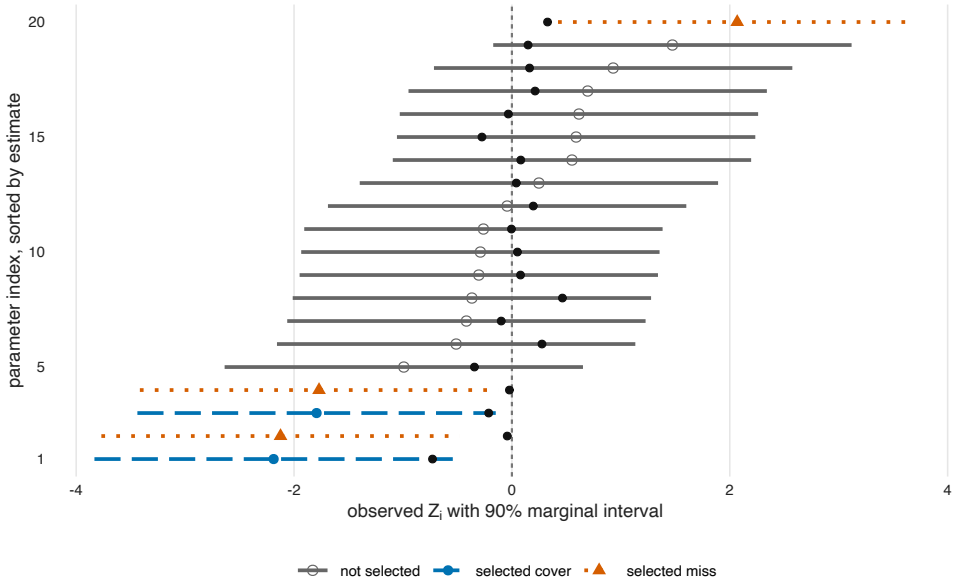



FIGURE 12.1. Marginal intervals after selection. Each horizontal segment is a 90 percent marginal interval; black dots mark the true $\theta_i$'s. Selecting intervals that exclude zero concentrates attention on noisy extremes, so selected intervals can miss far more often than their marginal confidence level suggests.

## 2. Targets After Selection

**Route: Core**

The first question in selective inference is: what parameter is the interval supposed to cover?

For a fixed family of parameters $\theta_1, \dots, \theta_m$, the selected set $\widehat{\mathcal{S}}$ changes only which intervals are reported. This is the setting of FCR control. The targets remain $\theta_i$, but the error criterion is evaluated over selected indices.

In regression, selection can change the estimand. Suppose $X \in \mathbb{R}^{n \times p}$ is fixed and

$$y \sim N(\mu, \sigma^2 I_n).$$

For a selected model $M \subseteq \{1, \dots, p\}$, the selected-model target is

$$\beta_M = (X_M^\top X_M)^{-1} X_M^\top \mu, \tag{25}$$

assuming $X_M^\top X_M$ is invertible. The coordinate $\beta_{j \bullet M}$ is the least-squares projection coefficient of the fixed mean vector $\mu$ onto the columns selected in $M$. This target differs from the full-model coefficient in general. Chapter 11 focused on pre-specified coordinates of a high-dimensional coefficient vector; this chapter focuses on targets created or chosen after selection.

## 3. Conditional Coverage and Its Limits

**Route: Core**

One natural goal is conditional coverage:

$$\mathbb{P}_\theta(\theta_i \in C_i \mid i \in \widehat{\mathcal{S}}) \geq 1 - \alpha. \tag{26}$$

This is attractive but often impossible without changing the interval or the conditioning rule. Consider

$$Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$$

and the usual 95 percent intervals $Y_i \pm 1.96$. Select intervals that exclude zero:

$$\widehat{\mathcal{S}} = \{i : 0 \notin [Y_i - 1.96, Y_i + 1.96]\}.$$

If $\mu = 0$, every selected interval excludes the true value. Thus

$$\mathbb{P}_0(\mu \in C_i \mid i \in \widehat{\mathcal{S}}) = 0$$

whenever the conditioning event has positive probability. Bonferroni widening can control the probability of at least one error, but it does not solve this conditional problem when the selection rule is itself defined by excluding the null value.

This example is the interval analogue of conditioning on making a rejection under the global null. Conditional error rates can become degenerate because the conditioning event may select precisely the cases in which the nominal procedure failed.

## 4. False Coverage Rate

**Route: Core**

Following Benjamini and Yekutieli [22], let

$$R_{\text{CI}} = |\widehat{\mathcal{S}}|$$

be the number of selected intervals, and let

$$V_{\text{CI}} = |\{i \in \widehat{\mathcal{S}} : \theta_i \notin C_i\}|$$

be the number of selected intervals that fail to cover. The FCR is

$$\text{FCR} = \mathbb{E}\left[\frac{V_{\text{CI}}}{R_{\text{CI}} \vee 1}\right]. \tag{27}$$

This is the confidence-interval analogue of the FDR. It does not promise that every selected interval has conditional coverage (26); it controls the expected fraction of noncovering intervals among the selected intervals.

Two simple procedures control FCR but can be conservative. If all $m$ intervals are reported, marginal $1 - \alpha$ intervals give

$$\text{FCR} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_\theta(\theta_i \notin C_i) \le \alpha.$$

If only selected intervals are reported, simultaneous coverage also suffices:

$$\mathbb{P}_\theta(\theta_i \in C_i \text{ for all } i = 1, \dots, m) \ge 1 - \alpha \quad \Longrightarrow \quad \text{FCR} \le \alpha.$$

Bonferroni intervals are an example. They protect against any selected interval failing, so they also protect the average selected fraction. The price is width.

Figure 12.2 separates three notions visually: naive selected intervals can fail near the selection boundary, Bonferroni protects every interval simultaneously, and the BY adjustment targets the average noncoverage fraction among those actually reported.

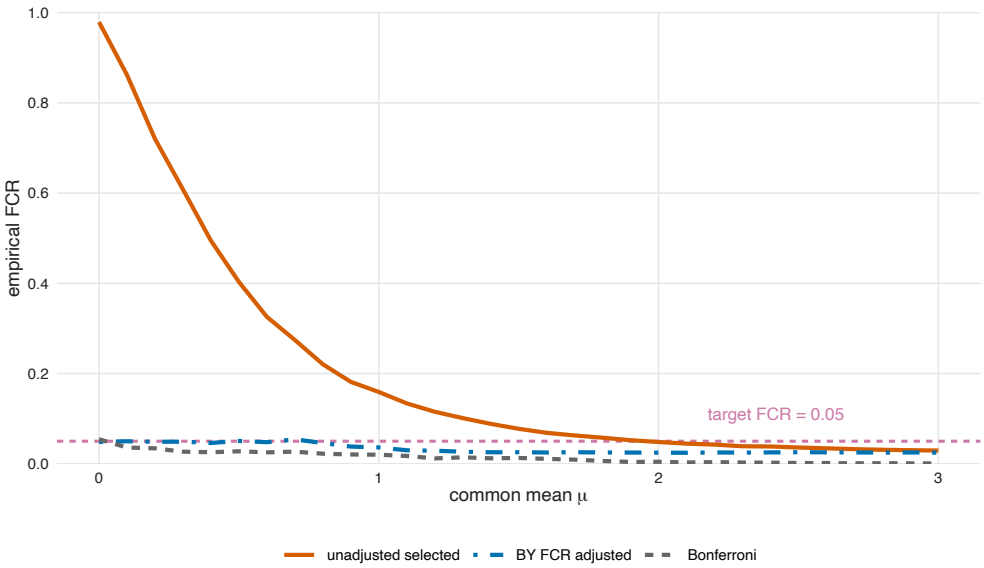



FIGURE 12.2. Empirical FCR in the Gaussian common-mean example with $m = 80$, $\alpha = 0.05$, based on 2,500 independent simulation runs at each $\mu$. Intervals selected for excluding zero can have very large FCR when the true mean is near zero. The BY FCR adjustment controls FCR while being less conservative than always using Bonferroni intervals.

## 5. The Benjamini–Yekutieli FCR Adjustment

**Route: Core**

The same paper proposes an FCR-controlling adjustment that adapts the interval width to the number of selected parameters. Let

$$T = (T_1, \dots, T_m)$$

be statistics used by an arbitrary selection rule $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(T)$. For each $i$, define

$$T^{(i:t)} = (T_1, \dots, T_{i-1}, t, T_{i+1}, \dots, T_m),$$

and

$$R^{(i)} = \min\big(\{|\widehat{\mathcal{S}}(T^{(i:t)})| : i \in \widehat{\mathcal{S}}(T^{(i:t)}),\ t \in \mathbb{R}\} \cup \{\infty\}\big). \tag{28}$$

Thus $R^{(i)}$ is the smallest number of selections that could occur while holding all other statistics fixed and varying $T_i$ in a way that still selects $i$. The set on the right of (28) is nonempty on the event $i \in \widehat{\mathcal{S}}(T)$, since the realized $t = T_i$ is feasible there; the value is $\infty$ only if no possible value of $T_i$ can select $i$. An unselected index can therefore still have finite $R^{(i)}$. The definition depends on $T_{-i}$ alone, while the reported interval uses it only when $i$ is selected. A selection rule is called *simple* if varying $T_i$ while keeping $i$ selected does not change the number of selections. For such rules, $R^{(i)} = R_{\mathrm{CI}}$ whenever $i$ is selected. The minimum in (28) also handles rules without this property.

For example, if $\widehat{\mathcal{S}} = \{i : |T_i| > c\}$ with fixed $c > 0$, then

$$R^{(i)} = 1 + \#\{j \neq i : |T_j| > c\}.$$

Varying $T_i$ can always select $i$, while the other decisions remain fixed. This formula holds even when the observed $|T_i| \le c$, so that $i$ is currently unselected.

Suppose $C_i(\gamma)$ denotes a marginal interval with noncoverage probability at most $\gamma$:

$$\mathbb{P}_\theta\{\theta_i \notin C_i(\gamma)\} \le \gamma.$$

For each $i \in \widehat{\mathcal{S}}$, report

$$C_i\left(\frac{\alpha R^{(i)}}{m}\right). \tag{29}$$

If many parameters are selected, the adjustment is mild. If only one parameter is selected, the interval becomes Bonferroni-like.

**Theorem 12.1: Benjamini–Yekutieli FCR control**

Assume $T_1, \dots, T_m$ are independent, and assume that for each $i$, $C_i(\gamma)$ is constructed from $T_i$ and prespecified quantities only (in particular, not from $T_{-i}$) and has marginal noncoverage at most $\gamma$. Then the selected intervals in (29) satisfy

$$\mathrm{FCR} \le \alpha.$$

**Proof of Theorem 12.1**

Set $\Delta_i = 0$ when $i \notin \widehat{\mathcal{S}}$; otherwise let

$$\Delta_i = \frac{\mathbf{1}\{\theta_i \notin C_i(\alpha R^{(i)}/m)\}}{R_{\mathrm{CI}}}.$$

Thus no interval is evaluated at an infinite level, and $\mathrm{FCR} = \sum_{i=1}^m \mathbb{E}[\Delta_i]$. On the event $\{i \in \widehat{\mathcal{S}}, R^{(i)} = k\}$, the definition of $R^{(i)}$ implies $R_{\mathrm{CI}} \ge k$. Therefore

$$\Delta_i \le \sum_{k=1}^m \frac{\mathbf{1}\{\theta_i \notin C_i(\alpha k/m),\ R^{(i)} = k\}}{k}.$$

Condition on the vector of all statistics except $T_i$. The value $R^{(i)}$ is a function of $T_{-i}$ alone, so it is fixed by this conditioning. The interval $C_i(\gamma)$ is a function of $T_i$ and prespecified ingredients, and under the marginal coverage assumption, $\mathbb{P}_\theta\{\theta_i \notin C_i(\gamma)\} \le \gamma$ when the law of $T_i$ is the assumed marginal. Because the $T_i$'s are independent, the conditional law of $T_i$ given $T_{-i}$ equals its marginal law, so

$$\mathbb{P}_\theta\{\theta_i \notin C_i(\alpha k/m) \,|\, T_{-i}\} \le \frac{\alpha k}{m} \quad \text{on} \quad \{R^{(i)} = k\}.$$

Hence

$$\mathbb{E}[\Delta_i \mid T_{-i}] \le \sum_{k=1}^m \frac{\mathbf{1}\{R^{(i)} = k\}}{k}\frac{\alpha k}{m} = \frac{\alpha}{m}\mathbf{1}\{R^{(i)} < \infty\} \le \frac{\alpha}{m}.$$

Taking expectations and summing over $i$ gives $\mathrm{FCR} \le \alpha$. □

Dependence extensions need additional structure. Benjamini and Yekutieli [22] obtain the same bound for PRDS statistics when selection and intervals are also *concordant*. Fix $i$, $k$, the interval level $\gamma$, and the true parameter $\theta_i$. The requirement concerns two sets:

$$\{t_{-i} : R^{(i)}(t_{-i}) \ge k\}, \qquad \{t_i : \theta_i \notin C_i(\gamma; t_i)\},$$

where $C_i(\gamma; t_i)$ denotes the interval computed when $T_i = t_i$. Both sets must be increasing, or both decreasing. In the increasing case, once $R^{(i)} \geq k$, raising any of the other statistics cannot make this inequality fail. Similarly, once the interval misses $\theta_i$, raising $t_i$ cannot restore coverage. In the decreasing case, both statements hold when the statistics are lowered instead. For $T_i \sim N(\theta_i, 1)$, selecting when $T_i > c$ and reporting $[T_i - z_{1-\gamma}, \infty)$ gives increasing sets: $R^{(i)} = 1 + \#\{j \neq i : T_j > c\}$, and noncoverage occurs when $T_i > \theta_i + z_{1-\gamma}$. A two-sided interval can miss in either tail, so its noncoverage set need not be monotone. Under arbitrary dependence, their harmonic correction applies to nested marginal intervals: $C_i(\gamma_1) \supseteq C_i(\gamma_2)$ whenever $\gamma_1 \leq \gamma_2$. For such intervals, replace $\alpha$ in (29) by $\alpha/H_m$, where $H_m = \sum_{k=1}^{m} 1/k$. These conditions explain why the independence proof cannot simply be reused for dependent statistics.

Controlling FCR bounds the expected fraction of reported intervals that miss their stated targets. It does not by itself correct selection bias in the estimates at their centers. For example, estimates selected for being unusually large are often large partly because of noise; widening intervals around those estimates does not correct their upward bias. A coverage guarantee also refers to a particular target. If model selection changes the regression coefficient being estimated, we must identify that new target and justify coverage for it. These issues motivate methods that adjust the estimates or explicitly account for the selected target.

## 6. POSI: Protection Against Arbitrary Selection

**Route: Advanced**

FCR concerns a fixed family of targets. POSI addresses a different problem: after looking at the data, an analyst may choose a regression model by an unknown or unreported procedure. Berk et al. [24] protect against this by constructing simultaneous intervals over all selected-model targets.

Let $\mathcal{M} = \{M \subseteq \{1, \ldots, p\} : M \neq \emptyset,\ X_M^\top X_M \text{ invertible}\}$ be the family of full-rank submodels. For every $M \in \mathcal{M}$ and $j \in M$, let

$$\eta_{j\bullet M}^\top = e_j^\top (X_M^\top X_M)^{-1} X_M^\top,$$

where $e_j \in \mathbb{R}^{|M|}$ is the local basis vector picking out $j$ within $M$, so that

$$\widehat{\beta}_{j\bullet M} = \eta_{j\bullet M}^\top y, \qquad \beta_{j\bullet M} = \eta_{j\bullet M}^\top \mu.$$

Define the standardized error

$$Z_{j\bullet M} = \frac{\eta_{j\bullet M}^\top (y - \mu)}{\sigma \|\eta_{j\bullet M}\|_2}.$$

The POSI constant $K_{1-\alpha}(X)$ is the $1-\alpha$ quantile of

$$\max_{M \in \mathcal{M}} \max_{j \in M} |Z_{j\bullet M}|.$$

Then the intervals

$$\widehat{\beta}_{j\bullet \widehat{M}} \pm K_{1-\alpha}(X)\, \sigma \|\eta_{j\bullet \widehat{M}}\|_2 \tag{30}$$

cover all selected-model coefficients simultaneously, no matter how $\widehat{M} \in \mathcal{M}$ was chosen. If $\sigma$ is unknown, a $t$-based version needs a variance estimate independent of the numerator contrasts, with a specified reference law. For example, if $\nu\widehat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$ independently of all the contrasts, calibrate the maximum of the resulting $t$-statistics using that same variance estimate. Full-model residual degrees of freedom are $\nu = n - \operatorname{rank}(X)$ only when the mean lies in the full model and $\nu > 0$; they are not automatically available when $p > n$.

**Theorem 12.2: POSI simultaneous protection**

Under $y \sim N(\mu, \sigma^2 I_n)$ with fixed design and $\sigma > 0$, if $K_{1-\alpha}(X)$ is chosen as above with known $\sigma$, then for every data-dependent model selection rule $\widehat{M} = \widehat{M}(y) \in \mathcal{M}$, the intervals $C_{j\bullet\widehat{M}}$ in (30) satisfy

$$\mathbb{P}\left\{\beta_{j\bullet\widehat{M}} \in C_{j\bullet\widehat{M}} \text{ for all } j \in \widehat{M}\right\} \geq 1 - \alpha.$$

**Proof of Theorem 12.2**

The event

$$\max_{M} \max_{j \in M} |Z_{j\bullet M}| \leq K_{1-\alpha}(X)$$

implies simultaneous coverage for every model and every coordinate in that model. It therefore implies coverage for the random subset chosen by any selection rule $\widehat{M}(y)$. The probability of the event is at least $1 - \alpha$ by the definition of $K_{1-\alpha}(X)$. □

Computing $K_{1-\alpha}(X)$ requires considering many model-coordinate pairs, but it can be approximated by Monte Carlo simulation: draw $y^{(b)} - \mu^{(b)} \sim N(0, \sigma^2 I_n)$ for $b = 1, \ldots, B$, compute the $Z_{j\bullet M}$ for every $M$ and $j \in M$, record the maximum absolute value, and take its empirical $1 - \alpha$ quantile. The constant can range from roughly

$$\sqrt{2 \log p} \quad \text{to} \quad \sqrt{p}$$

depending on the design and the selection geometry. Orthogonal designs are near the lower end; designs and selection strategies that hunt aggressively over model spaces can approach the upper end. POSI is attractive when the selection process is hard to formalize, but the intervals can be much wider than intervals tailored to a specific selection event.

Figure 12.3 illustrates how the desired protection affects interval width. Its dashed vertical line is the ordinary two-sided 95 percent normal cutoff, $z_{0.975} \approx 1.96$. The other endpoints are illustrative values, not computed constants or a universal ordering of widths. Actual widths depend on the design, the selection rule, and the target; a conditional interval can even be very long when the observed statistic is close to a selection boundary.

Sample splitting is a simpler alternative: use one part of the data to select questions and an independent part to perform inference. Its validity relies on an exchangeability or independence structure that justifies the split. Designed experiments, time series, spatial data, and clustered observations may not allow arbitrary splitting without changing the estimand or the reference law.

## 7. Polyhedral Selective Inference for the Lasso

**Route: Core**

Chapter 11 treated inference for pre-specified coordinates in a high-dimensional linear model. Here the target is a selected-model coefficient after the Lasso has selected variables. The procedure in Lee et al. [113] assumes that the Lasso tuning parameter $\lambda$ is fixed before looking at $y$. CV is itself a selection procedure; treating a cross-validated $\lambda$ as fixed changes the conditioning event and is not covered by the basic theorem.

Consider

$$y \sim N(\mu, \sigma^2 I_n),$$




FIGURE 12.3. A schematic comparison of interval width multipliers. The dashed vertical line at 1.96 marks the ordinary two-sided 95 percent normal cutoff. The other endpoints illustrate different types of protection; they are not calculated interval widths and do not assert that one method is always shorter than another. POSI protects all allowed model choices, whereas FCR and conditional methods address specified selection problems.

with fixed design $X$. For fixed $\lambda > 0$, the Lasso estimator is

$$\widehat{\beta}^{\lambda} \in \arg\min_{b} \left\{ \frac{1}{2}\|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}.$$

This section uses an unnormalized squared-error loss to simplify the KKT equations. The same fit in (6), whose loss is divided by $n$, uses penalty $\lambda/n$. Let

$$\widehat{M} = \{j : \widehat{\beta}_j^{\lambda} \neq 0\}, \qquad \widehat{s} = \operatorname{sign}(\widehat{\beta}^{\lambda}_{\widehat{M}}).$$

The KKT conditions are

$$X_j^{\top}(y - X\widehat{\beta}^{\lambda}) = \lambda \operatorname{sign}(\widehat{\beta}_j^{\lambda}) \quad (j \in \widehat{M}),$$

and

$$|X_j^{\top}(y - X\widehat{\beta}^{\lambda})| \leq \lambda \quad (j \notin \widehat{M}).$$

For a fixed active set $M$ and sign vector $s \in \{\pm 1\}^{|M|}$, three ingredients pin down the selection event: the active-set KKT equations expressed in $y$, the inactive-set inequalities expressed in $y$, and the strict sign constraint $\operatorname{diag}(s)\widehat{\beta}^{\lambda}_M > 0$ that ensures the active coefficients carry the prescribed signs. After eliminating $\widehat{\beta}^{\lambda}$ using the KKT identity, the equality and inequality parts together can be packaged, up to Gaussian-null boundary events that have probability zero under $y \sim N(\mu, \sigma^2 I_n)$, as the polyhedral event

$$\{\widehat{M} = M, \widehat{s} = s\} = \{Ay \leq b\}, \tag{31}$$

where $A$ and $b$ depend on $X$, $\lambda$, $M$, and $s$, but not on the unknown mean $\mu$.

Figure 12.4 shows the two reductions used by the selective pivot: KKT inequalities restrict $y$ to one selected polyhedron, and conditioning on the nuisance projection reduces that region to a line segment for the target contrast.

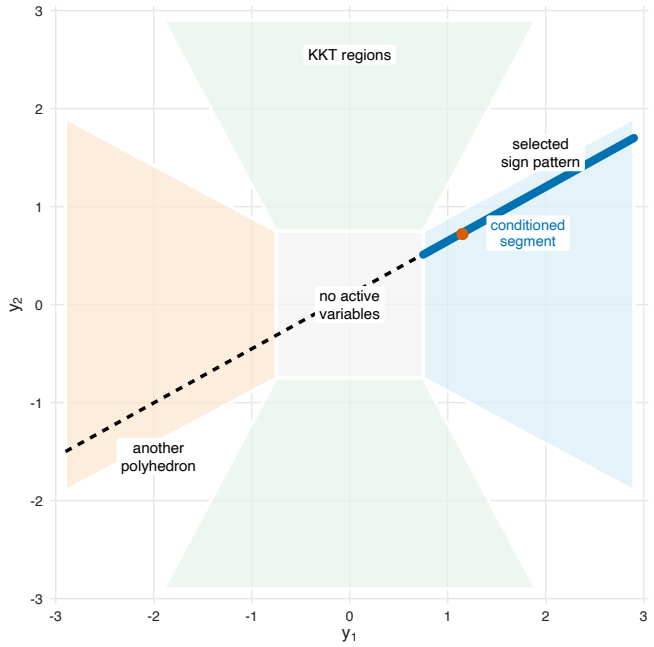



FIGURE 12.4. Schematic regions for different Lasso selection outcomes in response space. Gray represents no active variables; blue represents the selected model and sign pattern; orange and green represent other outcomes. The two green regions are distinct regions drawn in the same color. These colors identify selection outcomes, not significance levels. Within the blue region, the orange point is the observed response and the solid line is the set left after fixing the nuisance projection. The dashed continuation leaves the selected region. The geometry is illustrative rather than an exact partition for a specified design.

Fix a selected model $M$, a selected coordinate $j \in M$, and the contrast

$$\eta^\top = e_j^\top (X_M^\top X_M)^{-1} X_M^\top,$$

using the selected-model notation from Section 6. The target is

$$\eta^\top \mu = \beta_{j \bullet M}.$$

Without selection, $\eta^\top y$ is normal with mean $\eta^\top \mu$ and variance $\sigma^2 \|\eta\|_2^2$. After selection, we condition on $\{Ay \le b\}$. To obtain a one-dimensional pivot, also condition on the nuisance projection

$$P_{\eta^\perp} y = \left( I_n - \frac{\eta \eta^\top}{\|\eta\|_2^2} \right) y.$$

Because $y$ is Gaussian, $\eta^\top y$ is independent of this orthogonal projection.

Let $z = P_{\eta^\perp} y$ be the observed nuisance component. Along the line

$$y = z + \frac{\eta}{\|\eta\|_2^2} t, \qquad t = \eta^\top y,$$

the polyhedral constraint $Ay \le b$ becomes an interval constraint

$$t \in [V^-(z), V^+(z)].$$

Explicitly, for each row $a_\ell^\top$ of $A$,

$$a_\ell^\top z + \frac{a_\ell^\top \eta}{\|\eta\|_2^2} t \le b_\ell,$$

which gives an upper bound on $t$ when $a_\ell^\top \eta > 0$, a lower bound when $a_\ell^\top \eta < 0$, and a feasibility check $a_\ell^\top z \le b_\ell$ when $a_\ell^\top \eta = 0$. Intersecting,

$$V^-(z) = \max_{\ell:\, a_\ell^\top \eta < 0} \frac{\|\eta\|_2^2 (b_\ell - a_\ell^\top z)}{a_\ell^\top \eta}, \qquad V^+(z) = \min_{\ell:\, a_\ell^\top \eta > 0} \frac{\|\eta\|_2^2 (b_\ell - a_\ell^\top z)}{a_\ell^\top \eta},$$

with $V^-(z) = -\infty$ and $V^+(z) = +\infty$ when the corresponding index sets are empty.

**Theorem 12.3: Truncated-normal selective pivot**

Under the fixed-$\lambda$ Gaussian lasso setting, condition on $\{\widehat{M} = M, \widehat{s} = s\}$ and on $P_{\eta^\perp} y = z$. Then

$$\eta^\top y \sim \mathrm{TN}\left(\eta^\top \mu,\, \sigma^2 \|\eta\|_2^2,\, [V^-(z), V^+(z)]\right),$$

where TN denotes a normal distribution truncated to the displayed interval. Therefore, if

$$F^{\nu,\tau^2}_{[V^-,V^+]}$$

is the CDF of $N(\nu, \tau^2)$ truncated to $[V^-, V^+]$, then

$$F^{\eta^\top \mu, \sigma^2 \|\eta\|_2^2}_{[V^-(z), V^+(z)]}(\eta^\top y) \sim \mathrm{Unif}(0,1)$$

under the conditional law.

**Proof of Theorem 12.3**

Decompose $y$ into the scalar component $t = \eta^\top y$ and the nuisance component $z = P_{\eta^\perp} y$:

$$y = z + \frac{\eta}{\|\eta\|_2^2} t.$$

For a Gaussian vector with covariance $\sigma^2 I_n$, the scalar $\eta^\top y$ is independent of $P_{\eta^\perp} y$, with marginal distribution

$$\eta^\top y \sim N(\eta^\top \mu, \sigma^2 \|\eta\|_2^2).$$

After conditioning on $P_{\eta^\perp} y = z$, the only remaining randomness is therefore the one-dimensional variable $t$. The selection event $\{Ay \le b\}$, restricted to the line above, is exactly the interval constraint $t \in [V^-(z), V^+(z)]$, because each row of $A$ contributes one linear upper bound, lower bound, or feasibility check. Conditioning a normal random variable on belonging to that interval gives the stated truncated normal law. Applying its conditional CDF to the observed value is uniform by the probability integral transform for the truncated distribution. □

Inverting the pivot gives a selective confidence interval for $\eta^\top \mu$:

$$C_j = \left\{ \nu : \frac{\alpha}{2} \le F^{\nu, \sigma^2 \|\eta\|_2^2}_{[V^-(z), V^+(z)]}(\eta^\top y) \le 1 - \frac{\alpha}{2} \right\}.$$

This interval targets the selected-model coefficient $\beta_{j \bullet M}$, conditional on the model and sign pattern selected by the fixed-$\lambda$ Lasso. It is often shorter than POSI because the selection rule

is much more specific. The cost is that the analyst must commit to the selection rule, and the computation can become unstable when the conditioning polyhedron is narrow.

## 8. Selective Inference for Clustering

**Route: Advanced**

Selection can also create groups. Suppose independent observations

$$W_i \sim N(\mu_i, \sigma^2 I_d), \qquad i = 1, \dots, n,$$

with known $\sigma > 0$ are clustered by an algorithm $\mathcal{C}$. Write $W \in \mathbb{R}^{n \times d}$ for the matrix with rows $W_i^\top$, and $w$ for its observed value. After seeing the data, we may choose two clusters $\widehat{C}_1, \widehat{C}_2 \in \mathcal{C}(W)$ and test

$$H_{0,\widehat{C}_1,\widehat{C}_2} : \bar{\mu}_{\widehat{C}_1} = \bar{\mu}_{\widehat{C}_2}.$$

A naive Wald test treats the clusters as fixed. But under a global null, clustering will still tend to separate observations into groups with different sample means. The selected clusters are not fixed labels; they were produced by the same noise used for testing.

*Separate the difference being tested from the other information.* For the derivation, fix two nonempty disjoint sets $C_1, C_2$. We hold their memberships fixed in the calculation, but retain the selection step by considering only data sets for which the clustering returns these same groups. For any group $C$, write $\bar{W}_C = |C|^{-1} \sum_{i \in C} W_i$ and $\bar{\mu}_C = |C|^{-1} \sum_{i \in C} \mu_i$. Define the contrast vector $v = v(C_1, C_2)$ by

$$v_i = \frac{\mathbf{1}\{i \in C_1\}}{|C_1|} - \frac{\mathbf{1}\{i \in C_2\}}{|C_2|}.$$

Then

$$\bar{W}_{C_1} - \bar{W}_{C_2} = W^\top v, \qquad \|v\|_2^2 = |C_1|^{-1} + |C_2|^{-1}.$$

Let

$$P_v^\perp = I_n - \frac{v v^\top}{\|v\|_2^2}.$$

This projection removes the component in the direction $v$ from each column of $W$. In the remaining matrix $P_v^\perp W$, the two groups have the same sample mean; the removed component contains their mean difference. The vector $W^\top v$ has mean $\bar{\mu}_{C_1} - \bar{\mu}_{C_2}$ and covariance $\sigma^2 \|v\|_2^2 I_d$. It is independent of $P_v^\perp W$: the two are jointly Gaussian, and their cross-covariance is zero because $P_v^\perp v = 0$. The decomposition

$$W = P_v^\perp W + \frac{v (W^\top v)^\top}{\|v\|_2^2}$$

therefore separates the cluster mean difference from the remaining data.

Define the *separation* and its observed value by

$$\delta_{\text{sep}} = \|W^\top v\|_2, \qquad \delta_{\text{sep,obs}} = \|w^\top v\|_2.$$

Under the pairwise null, $W^\top v$ is a centered spherical Gaussian vector. Its length is independent of its direction, and

$$\delta_{\text{sep}} \sim \sigma \sqrt{|C_1|^{-1} + |C_2|^{-1}}\, \chi_d,$$

where $\chi_d$ is the length of a standard $N(0, I_d)$ vector. This is the reference law before accounting for clustering.

*Vary only the separation and rerun the clustering.* Why condition on more than the selected groups? The pairwise null equates two group-average means, but leaves the remaining mean structure unknown. The distribution conditional only on the selected groups can still depend on those unknown means. We therefore fix the orthogonal projection $P_v^\perp W$, whose mean contains

these nuisance quantities. Fixing the direction of the mean difference as well leaves only its length to vary. Under the null, the two independence properties above ensure that these extra conditions preserve the scaled chi law before selection.

Let $u = w^\top v / \|w^\top v\|_2$ be the observed direction of the mean difference; a zero difference has probability zero under the model. Condition on $P_v^\perp W = P_v^\perp w$ and on this direction. For a possible separation $r \geq 0$, the corresponding data matrix is

$$w(r) = P_v^\perp w + \frac{r}{\|v\|_2^2} v u^\top.$$

Indeed, $w(r)^\top v = ru$, so its separation is exactly $r$, and $w(\delta_{\text{sep,obs}}) = w$. Increasing $r$ moves the two groups apart along their observed mean-difference direction; decreasing $r$ moves them together. All observations in a given group receive the same shift, so the within-group configuration stays fixed, as do observations outside the two groups.

Now define the *allowed separations* explicitly:

$$\mathcal{S}(w, C_1, C_2) = \{r \geq 0 : C_1, C_2 \in \mathcal{C}(w(r))\}.$$

To check whether $r$ belongs to this set, cluster $w(r)$ and ask whether the same two groups appear. We require these groups to reappear; the other groups need not form the same partition. This set can contain several disjoint intervals. To see why it describes the selection event exactly, substitute the fixed projection and direction into the decomposition:

$$W = P_v^\perp w + \frac{\delta_{\text{sep}}}{\|v\|_2^2} v u^\top = w(\delta_{\text{sep}}).$$

Thus, under this conditioning,

$$\{C_1, C_2 \in \mathcal{C}(W)\} = \{\delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)\}.$$

Selection of the two groups is now a restriction on a single scalar, the separation [67].

*Use the null law restricted to the allowed separations.* Under the pairwise null, the separation is independent of the fixed projection and direction. Conditioning further on the clustering event therefore restricts its scaled chi law to $\mathcal{S}(w, C_1, C_2)$. The selective p-value is the upper-tail probability in that restricted law:

$$p_{\text{sel}} = \frac{\mathbb{P}\{\delta_{\text{sep}} \geq \delta_{\text{sep,obs}},\ \delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)\}}{\mathbb{P}\{\delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)\}}, \tag{32}$$

where both probabilities use the scaled chi reference law above and treat $w, C_1, C_2$ as fixed. The denominator keeps only separations that would reproduce the selected groups; the numerator also requires separation at least as large as observed.

For a direct check of validity, fix the projection and direction, and let $F_{\mathcal{S}}$ be the CDF of the scaled chi law restricted to the resulting set $\mathcal{S}$. The tail ratio above is

$$p_{\text{sel}} = 1 - F_{\mathcal{S}}(\delta_{\text{sep,obs}}).$$

Under the pairwise null and conditional on the selected groups, the separation has this continuous CDF. The probability integral transform therefore makes $F_{\mathcal{S}}(\delta_{\text{sep}})$ uniform on $[0, 1]$. Under this conditional law,

$$\mathbb{P}\{1 - F_{\mathcal{S}}(\delta_{\text{sep}}) \leq \alpha\} = 1 - (1 - \alpha) = \alpha.$$

This holds for almost every projection and direction compatible with selection. Averaging over their distribution among the selected data sets, by the law of total probability, gives

$$\mathbb{P}\{p_{\text{sel}} \leq \alpha \mid C_1, C_2 \in \mathcal{C}(W)\} = \alpha,$$

for a selection event of positive probability. The additional conditioning has made the calculation possible while preserving the desired conditional error rate. Unknown means in the orthogonal

projection do not enter the scaled chi law. If an additional data-dependent rule chooses which pair of clusters to compare, that rule must also enter the selection event, or multiplicity across the examined pairs must be handled separately.

Figure 12.5 highlights why an ordinary between-cluster test is invalid: the same separation that defines the test statistic helped create the clusters, so the reference law must be truncated to separations that reproduce the selected pair of groups.

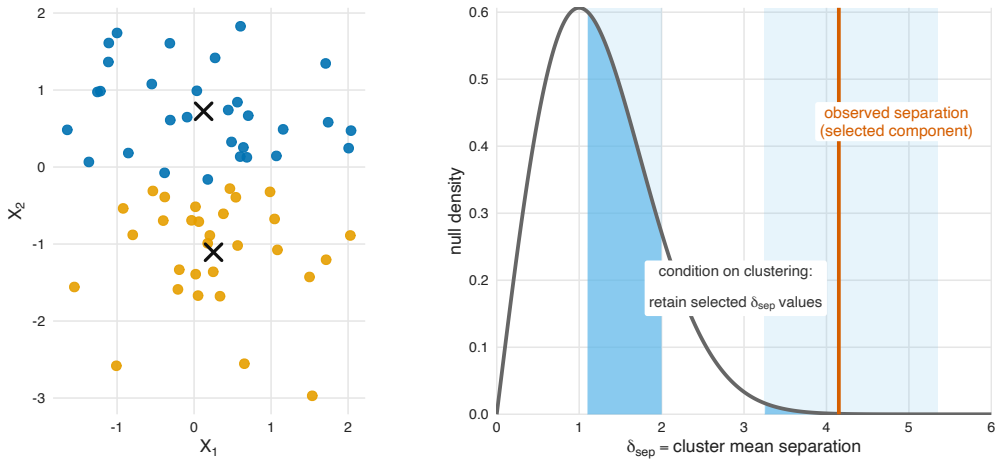



FIGURE 12.5. Selective inference for clustering. Even under a null model, clustering creates separated groups. Conditional inference varies the cluster separation $\delta_{\text{sep}}$ along a line determined by the observed nuisance information and evaluates the null law truncated to values that reproduce the selected clustering event. In the schematic right panel, that selected support has two disconnected components, and the observed separation lies in the right component.

For Monte Carlo computation, draw separations independently from the scaled chi law, reconstruct $w(r)$, and rerun the clustering for each draw. Use the same algorithm and settings each time, holding any independent algorithmic random seed fixed. Among the draws that reproduce $C_1, C_2$, the fraction with $r \geq \delta_{\text{sep,obs}}$ estimates (32). This is a numerical approximation: if very few draws are retained, more draws or a more efficient sampling method are needed. No retained draws means that this estimate is undefined, not that the p-value is zero. The same principle applies to hierarchical clustering [67] and $k$-means [41], although the allowed separations depend on the algorithm. Whenever the set can be computed explicitly, one can integrate the chi density over its intervals instead of simulating.

## 9. Data Thinning and Data Fission

**Route: Advanced**

Sample splitting is the simplest reusable device for selective inference: split the data into a discovery half and an inference half, choose a target on the discovery half, and conduct standard inference on the inference half. When the two halves are independent and the inference method is valid for the held-out data and chosen target, selection on the first half does not invalidate inference on the second. This avoids the selection-event bookkeeping of Section 7. It also has a fundamental limitation. When the inferential target is the *latent structure* of the data — clusters, principal components, change-points, or any structure jointly identified across observations —

splitting observations into two halves changes the target. The clusters estimated on the training half may not map cleanly to the inference half, and the parameter we wished to test ("the mean of cluster 1 estimated by $k$-means on the full data") is not the same as the parameter we can test ("the mean of cluster 1 estimated by $k$-means on the training half").

*Data thinning* and *data fission* address this limitation by splitting *within* each observation rather than *across* observations.

**Data thinning for convolution-closed distributions.** Neufeld et al. [140] introduce data thinning for distributions closed under convolution. Suppose $X$ has a distribution in a parametric family $\{P_\theta : \theta \in \Theta\}$ such that the convolution of two members of the family is again in the family. Examples include Gaussian, Poisson, negative binomial, and gamma families. Thinning chooses a parameter $\epsilon \in (0, 1)$ and constructs two independent variables $(X^{(1)}, X^{(2)})$ by

$$X^{(1)} \sim P_{\epsilon\theta}, \quad X^{(2)} \sim P_{(1-\epsilon)\theta}, \quad X = X^{(1)} + X^{(2)}.$$

Because the family is convolution-closed and the parameter is additive, the sum $X^{(1)} + X^{(2)}$ reproduces the distribution of $X$. More importantly, $X^{(1)}$ and $X^{(2)}$ are independent.

**Theorem 12.4: Independence of thinned components**

Let $X \sim P_\theta$ belong to a convolution-closed family with additive parameter $\theta$: that is, if $Y_1 \sim P_{\theta_1}$ and $Y_2 \sim P_{\theta_2}$ are independent then $Y_1 + Y_2 \sim P_{\theta_1+\theta_2}$. Fix $\epsilon \in (0, 1)$, and assume $\epsilon\theta, (1-\epsilon)\theta \in \Theta$. Then there is a distributional split $(X^{(1)}, X^{(2)})$ with $X^{(1)} \perp X^{(2)}$, $X^{(1)} + X^{(2)} \overset{d}{=} X$, $X^{(1)} \sim P_{\epsilon\theta}$, and $X^{(2)} \sim P_{(1-\epsilon)\theta}$. An implementable thinning mechanism from an observed $X$ is obtained by sampling from the conditional law of $(Y_1, Y_2)$ given $Y_1 + Y_2 = X$ when that conditional law is known and does not involve unknown nuisance parameters, or when those parameters are known.

**Proof of Theorem 12.4**

First draw $Y_1 \sim P_{\epsilon\theta}$ and $Y_2 \sim P_{(1-\epsilon)\theta}$ independently. Convolution closure gives $Y_1 + Y_2 \sim P_\theta$, so the sum has the same distribution as $X$.
We can generate the same pair in the opposite order: first draw the total $X \sim P_\theta$, then, given $X = x$, draw the parts from the conditional distribution of $(Y_1, Y_2)$ given $Y_1 + Y_2 = x$. The law of total probability shows that these two ways of generating the pair give the same joint distribution. If the conditional splitting law is available and has no unknown parameters, the second construction can be applied to the observed total. Averaging over the random total therefore recovers the distribution of the independent pair we started with. Thus the thinned components are independent unconditionally, with marginals $P_{\epsilon\theta}$ and $P_{(1-\epsilon)\theta}$. □

Worked example: Poisson thinning. Let $X \sim \text{Poisson}(\lambda)$. After observing a count $X = x$, assign each of its $x$ events to the first part with probability $\epsilon$, independently. Equivalently, draw

$$X^{(1)} \mid X = x \sim \text{Binomial}(x, \epsilon), \qquad X^{(2)} = x - X^{(1)}.$$

The two parts sum to the observed count. For $x > 0$, they are dependent conditional on $X = x$: knowing one determines the other. The independence claim is *unconditional*, averaging over both the original Poisson count and the random split. To check it, let $a, b$ be nonnegative integer

values of the two parts. The event $\{X^{(1)} = a, X^{(2)} = b\}$ requires the original count to be $a+b$, so

$$\begin{aligned}\mathbb{P}(X^{(1)} = a, X^{(2)} = b) &= \mathbb{P}(X = a+b)\mathbb{P}(X^{(1)} = a \mid X = a+b) \\ &= e^{-\lambda}\frac{\lambda^{a+b}}{(a+b)!}\binom{a+b}{a}\epsilon^a(1-\epsilon)^b \\ &= \left[e^{-\epsilon\lambda}\frac{(\epsilon\lambda)^a}{a!}\right]\left[e^{-(1-\epsilon)\lambda}\frac{((1-\epsilon)\lambda)^b}{b!}\right].\end{aligned}$$

The last line is the product of two Poisson mass functions. Thus the parts are independent, with means $\epsilon\lambda$ and $(1-\epsilon)\lambda$. The splitting rule itself does not require knowing $\lambda$.

Other families require their own conditional splitting rules and parameter conditions; these are given in Neufeld et al. [140].

The consequence for inference is that we can perform a data-driven step (clustering, principal component analysis (PCA), change-point detection, feature selection) on $X^{(1)}$ and then apply an inference method valid for the distribution of $X^{(2)}$ and the chosen target. Independence protects this second step from selection on $X^{(1)}$; exactness still requires a valid reference law. The two matrices have the same dimension as the original data, so latent structures identified on $X^{(1)}$ map directly to $X^{(2)}$.

**Example 12.5: Single-cell clustering**

Let $X_{ij}$ be the read count of gene $j$ in cell $i$. Assume all counts are independent Poisson variables conditional on a fixed mean matrix $(\mu_{ij})$. Use independent binomial randomizations for all entries to compute the thinned counts $X^{(1)}_{ij} \sim \text{Poisson}(\epsilon\mu_{ij})$ by sampling $X^{(1)}_{ij} \mid X_{ij} \sim \text{Binomial}(X_{ij}, \epsilon)$ and setting $X^{(2)}_{ij} = X_{ij} - X^{(1)}_{ij}$. By Theorem 12.4 the matrices $X^{(1)}$ and $X^{(2)}$ are independent. Cluster the cells using $X^{(1)}$ to obtain a partition $\widehat{\mathcal{C}} = (\widehat{C}_1, \ldots, \widehat{C}_g)$. Any clustering algorithm may be used, such as $k$-means or hierarchical clustering. The algorithm, preprocessing, and choice of the number of clusters may all be chosen using $X^{(1)}$, but must not use $X^{(2)}$. This flexibility follows from independence, rather than from a special property of one clustering algorithm [140]. Now test, on $X^{(2)}$, the hypothesis that two estimated clusters $\widehat{C}_a$ and $\widehat{C}_b$ share the same per-gene Poisson mean. Because $\widehat{\mathcal{C}}$ is a function of $X^{(1)}$ alone and $X^{(2)}$ is independent of $X^{(1)}$, conditioning on the partition leaves the law of $X^{(2)}$ unchanged. For each realized partition, treat the group indices as fixed and use a test valid under the proposed Poisson null, with inference-view means $(1-\epsilon)\mu_{ij}$. For example, equality of the average rates for a prespecified gene can be tested by conditioning the two group totals on their sum, as in Exercise 12.16. If many genes or cluster pairs are tested, their multiplicity must still be handled.

**Remark 12.6: Why splitting cells changes the target**

A natural alternative is to split cells into two halves, cluster on one half, and test on the other. But the clusters identified on the training half are defined by the cells in that half; they may not exist as identifiable structures on the test half, and any rule for projecting them onto the test half (e.g., assign each test cell to the nearest training centroid) introduces a selection event in the projection that data thinning sidesteps by keeping all observations in both $X^{(1)}$ and $X^{(2)}$.

Dharamshi et al. [49] extend the construction to a broader class of distributions by replacing the additive parameter requirement with a sufficient-statistic decomposition. This generalization covers uniform, beta, and Wishart distributions and recovers the original thinning construction for natural exponential families.

**Data fission.** Leiner et al. [119] propose a closely related but more permissive framework called *data fission*. Rather than requiring an exact convolution decomposition, data fission decomposes a single observation into two parts $(f(X), g(X))$ such that either (i) $f(X)$ and $g(X)$ are independent with known distributions, or (ii) the conditional distribution of $g(X)$ given $f(X)$ is known in closed form. The second condition weakens the independence requirement of data thinning, in exchange for a step that uses the conditional reference law for inference. This is an adjustment for dependence between the two views, distinct from the Lasso bias correction of Chapter 11; any unknown nuisance parameters in that conditional law must still be handled.

**Example 12.7: Gaussian fission**

Let $X \sim N(\mu, \tau^2)$ with $\tau^2 > 0$ known. Generate an external $Z \sim N(0, \sigma^2)$, with $\sigma^2 > 0$, independent of $X$ and define

$$f(X) = X + Z, \qquad g(X) = X - \frac{\tau^2}{\sigma^2} Z.$$

We claim $f(X) \perp g(X)$ and compute their marginals.

**Proof of the claim in Example 12.7**

The pair $(f(X), g(X))$ is a linear transformation of the jointly Gaussian pair $(X, Z)$, hence jointly Gaussian. Independence then reduces to zero covariance:

$$\mathrm{Cov}(f(X), g(X)) = \mathrm{Cov}(X + Z,\, X - \tau^2 Z/\sigma^2) = \mathrm{Var}(X) - \frac{\tau^2}{\sigma^2}\,\mathrm{Var}(Z) = \tau^2 - \tau^2 = 0.$$

For the marginals, $f(X) = X + Z$ has mean $\mu$ and variance $\tau^2 + \sigma^2$; similarly,

$$\mathbb{E}[g(X)] = \mu, \qquad \mathrm{Var}(g(X)) = \mathrm{Var}(X) + \frac{\tau^4}{\sigma^4}\,\mathrm{Var}(Z) = \tau^2 + \frac{\tau^4}{\sigma^2}.$$

Selecting hypotheses based on $f(X)$ and conducting inference on $g(X)$ is valid because any function of $f(X)$ is independent of $g(X)$. □

The Gaussian construction illustrates the central trade-off: as $\sigma^2 \to \infty$, $f(X)$ becomes very noisy (less informative for selection) while $g(X)$ approaches $X$ (more informative for inference); as $\sigma^2 \to 0$, the reverse. Choosing $\sigma^2 = \tau^2$ makes $f$ and $g$ symmetric: both views have variance $2\tau^2$, so they are equally informative after the noise injection. Other distributions require their own randomization and reference law; the Gaussian coefficient $\tau^2/\sigma^2$ cannot simply be transferred to a Poisson or other non-Gaussian model.

**Remark 12.8**

Data thinning and data fission retain all observation labels in both views. They do not retain all the original information in each view: the allocation parameter trades information for selection against information for inference. Keeping the same cell indices allows a

selected partition to be tested in the second view, but the usefulness of that partition and the cost in power depend on the model and the allocation.

The two approaches account for selection in different ways. Polyhedral selective inference keeps the original data and derives a reference distribution conditional on the model and signs selected by the Lasso. It treats the design matrix as fixed, so no probability model for that matrix is needed; the exact pivot still requires the stated Gaussian response model and a fixed $\lambda$. Data thinning instead creates two independent parts under a specified distributional model. We may choose questions using any rule based only on the first part, then answer them using a test or interval valid for the second part. This avoids describing every possible outcome of the selection algorithm, but it does not remove the distributional assumptions needed for thinning and inference.

## 10. Selective Inference Beyond Fixed-$\lambda$ Models

**Route: Research frontier**

Polyhedral selective inference for the Lasso, as developed in Section 7, conditions on a particular fixed $\lambda$ and a particular selected active set and sign pattern. In practice, $\lambda$ is rarely fixed: it is chosen by CV, selected by an information criterion, or chosen adaptively from the data. The selection rule, once it includes the tuning step, is no longer a single polyhedral event, and the truncated-normal pivot of Section 7 does not apply.

**Cross-validated Lasso.** The selective-CV construction of Loftus and Taylor [130] represents squared-error foldwise validation criteria as quadratic forms in the data. Comparing candidate tuning parameters therefore introduces quadratic inequalities. These are combined with the linear or quadratic constraints that describe the training procedure's fitted-model path. Conditioning on the CV-selected $\lambda$, active set, and signs consequently gives a quadratically constrained selection region, not the single polyhedron that arises for a fixed-$\lambda$ Lasso fit. This difference explains both the additional bookkeeping and the harder computation: the reference law must be evaluated over the resulting quadratic region, often by specialized sampling or randomized selective-inference methods.

**Randomized selective inference and data carving.** A different approach adds independent randomization to the selection rule, so that the selected outcome reveals less about the data used for inference. Data carving reuses information from the discovery data that remains after conditioning on the selection event, together with any untouched inference data [63]. Both approaches can improve power over discarding all discovery data, but their reference laws must account for the actual randomization and selection. The amount of information recovered depends on that design.

**Selective inference for non-linear models.** Taylor and Tibshirani [175] extend selective inference to $\ell_1$-penalized likelihood models such as logistic and Poisson regression. The polyhedral characterization no longer applies exactly, but a local quadratic approximation around the selected solution yields an asymptotically valid selective pivot.

**Selective inference for change-points and graph saliency.** Recent extensions apply selective inference principles to selection events that are not polyhedral at all. Shiraishi et al. [162] develop selective inference for change-point detection by recurrent neural networks (RNNs): the selection event is the non-linear, temporally constrained sequence of states visited by the RNN, and the conditional distribution under the null is characterized by a sampling scheme. Nishino

et al. [142] apply selective inference to saliency maps of graph neural networks (GNNs): the selected nodes and edges are those flagged by a GNN explainability method, and the conditional distribution under the null accounts for the spatial dependence in the underlying graph. In both cases, the selective pivot is computed by simulating the algorithmic selection process under the null and using the empirical distribution as the reference law.

These extensions all share a common structure: identify the selection event, characterize the conditional distribution of the test statistic given that event, and invert. The polyhedral case of Section 7 is the cleanest instance; the algorithmic cases of change-point and graph saliency analysis are the most complex. Data thinning instead separates the selection view from an independent inference view. An arbitrary rule may use the selection view; inference on the other view must still use its correct sampling model and the chosen target.

## 11. Assumptions in Plain Language

**Route: Core**

Before reporting an interval, state which parameter it covers and how the data were used to choose that parameter. FCR bounds the expected fraction of reported intervals that miss their targets; it does not promise coverage for each interval conditional on its selection. POSI protects all coefficients in every allowed model at once, which can make its intervals wide. Lasso selective inference uses the particular model and sign pattern selected at a fixed $\lambda$. For clustering, the calculation must account for the algorithm and the rule used to choose the clusters being compared.

A common mistake is to use the data to choose an analysis, then calculate uncertainty as if that analysis had been chosen in advance. For example, choosing $\lambda$ by CV adds a selection step that the fixed-$\lambda$ Lasso calculation does not include.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Write down the parameter of interest and every choice made after looking at the data, including screening variables, choosing a penalty, or choosing clusters. Check that rerunning these steps gives the reported selection and that the inference calculation includes them. If p-values are computed by simulation, report their Monte Carlo uncertainty. |
| **What these data cannot show** | The final selected model does not reveal all the choices that led to it. A calculation that omits those choices can give misleading p-values or intervals. Statistical coverage also does not establish that the selected coefficient answers the scientific question. |
| **Safer fallback** | If the selection steps cannot be recovered, use new, independent data to test the chosen hypotheses or estimate the chosen parameters. Under the fixed-design Gaussian assumptions, POSI can instead protect against any choice within its specified collection of models. |

## 12. Failure Modes and Diagnostics

**Route: Core**

- Estimates selected for being unusually large often move closer to their true values in a new study: some of their original size was noise. This is regression to the mean. Widening intervals to control FCR does not by itself correct this bias in their centers.
- POSI can give unnecessarily wide intervals when only a small set of models was considered, but the calculation protects a much larger set.
- A Lasso selective interval must account for screening before fitting, CV, and any tuning done after inspecting results. The fixed-$\lambda$ calculation alone does not cover those extra choices.
- If the observed selection is very unlikely under the assumed model, the conditional calculation may be numerically difficult and its intervals very wide.
- For clustering, changing the distance, number of clusters, or preprocessing can change which groups are selected. The conditional test must account for the choices that were actually used.

Report enough detail to reproduce the selection, say whether the target was fixed in advance or depends on the selected model, and give the number of reported intervals $R_{\mathrm{CI}}$. Also report numerical difficulties and very wide intervals, rather than hiding them by changing the calculation after seeing its output.

## 13. Bibliographic Notes

**Route: Core**

Sorić's warning about selected confidence intervals is an early statement of the problem [167]. FCR and selected confidence interval adjustments were introduced by Benjamini and Yekutieli [22]. POSI was developed by Berk et al. [24]. The general selective-inference framework that organizes polyhedral and other conditional pivots, including the optimality viewpoint, is due to Fithian et al. [63]. Polyhedral selective inference for the Lasso is due to Lee et al. [113]; related sequential regression procedures are developed in Tibshirani et al. [177]. Selective inference after clustering is developed for hierarchical clustering in Gao et al. [67] and for $k$-means in Chen and Witten [41].

Data thinning for convolution-closed distributions is due to Neufeld et al. [140]; the sufficient-statistic generalization that covers uniform, beta, and Wishart families is in Dharamshi et al. [49]. Data fission, which decomposes a single observation into independent components or components with known conditional distribution, is from Leiner et al. [119]. Selective inference for CV-selected models was developed by Loftus and Taylor [130]; the related randomization and data-carving techniques are in Fithian et al. [63]. Post-selection inference for $\ell_1$-penalized likelihood models is from Taylor and Tibshirani [175]. Recent extensions to RNN change-point detection and GNN saliency are from Shiraishi et al. [162] and Nishino et al. [142].

## 14. Exercises

**Route: Core**

### Intermediate.

**Exercise 12.9: The $R^{(i)}$ construction**

Fix $1 \le k \le m$, and let $\widehat{\mathcal{S}}$ contain the indices of the $k$ largest $|T_i|$'s, breaking ties by a fixed index order. For a selected $i$, compute $R^{(i)}$ in (28) and justify your answer from its minimum-over-perturbations definition.

**Exercise 12.10: POSI target**

For a fixed design with three predictors, write the selected-model target $\beta_{j\bullet M}$ for each two-variable model $M$. Give an example in which $\beta_{j\bullet M}$ differs from the coefficient of $j$ in the full three-variable model.

**Exercise 12.11: KKT polyhedron**

Let $X$ be a fixed two-column design of full column rank and fix $\lambda > 0$. Minimize

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

For the event that the selected active set is $M = \{1\}$ with positive sign, write the KKT conditions as linear inequalities in $y$. Identify the strict inequality needed for the active coefficient to be positive.

### Computational.

**Exercise 12.12: Selected-interval simulation**

Reproduce the common-mean simulation in Figure 12.2 using independent $Y_i \sim N(\mu, 1)$, $\alpha = 0.05$, and $\mu \in \{0, 0.1, \dots, 3\}$. For the unadjusted and BY FCR methods, select coordinates with $|Y_i| > z_{1-\alpha/2}$. Keep the usual marginal intervals for the first method; for the second, apply (29) to the same selected set. For Bonferroni, report intervals $Y_i \pm z_{1-\alpha/(2m)}$ only when they exclude zero. Use $m \in \{10, 80, 500\}$ and 2,500 independent runs for each $(\mu, m)$. Compare empirical FCR, taking the noncoverage fraction to be zero when no intervals are reported, and report Monte Carlo SEs across runs. Explain how selection and interval width affect the three FCR curves.

**Exercise 12.13: POSI constants by simulation**

Choose a fixed design with $p = 5$ linearly independent columns and $n \ge 5$, and report the design. Enumerate all nonempty models $M$ and simulate the POSI statistic

$$\max_M \max_{j \in M} |Z_{j\bullet M}|.$$

In each repetition, use the same Gaussian noise vector to compute every $Z_{j\bullet M}$, so their joint dependence is preserved. Estimate the 95 percent quantile and report the number of Monte Carlo draws. Compare it with the Bonferroni cutoff for all model-coordinate pairs $(j, M)$, counting those pairs explicitly, and with the ordinary 1.96 cutoff.

**Exercise 12.14: Truncated-normal pivot**

Fix $\mu$, a nonzero contrast $\eta$, and a simple polyhedron $\{y : Ay \le b\}$ with appreciable probability under $N(\mu, I_n)$. Simulate $y \sim N(\mu, I_n)$ and retain the draws satisfying $Ay \le b$. For each retained draw, put $z = P_{\eta^\perp} y$ and compute $V^-(z)$ and $V^+(z)$. Use the normal law with mean $\eta^\top \mu$ and variance $\|\eta\|_2^2$, truncated to $[V^-(z), V^+(z)]$, and evaluate its CDF at the observed contrast $\eta^\top y$. Check that these CDF values among the retained draws are approximately uniform.

## Advanced.

**Exercise 12.15: Clustering truncation set**

Let $W = (W_1, W_2, W_3)^\top \sim N_3(\mu, I_3)$ and let the clustering rule be the fixed threshold split

$$C_1(W) = \{i : W_i \le 0\}, \qquad C_2(W) = \{i : W_i > 0\}.$$

For the observed vector $w = (-1.0, -0.4, 0.8)^\top$, the selected clusters are $C_1 = \{1, 2\}$ and $C_2 = \{3\}$. Define

$$v_i = \frac{\mathbf{1}\{i \in C_1\}}{|C_1|} - \frac{\mathbf{1}\{i \in C_2\}}{|C_2|}.$$

Test $H_0 : v^\top \mu = 0$ against $v^\top \mu < 0$, with the noise variance known to equal one. Here use the *signed* contrast $t = v^\top W$, which differs from the nonnegative separation $\delta_{\text{sep}}$ in the chapter. Condition on $P_v^\perp W = P_v^\perp w$ and write

$$W(t) = P_v^\perp w + \frac{t}{\|v\|_2^2} v, \qquad t \in \mathbb{R}.$$

Derive the interval of $t$ values for which the threshold split still returns the same two clusters. Compute the observed contrast and compare its naive lower-tail Gaussian p-value with the lower-tail selective p-value from the normal law truncated to your interval.

**Exercise 12.16: Poisson thinning**

Use the Poisson-thinning construction from Section 9. For a fixed $0 < \epsilon < 1$, draw $X_{ij}^{(1)} \mid X_{ij} \sim \text{Binomial}(X_{ij}, \epsilon)$ independently across entries and set $X_{ij}^{(2)} = X_{ij} - X_{ij}^{(1)}$. The worked example establishes the independence of the two thinned views.

For a type-I-error simulation, generate $n$ independent cells and $p$ independent genes under the global null

$$X_{ij} \sim \text{Poisson}(\lambda_j), \qquad i = 1, \dots, n, \quad j = 1, \dots, p,$$

with fixed positive $\lambda_j$'s. Choose and report $n$, $p$, $\epsilon$, the rates, the nominal level, the clustering rule, and the number of independent repetitions. Thin every entry independently, cluster the cells into two nonempty groups using only $X^{(1)}$, and select one gene $j_0$ before running the simulation. On $X^{(2)}$, test equality of the two cluster means for gene $j_0$ by conditioning the two group totals on their sum; under the null the first total is binomial with success probability equal to the first cluster's share of the $n$ cells. Repeat the entire simulation using the following exact two-sided convention: sum the probabilities of all binomial outcomes whose probability mass is no larger than that of the observed first-group total. Set the p-value to one when the combined total is zero. Report the estimated rejection rate and its Monte Carlo SE, and assess whether they are consistent with the nominal upper bound. Explain any conservatism in terms of the discreteness of this non-randomized test. Optionally randomize at the rejection boundary and compare the resulting size.

**Exercise 12.17: CV-selected lambda as a compound selection event**

For Lasso regression, fix $F$ CV folds and a finite tuning grid $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$, using fixed tie-breaking rules. Assume each foldwise fit is unique and its active design has full column rank. Use the quadratic validation comparisons described in Section 10 to write a schematic expression for the event "CV selects $\lambda_k$" as a union over foldwise active-set and sign configurations. For each fixed configuration, express its contribution as the intersection of the training-fit constraints and the validation-error comparisons, using the linear training constraints and quadratic comparisons established in the chapter.

If each foldwise fit has at most $A$ possible active-set and sign configurations, give an upper bound on the number of configurations across all $F$ folds and $K$ tuning values before infeasible combinations are discarded. Explain why the bound can substantially overcount the nonempty selection regions.

CHAPTER 13

# Prediction-Powered and Active Statistical Inference

A satellite model can assign a forest-loss score to thousands of land parcels in seconds. Verifying whether a parcel actually lost forest during a specified period may require a field visit or a careful assessment of higher-resolution images. The prediction is cheap; the measurement that defines the scientific outcome is expensive. How can the two be combined to estimate the proportion of parcels that experienced forest loss?

Averaging the model's scores is tempting. With enough parcels, the average will be very precise. Precision, however, does not remove bias. A model that confuses seasonal changes with forest loss can produce a narrow interval around the wrong quantity. Using only verified parcels avoids that particular mistake, but discards information that might substantially improve precision.

PPI uses predictions to supply information and observed outcomes to correct their errors. Active statistical inference adds a design question: which outcomes should be measured? The methods in this chapter address population parameters, such as a prevalence, a mean, or a regression coefficient. They do not primarily aim to improve predictions for individual parcels.

For the core route, read Sections 1–3, then continue with label acquisition in Section 7 and the closing discussion. Sections 4–6 are optional: they explain regression, cross-fitting, and recalibration. The prerequisites are expectation, conditional probability, variance, and the CLT. We give complete calculations for the elementary estimators; the more general large-sample theory is described with references. The final discussion connects individual confidence statements to multiple testing and to the sequential methods of Chapter 8.

## 1. Cheap Predictions, Expensive Measurements

**Route: Core**

Fix a region, a period, and a rule for what counts as forest loss. Let $Y = 1$ if a parcel satisfies this rule and $Y = 0$ otherwise. Let $X$ contain the satellite measurements and other information available without verifying the outcome. Our target is

$$\mu = \mathbb{E}[Y] = \mathbb{P}(Y = 1).$$

The probability refers to the specified population of parcels. If parcels have different areas, their unweighted prevalence is not the proportion of *land area* lost. Estimating area loss requires an area-weighted target and an appropriate sampling design. The target must be settled before choosing an estimator.

For the basic development, observations are independent draws from a common population. This is a teaching model, not a claim that adjacent satellite pixels are independent. Spatial clusters, unequal selection probabilities, and repeated images of the same parcel require design-aware uncertainty calculations. We return to the sampling design later.

We use two samples:

$$\mathcal{D}_L = \{(X_i, Y_i) : i = 1, \dots, n\}, \qquad \mathcal{D}_U = \{\widetilde{X}_j : j = 1, \dots, N\}.$$

The first has verified outcomes and is called the *labeled sample*. The second contains only the inexpensive measurements and is called the *unlabeled sample*. The two samples are independent, and their feature vectors have the same distribution. Usually $N$ is much larger than $n$.

A prediction rule $f$ assigns a score $f(x)$ to a feature vector $x$. For a binary outcome it can be a probability score in $[0, 1]$; thresholding that score to obtain a binary prediction is not necessary. Initially, $f$ is fixed or trained on independent data. All expectations and variances below may be read conditional on that independent training data. No assumption that $f(x) = \mathbb{E}[Y \mid X = x]$ is imposed.

Write

$$\bar{Y}_L = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \bar{f}_L = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad \bar{f}_U = \frac{1}{N}\sum_{j=1}^{N} f(\widetilde{X}_j).$$

The labeled average $\bar{Y}_L$ estimates the desired outcome mean. The prediction averages $\bar{f}_L$ and $\bar{f}_U$ both estimate $\mathbb{E}[f(X)]$, which generally differs from $\mu$. Their difference is useful because its expectation is zero.

If the prediction-only estimator is $\bar{f}_U$, its bias for $\mu$ is

$$\mathbb{E}[\bar{f}_U] - \mu = \mathbb{E}[f(X) - Y].$$

Increasing $N$ decreases its variance but leaves this bias unchanged. An ordinary confidence interval around $\bar{f}_U$ describes uncertainty about $\mathbb{E}[f(X)]$, not automatically about $\mathbb{E}[Y]$.

**Example 13.1: A small bias with a large prediction sample**

Suppose the forest-loss prevalence is 0.15, the mean prediction is 0.18, and the standard deviation of predictions is 0.30. With $N = 10{,}000$, the SE of their average is 0.003. A typical prediction-only normal interval has half-width about 0.006. The prediction bias, 0.03, is five times this half-width. More predictions make the discrepancy easier to see; they do not correct it. These are illustrative population quantities, not results from a forest survey.

## 2. Correcting Predictions with Observed Errors

**Route: Core**

Next, we correct this bias using the labeled residuals $R_i = Y_i - f(X_i)$. Let $\bar{R}_L = n^{-1}\sum_i R_i$. A positive residual means that the prediction was too small. We add the estimated average residual to the large-sample prediction average:

$$\widehat{\mu}_{\mathrm{PPI}} = \bar{f}_U + \bar{R}_L = \bar{Y}_L + (\bar{f}_U - \bar{f}_L). \tag{33}$$

The first form explains the correction. The second shows that PPI adjusts the labeled mean by a difference whose expectation is zero. This is the mean-estimation form of Angelopoulos et al. [3]. In that literature the mean *rectifier* is $\mathbb{E}[f(X) - Y]$. Subtracting its estimate $\bar{f}_L - \bar{Y}_L = -\bar{R}_L$ is equivalent to adding the average residual used here.

**Example 13.2: Correcting the forest-loss estimate**

Suppose a study produces the following hypothetical summaries:

$$n = 200, \quad N = 10{,}000, \qquad \bar{Y}_L = 0.17, \quad \bar{f}_L = 0.20, \quad \bar{f}_U = 0.18.$$

The labeled parcels show an average residual of $0.17 - 0.20 = -0.03$. The corrected estimate is therefore

$$\widehat{\mu}_{\text{PPI}} = 0.18 - 0.03 = 0.15.$$

The correction is learned from prediction errors on the labeled parcels, not from the difference between their observed outcomes and predictions on unrelated parcels. Equivalently, the labeled parcels have slightly larger predictions compared with the large sample; PPI adjusts their outcome mean by $0.18 - 0.20 = -0.02$.

The adjustment is not an assumption that every parcel has the same prediction error. Errors may vary with terrain, cloud cover, or image quality. What matters for this estimator is that the labeled sample represents the population average of those errors.

**Proposition 13.3: Mean correction and its variance**

Assume the two samples in Section 1 are independent i.i.d. samples from the stated common population, $f$ is fixed independently of them, and $\mathbb{E}[Y^2] + \mathbb{E}[f(X)^2] < \infty$. Then

$$\mathbb{E}[\widehat{\mu}_{\text{PPI}}] = \mu,$$

and, with $\sigma_f^2 = \text{Var}(f(X))$ and $\sigma_R^2 = \text{Var}(Y - f(X))$,

$$\text{Var}(\widehat{\mu}_{\text{PPI}}) = \frac{\sigma_f^2}{N} + \frac{\sigma_R^2}{n}. \tag{34}$$

**Proof of Proposition 13.3**

The expectation is

$$\mathbb{E}[f(X)] + \mathbb{E}[Y - f(X)] = \mathbb{E}[Y].$$

The two averages in (33) use independent samples, so their covariance is zero. The variance of each average is the variance of one summand divided by its sample size, giving (34). □

The correction removes prediction bias by estimating the average prediction error. The same fixed function must be used on both samples. If different models are used, the two prediction terms may have different expectations, so their biases need not cancel.

To estimate the uncertainty, use the sample variances (for $n, N \geq 2$)

$$s_R^2 = \frac{1}{n-1}\sum_{i=1}^{n}(R_i - \bar{R}_L)^2, \qquad s_{f,U}^2 = \frac{1}{N-1}\sum_{j=1}^{N}(f(\widetilde{X}_j) - \bar{f}_U)^2.$$

The estimated SE is

$$\widehat{\text{se}}_{\text{PPI}} = \left(\frac{s_R^2}{n} + \frac{s_{f,U}^2}{N}\right)^{1/2}.$$

Both terms matter. Treating the prediction average as known can appreciably underestimate uncertainty when $N$ is not very large or predictions are highly variable.

**Proposition 13.4: A large-sample confidence interval**

Under Proposition 13.3, let $n, N \to \infty$, with $n/N \to r \in [0, \infty)$, and assume $\sigma_R^2 + r\sigma_f^2 > 0$. For fixed $\alpha \in (0, 1)$, the interval

$$C_{\text{PPI}} = \left[\widehat{\mu}_{\text{PPI}} \ \pm \ z_{1-\alpha/2}\widehat{\text{se}}_{\text{PPI}}\right] \tag{35}$$

has coverage tending to $1 - \alpha$. Here $z_u$ is the $u$ quantile of a standard normal distribution, and brackets with $\pm$ denote the two endpoints.

**Proof of Proposition 13.4**

Apply the CLT to each centered average in (33). Their independence makes their limiting variances add. If $r = 0$, both sample sizes grow but $N$ grows faster than $n$. The prediction-average uncertainty is then negligible on the $\sqrt{n}$ scale; the correction supplies the limiting variance. The sample variances converge to their population values by the law of large numbers. Dividing the centered estimator by its estimated SE therefore gives a limiting standard normal variable. □

**Example 13.5: Uncertainty in the corrected prevalence**

Continue the hypothetical study with $s_R = 0.24$ and $s_{f,U} = 0.30$. Then

$$\widehat{\text{se}}_{\text{PPI}} = \sqrt{0.24^2/200 + 0.30^2/10{,}000} \approx 0.0172.$$

The approximate 95 percent interval is $[0.116, 0.184]$. Using only the 200 binary labels gives a conventional SE approximately $\sqrt{0.17(1 - 0.17)/200} = 0.0266$. The two intervals have different centers in this realization. The numerical comparison illustrates a possible gain; it does not establish a general ordering of observed interval widths.

For a prevalence, intersecting the confidence set with $[0, 1]$ does not remove the true parameter when it was already covered. This intersection can be empty if the entire untruncated interval lies outside $[0, 1]$.

To assess precision, compare PPI with the labeled-only variance $\text{Var}(Y)/n$. PPI is more precise when

$$\text{Var}(Y - f(X)) + \frac{n}{N}\text{Var}(f(X)) < \text{Var}(Y). \tag{36}$$

Thus prediction quality matters through variability of the correction, as well as through the size of the prediction sample. It does not enter as an assumption of correctness of the model.

Adding a constant to every prediction changes neither the PPI estimate nor its variance: the extra constant in $\bar{f}_U$ cancels the one subtracted in $\bar{R}_L$. This explains how an offset predictor can be useful. Conversely, a variable prediction uncorrelated with $Y$ can add noise rather than remove it.

There is also a limit to what more unlabeled observations accomplish. With $n$ fixed and $N$ increasing, the variance approaches $\sigma_R^2/n$. Unless the prediction errors have negligible variability, the labeled sample limits the precision of the estimate. More predictions do not reduce the uncertainty in a correction based on only a few labels.

**Remark 13.6: Exact identities and approximate intervals**

Unbiasedness and (34) hold at finite sample sizes under the stated assumptions. The normal approximation in (35) does not. The original PPI framework also constructs confidence sets by combining valid bounds for prediction and correction terms. Finite-sample bounds can yield finite-sample coverage, for example with bounded outcomes and appropriate concentration inequalities. Replacing those bounds by estimated normal SEs gives the large-sample procedure taught here.

## 3. How Much Should Predictions Contribute?

**Route: Core**

Next, we choose the predictions' weight. Basic PPI uses coefficient one, which need not minimize variance. Consider instead a fixed real number $\lambda$ and

$$\widehat{\mu}_\lambda = \bar{Y}_L + \lambda(\bar{f}_U - \bar{f}_L) = \lambda \bar{f}_U + \overline{Y - \lambda f}_L. \tag{37}$$

Every fixed $\lambda$ preserves the expectation $\mu$. The value $\lambda = 0$ ignores predictions and $\lambda = 1$ recovers basic PPI. In PPI++, Angelopoulos et al. [6] choose this coefficient to reduce variance. They call this choice *power tuning*.

For a concrete comparison, suppose $\mathrm{Var}(Y) = 1.64$, $\mathrm{Var}(f(X)) = 1$, and $\mathrm{Cov}(Y, f(X)) = 0$. With $n = 200$ and $N = 5{,}000$, labels alone give variance $1.64/200 = 0.0082$. Basic PPI adds prediction noise from both samples, giving $1.64/200 + 1/200 + 1/5{,}000 = 0.0134$. Setting $\lambda = 0$ avoids that increase. The calculation below tells us which weight is best when predictions do carry information.

Let $\sigma_Y^2 = \mathrm{Var}(Y)$ and $c_{Yf} = \mathrm{Cov}(Y, f(X))$. Expanding $\mathrm{Var}(Y - \lambda f(X))$ gives

$$\mathrm{Var}(\widehat{\mu}_\lambda) = \frac{1}{n}\left\{\sigma_Y^2 - 2\lambda c_{Yf} + (1 + n/N)\lambda^2 \sigma_f^2\right\}. \tag{38}$$

The covariance term reduces variance when $\lambda$ has the same sign as $c_{Yf}$. The quadratic term increases with $\lambda^2$, so choosing too large a coefficient can increase the variance.

**Proposition 13.7: The best fixed contribution for a mean**

Under the sampling assumptions of Proposition 13.3, if $\sigma_f^2 > 0$, the variance in (38) is minimized at

$$\lambda^\star = \frac{c_{Yf}}{(1 + n/N)\sigma_f^2}. \tag{39}$$

Its minimum equals

$$\frac{1}{n}\left\{\sigma_Y^2 - \frac{c_{Yf}^2}{(1 + n/N)\sigma_f^2}\right\},$$

which is no greater than the labeled-only variance. If $f(X)$ is constant, all the estimators in (37) equal $\bar{Y}_L$.

**Proof of Proposition 13.7**

Differentiate the quadratic in braces in (38): its derivative is $-2c_{Yf} + 2(1 + n/N)\lambda\sigma_f^2$. Setting it to zero gives (39); the positive quadratic coefficient makes this the minimum. Substitute the minimizing value to obtain the displayed variance. For a constant predictor, $\bar{f}_U - \bar{f}_L = 0$ identically. □

The factor $1 + n/N$ accounts for estimating the prediction mean. When $N$ is much larger than $n$, that source of uncertainty is small. When the sample sizes are comparable, a smaller coefficient can be preferable even for a good model. A negative covariance can justify a negative coefficient for mean estimation: a reversed score can still carry information. For general loss-based optimization, unrestricted coefficients may cause computational difficulties; PPI++ treats these separately. The scalar mean calculation itself needs no convexity restriction on $\lambda$.

**Example 13.8: Combining labels with perfect predictions**

If $f(X) = Y$ almost surely and $\mathrm{Var}(Y) > 0$, then $\lambda^\star = N/(N + n)$. Equation (37) becomes the mean of all $N + n$ outcomes: the $N$ predicted outcomes are exact in this special case. Basic PPI with $\lambda = 1$ uses only their mean and discards the additional information in the labeled mean.

In practice, we usually need to estimate the optimal coefficient. Estimate $c_{Yf}$ from labeled pairs and $\sigma_f^2$ from the pooled $n + N$ predictions. Denote these estimates by $s_{Yf,L}$ and $s^2_{f,\mathrm{pool}}$. The mean version of PPI++ uses

$$\widehat{\lambda} = \frac{s_{Yf,L}}{(1 + n/N)s^2_{f,\mathrm{pool}}}.$$

For an exactly constant predictor set $\widehat{\lambda} = 0$. For an almost constant predictor, the denominator is small, so small changes in the estimated covariance can cause large changes in the coefficient.

The estimated SE is obtained from

$$\widehat{v}_{\widehat{\lambda}} = \frac{s^2_{Y - \widehat{\lambda}f,L}}{n} + \frac{\widehat{\lambda}^2 s^2_{f,U}}{N},$$

where the first sample variance is computed from the labeled adjusted outcomes. Use $\widehat{\mu}_{\widehat{\lambda}} \pm z_{1-\alpha/2}\sqrt{\widehat{v}_{\widehat{\lambda}}}$. For its large-sample justification, keep $f$ fixed, assume finite second moments and $\sigma_f^2 > 0$, and let $n, N \to \infty$ with $n/N \to r \in [0, \infty)$. If $\widehat{\lambda}$ consistently estimates the optimal coefficient and the limiting variance is positive, the tuned estimator has the same first-order normal limit as the estimator using the known optimal coefficient. The interval then has asymptotic coverage $1 - \alpha$. To see why estimating the coefficient has no first-order effect, use the exact identity

$$\widehat{\mu}_{\widehat{\lambda}} - \widehat{\mu}_{\lambda^\star} = (\widehat{\lambda} - \lambda^\star)(\bar{f}_U - \bar{f}_L).$$

The weight-estimation error tends to zero. It multiplies a difference of prediction averages that is already close to zero, with fluctuations of order $n^{-1/2}$ when $n/N$ stays bounded. Their product is therefore smaller than the main estimation error.

**Remark 13.9: Known and estimated coefficients**

The variance comparison in Proposition 13.7 uses a *known* optimal coefficient; this is often called an oracle comparison. In a small labeled sample, estimating that coefficient can introduce enough error to offset the variance reduction. The resulting estimator need not be exactly unbiased, even though it is unbiased for every fixed coefficient. The asymptotic comparison therefore does not imply improvement in every dataset. Mani et al. [132] give finite-sample analyses demonstrating this cost of tuning.

Figure 13.1 compares these choices. Useful and offset predictions give the same PPI intervals because the constant offset cancels. With $n = 200$, the mean width is about 0.229 for both basic and tuned PPI, compared with 0.354 using labels alone. Their empirical coverage is close to 95 percent in this experiment. Uninformative predictions instead increase basic PPI's mean width to about 0.453; tuning brings it back to about 0.354.

The orange prediction-only intervals remain narrow in the offset, uninformative, and misleading settings, but have essentially zero coverage because their centers estimate the wrong mean. In the uninformative case, $f(X, Z) = 1.6 + Z$ is independent of $Y$. Basic PPI corrects the bias but adds prediction noise; tuning reduces its coefficient toward zero and recovers approximately the labeled-only precision.

Readers following the core route can now continue with Section 7. The next three sections extend the correction idea to other targets and to predictors learned from available labels.

## 4. Estimating Equations

**Route: Advanced**

A proportion is already a mean of indicators, so the preceding calculations apply directly to prevalence. To extend the idea, think of estimation as choosing a parameter so that an average discrepancy equals zero. For a mean,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta) = 0 \qquad \Longleftrightarrow \qquad \theta = \bar{Y}_L.$$

A regression coefficient or a quantile uses a different discrepancy. PPI corrects that discrepancy before solving the equation for the parameter.

Let $\theta$ be a fixed-dimensional parameter, and let $\psi_\theta(X, Y)$ be a scalar or vector function with

$$\mathbb{E}[\psi_{\theta_0}(X, Y)] = 0$$

at the target $\theta_0$. Such a function is called an *estimating function*. For a mean, $\psi_\theta(X, Y) = Y - \theta$. If $Y$ is replaced by $f(X)$, the resulting function can be averaged over the unlabeled sample. As before, a fixed coefficient $\lambda$ controls its contribution. The labeled sample estimates the difference between the function evaluated at the observed outcome and $\lambda$ times the function evaluated at its prediction. Combining these averages gives

$$0 = \frac{\lambda}{N}\sum_{j=1}^{N}\psi_\theta(\widetilde{X}_j, f(\widetilde{X}_j)) + \frac{1}{n}\sum_{i=1}^{n}\{\psi_\theta(X_i, Y_i) - \lambda\psi_\theta(X_i, f(X_i))\}. \tag{40}$$

For fixed $\theta$ and $\lambda$, the expectation of its right-hand side is $\mathbb{E}[\psi_\theta(X, Y)]$. Prediction terms cancel because they have the same distribution in the two samples.

A solution of a nonlinear empirical equation need not itself be exactly unbiased. The cancellation identifies the population equation. For its solution to be consistent and approximately normal, the population equation must distinguish the target from other parameter values.

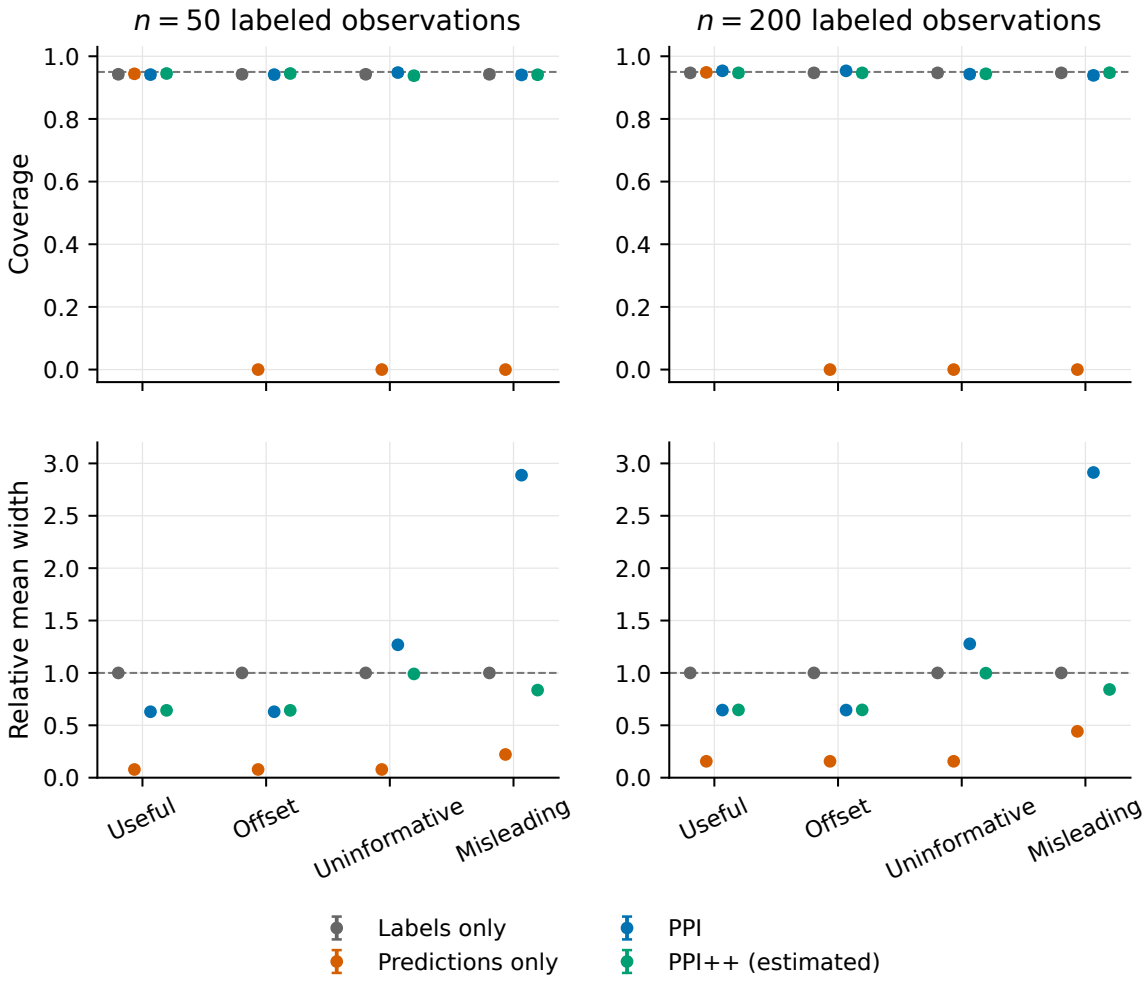



FIGURE 13.1. PPI precision depends on prediction quality; prediction bias can invalidate prediction-only intervals. Independent labeled and unlabeled samples follow $Y = 1 + X + 0.8\varepsilon$, with independent standard normal $X, Z, \varepsilon$. Both $X$ and $Z$ are observed features. Predictions are $1 + X$, $1.6 + X$, $1.6 + Z$, or $2 - 2X + 2Z$, respectively. The "uninformative" predictor is independent of $Y$ and has mean 1.6 rather than the target $\mu = 1$. Each setting uses 2,000 repetitions, $N = 5{,}000$, and the displayed labeled sample size. Intervals use normal 95 percent critical values; PPI++ estimates its coefficient. Lower panels show mean width relative to labeled-only inference; error bars show two Monte Carlo SEs. Shared simulated samples make the labeled-only points coincide across settings. Assess width together with coverage.

Appropriate smoothness and moment conditions, together with a nonsingular derivative at the target, are also needed.

First, consider the least-squares association with forest loss. Let $z(X) \in \mathbb{R}^d$ contain a small, prespecified set of regressors, including an intercept. Assume $H = \mathbb{E}[z(X)z(X)^\top]$ is nonsingular. Define

$$\theta_0 = \arg\min_\theta \mathbb{E}[(Y - z(X)^\top\theta)^2], \qquad \psi_\theta(X, Y) = z(X)(Y - z(X)^\top\theta).$$

The prediction model may use many more features than these $d$ regressors. The target is the best linear approximation to $Y$ in population mean squared error, whether or not the conditional mean is linear.

With $\lambda = 1$, equation (40) simplifies to

$$\left\{\frac{1}{N}\sum_j \widetilde{z}_j\widetilde{z}_j^\top\right\}\widehat{\theta} = \frac{1}{N}\sum_j \widetilde{z}_j f(\widetilde{X}_j) + \frac{1}{n}\sum_i z_i\{Y_i - f(X_i)\},$$

where $z_i = z(X_i)$ and $\widetilde{z}_j = z(\widetilde{X}_j)$. Each component of $z_i$ multiplies the prediction error before averaging. The intercept uses the average error; a slope also uses how that error varies with its regressor. Correcting only the overall mean error therefore does not generally correct a slope estimate.

Next, we derive standard errors. For a fixed $\lambda$, define $a_\theta(X,Y) = \psi_\theta(X,Y) - \lambda\psi_\theta(X,f(X))$ and $b_\theta(X) = \lambda\psi_\theta(X,f(X))$. These are the labeled correction and unlabeled prediction terms in (40), so that equation is the sum of their sample averages. More explicitly, compute $a_i = a_{\widehat{\theta}}(X_i,Y_i)$ on the labeled sample and $b_j = b_{\widehat{\theta}}(\widetilde{X}_j)$ on the unlabeled sample. With $\bar{a} = n^{-1}\sum_i a_i$ and $\bar{b} = N^{-1}\sum_j b_j$, their sample covariance matrices are

$$\widehat{\mathrm{Cov}}_L(a_{\widehat{\theta}}) = \frac{1}{n-1}\sum_i (a_i-\bar{a})(a_i-\bar{a})^\top, \qquad \widehat{\mathrm{Cov}}_U(b_{\widehat{\theta}}) = \frac{1}{N-1}\sum_j (b_j-\bar{b})(b_j-\bar{b})^\top.$$

Let $\widehat{H}$ consistently estimate the negative derivative of the population equation at its target,

$$H_0 = -\left.\frac{\partial}{\partial\theta^\top}\mathbb{E}[\psi_\theta(X,Y)]\right|_{\theta=\theta_0}.$$

For this regression, $H_0 = H$, and the empirical average $\widehat{H} = N^{-1}\sum_j \widetilde{z}_j\widetilde{z}_j^\top$ estimates it. Near $\theta_0$, multiplying the error in the corrected equation by $H_0^{-1}$ gives the first-order parameter error. At $\theta_0$, the two sample averages are independent, so their covariance matrices add. Estimating these quantities at $\widehat{\theta}$ gives

$$\widehat{\mathrm{Cov}}(\widehat{\theta}) = \widehat{H}^{-1}\left\{\frac{\widehat{\mathrm{Cov}}_L(a_{\widehat{\theta}})}{n} + \frac{\widehat{\mathrm{Cov}}_U(b_{\widehat{\theta}})}{N}\right\}\widehat{H}^{-\top}.$$

The middle term measures variability of the corrected equation. The outer matrices convert error in the equation into error in the coefficients. This arrangement is called a sandwich variance formula, with two independent sources of uncertainty. It requires fixed $d$, a consistent solution of the estimating equation, finite second moments of $a_{\theta_0}$ and $b_{\theta_0}$, and the local regularity conditions stated above. For smooth loss-based estimating equations, PPI++ also permits a consistently estimated weight under its regularity conditions [6]. This is not a high-dimensional coefficient guarantee of the kind developed in Chapter 11.

For quantiles, we instead correct counts below a threshold. For a continuous outcome and a quantile level $\tau \in (0,1)$, a quantile $\theta_0$ solves $\mathbb{P}(Y \le \theta_0) = \tau$. The relevant estimating function is $\mathbf{1}\{Y \le \theta\} - \tau$, where the indicator is 1 when $Y \le \theta$ and 0 otherwise. Basic PPI therefore corrects a predicted CDF:

$$\widehat{F}_{\mathrm{corr}}(t) = \frac{1}{N}\sum_j \mathbf{1}\{f(\widetilde{X}_j) \le t\} + \frac{1}{n}\sum_i \big(\mathbf{1}\{Y_i \le t\} - \mathbf{1}\{f(X_i) \le t\}\big).$$

The first average estimates the fraction of predictions below $t$. The second corrects the difference between that fraction and the fraction of outcomes below $t$. For each fixed $t$, the expectation of their sum is $\mathbb{P}(Y \le t)$. A quantile procedure uses this corrected equation or inverts tests of it;

quantile regularity, such as positive density near a unique target quantile, matters for normal approximations. The corrected empirical function need not be monotone in a finite sample. Simply subtracting the average prediction error from a predicted quantile has no corresponding general justification.

## 5. Cross-Prediction

**Route: Advanced**

Next, we allow the available labels to be used for training as well as inference. A model that memorizes those labels has zero training residuals even when its errors on new parcels are substantial. Using those training residuals as the PPI correction would miss the error that needs to be corrected.

The simplest remedy is sample splitting: train on one subset, freeze the model, and use a different subset for correction. Conditional on the training subset, the basic theory applies to the correction subset and the independent unlabeled data. Its cost is that only some of the labels remain for correction. All preprocessing, feature selection, and tuning that use outcomes belong inside the training step.

Cross-prediction reuses labels through held-out predictions [206]. It uses the same held-out folding pattern as CV. Here the held-out residuals correct a population estimate, rather than merely assess prediction performance. For clarity, divide the labeled observations into $K$ equal folds $I_1, \ldots, I_K$, with fixed $K$, and train $f^{(-k)}$ using all labeled folds except $I_k$. The mean estimator is

$$\widehat{\mu}_{\text{cross}} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{j=1}^{N} f^{(-k)}(\widetilde{X}_j) + \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \{Y_i - f^{(-k)}(X_i)\}. \tag{41}$$

Each label supplies a correction only for a model that did not train on it. The prediction term averages the same fold-specific models on the unlabeled sample, so the two sides of the correction use matching prediction rules. The left panel of Figure 13.2 shows the folding pattern.

**Algorithm 13.1: Cross-prediction for a mean**

1. Fix the folds before examining outcome-dependent fitting results.
2. For each fold, fit the model, including preprocessing and tuning, on the other folds. Predict both its held-out fold and the independent unlabeled sample.
3. Average the held-out residuals and the fold-averaged unlabeled predictions as in (41).
4. Estimate the two variance components in the interval below. Its justification requires the stated stability condition and consistent variance estimates; the fitted folds are not independent experiments.

For confidence intervals, dependence matters. Although every outcome is evaluated by a model trained without it, the training sets overlap and all models predict the same unlabeled observations. The fold estimates are therefore dependent. Their covariance cannot be omitted when estimating the variance of their average.

To study this dependence, let $\bar{f}_n(x) = \mathbb{E}_{\text{train}}[f^{(-1)}(x)]$, where the expectation averages over fresh training samples of size $n - n/K$. This is a theoretical comparison rule, not the average of

the $K$ models fitted to the observed data. The stability condition of Zrnic and Candès [206] is

$$\mathbb{E}_{\text{train}}\left[\sqrt{K\ \mathrm{Var}_X\left(f^{(-1)}(X) - \bar{f}_n(X) \mid f^{(-1)}\right)}\right] \longrightarrow 0.$$

Here $X$ is a new feature vector, independent of the training data. The conditional variance measures how the difference between the fitted and average rules varies across such feature vectors. To see why it enters, write $d_k = f^{(-k)} - \bar{f}_n$. The difference between the original estimator and the estimator using $\bar{f}_n$ is an average of terms

$$\frac{1}{N}\sum_j d_k(\widetilde{X}_j) - \frac{1}{|I_k|}\sum_{i\in I_k} d_k(X_i).$$

Conditional on the training data for fold $k$, both averages have the same mean. Stability makes their remaining fluctuation negligible on the $n^{-1/2}$ scale. The estimator then behaves like a sum of two independent averages using $\bar{f}_n$. The factor $K$ explains why allowing the number of folds to grow requires faster decay of the fitting variability.

Let $\widehat{\sigma}^2_{\text{pred}}$ and $\widehat{\sigma}^2_{\text{res}}$ consistently estimate $\mathrm{Var}(\bar{f}_n(X))$ and $\mathrm{Var}(Y - \bar{f}_n(X))$, respectively. For a concrete sufficient condition for normal approximation, assume $\mathbb{E}[Y^2] < \infty$ and that, for some fixed square-integrable function $f_\infty$,

$$\mathbb{E}[(\bar{f}_n(X) - f_\infty(X))^2] \longrightarrow 0.$$

Assume also stability, $n, N \to \infty$ with $n/N$ tending to a finite limit, and positive limits for both variances. The mean-square convergence lets us replace $\bar{f}_n$ by $f_\infty$: the two averages of their difference have equal expectations, and their centered difference has variance $o(1/n)$. The ordinary CLT then applies to the two independent averages based on $f_\infty$. With the consistent variance estimates, an asymptotic $1-\alpha$ interval is

$$\widehat{\mu}_{\text{cross}} \ \pm\ z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{\text{res}}}{n} + \frac{\widehat{\sigma}^2_{\text{pred}}}{N}}.$$

Zrnic and Candès [206] propose bootstrap fits to estimate these components. Predictions on unlabeled observations are averaged over those fits. An *out-of-bag* residual uses an observation omitted from the corresponding bootstrap training sample. This keeps its outcome out of the fit used to predict it.

## 6. Recalibration

**Route: Advanced**

In Section 3, we changed the contribution of the prediction $f(X)$ by multiplying it by a coefficient $\lambda$. We can instead learn a function that converts $f(X)$ into a better prediction of $Y$. This is called *recalibration*. For example, among parcels with a forest-loss score of 0.8, the observed loss rate might be only 0.6. A fitted curve can account for such a relationship when a single multiplier cannot. The prediction is sometimes called a *surrogate*: an inexpensive quantity used in place of a measured outcome.

**Example 13.10: An informative score on the wrong scale**

Suppose a model supplies a score $S \in [0,1]$, while the actual probability of forest loss among parcels with score $s$ is

$$\mathbb{P}(Y = 1 \mid S = s) = 0.10 + 0.80s^2.$$

The score ranks risk correctly, but it is not a calibrated probability. For scores $0.2, 0.5, 0.8$, the probabilities are $0.132, 0.300, 0.612$. A scalar multiplier cannot reproduce this curved relationship. Learning the curve on separate labeled data can provide a more useful surrogate. In this split-sample construction, its fitted values are corrected using labels not used to fit the curve. The right panel of Figure 13.2 illustrates the fitted relationship.

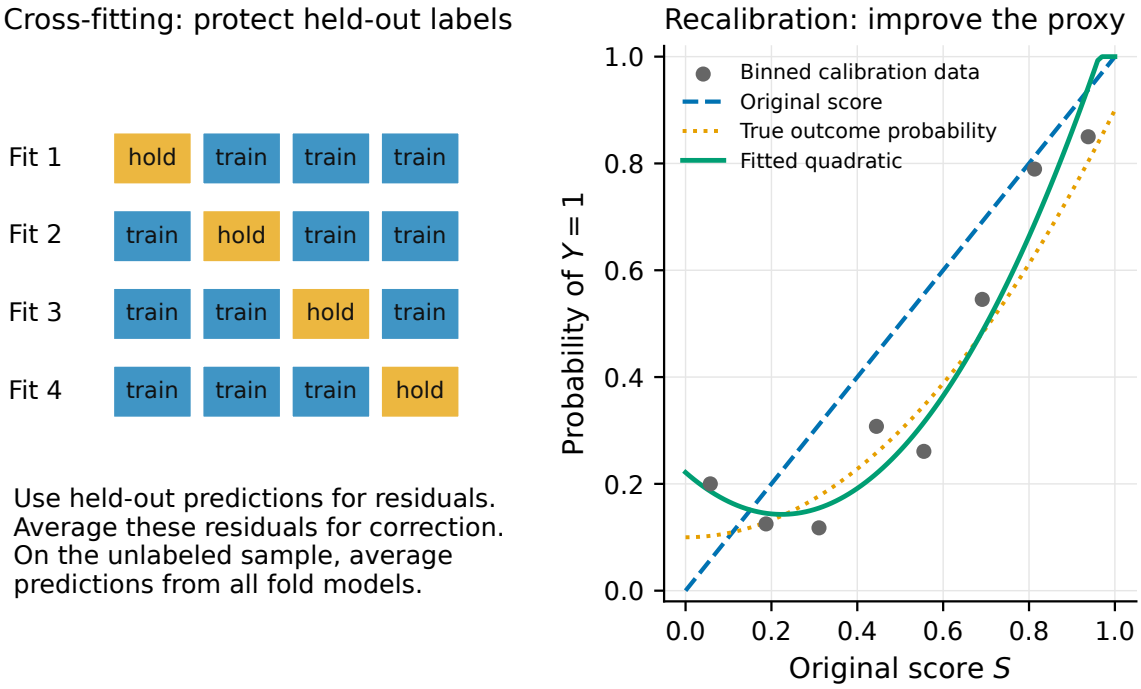



FIGURE 13.2. Cross-fitting and recalibration. Left: each outcome is predicted by a model trained on the other folds, and can train models used for other observations. Right: recalibration estimates the outcome probability given the score, here $0.10 + 0.80s^2$. The fitted quadratic uses 200 independent simulated calibration pairs with $S \sim \text{Unif}(0, 1)$ and binary outcomes having that success probability; points summarize eight score bins. The fitted curve is clipped to $[0, 1]$ for display. In both procedures, the residual correction uses outcomes held out from the corresponding prediction fit.

For mean estimation, we now replace $f$ by a fixed recalibrated function $h$. This preserves the identity $\mathbb{E}[h(X)] + \mathbb{E}[Y - h(X)] = \mu$. The conditional-mean predictor $\eta(X) = \mathbb{E}[Y \mid X]$ has a special role because

$$\mathbb{E}[(Y - h(X))^2] = \mathbb{E}[\text{Var}(Y \mid X)] + \mathbb{E}[(\eta(X) - h(X))^2].$$

This decomposition explains why learning a better surrogate can reduce residual mean squared error. It does not by itself minimize the entire finite-$N$ PPI variance, which also includes the prediction-average term. The coefficient $\lambda$ can still be chosen to reduce that total variance.

Recalibrated prediction-powered inference (RePPI) takes this idea beyond means [99]. The useful surrogate for a general parameter is related to the conditional expectation of its *estimating function*. A model trained to predict outcomes need not provide the best approximation to that

target-specific quantity. For a quantile, for example, one wants information about $\mathbb{P}(Y \le t \mid X)$, not only about $\mathbb{E}[Y \mid X]$. We next describe the three-fold construction for smooth loss-based estimating equations. Split the labels into equal folds $D_1, D_2, D_3$. Use $D_3$ to obtain an initial estimate $\tilde{\theta}$. On $D_2$, compute the responses $\psi_{\tilde{\theta}}(X_i, Y_i)$ and fit a function $\widehat{g}(X)$ to predict them from the available features and model predictions. On $D_1$, estimate its coefficient:

$$\widehat{M} = \widehat{\mathrm{Cov}}_{D_1}(\psi_{\tilde{\theta}}(X, Y), \widehat{g}(X))\widehat{\mathrm{Cov}}_{D_1}(\widehat{g}(X))^{-1}.$$

These are sample cross-covariance and covariance matrices on $D_1$; the second must be invertible. Then solve

$$0 = \frac{1}{|D_1|}\sum_{i \in D_1} \psi_\theta(X_i, Y_i) + \frac{N}{n+N}\widehat{M}\left\{\frac{1}{N}\sum_j \widehat{g}(\widetilde{X}_j) - \frac{1}{|D_1|}\sum_{i \in D_1}\widehat{g}(X_i)\right\}.$$

Here $n$ is the total labeled sample size. Rotate the three roles and average the resulting parameter estimates. Each correction fold uses a function fitted on another fold; its coefficient is estimated on the correction fold itself. For a mean, this reduces to a labeled mean plus a weighted difference of recalibrated prediction averages.

The asymptotic variance comparison in Ji et al. [99] requires the fitted function to converge in mean square to a fixed square-integrable function, nonsingular population curvature, and their smooth-loss regularity conditions. Under those assumptions the limiting variance is no greater than that of labeled-only estimation. The quantile example above explains which quantity to predict; it is not covered by this smooth-loss statement.

For mean estimation, a simpler alternative is to learn the recalibration on a separate training sample, freeze it, and apply ordinary PPI on fresh data. Residuals from the same data used to fit the curve can underestimate its error on new observations.

## 7. Choosing Which Outcomes to Measure

**Route: Core**

Thus far, the labeled sample was given. Now suppose we can choose which parcels to verify. If prediction errors vary with cloud cover, measuring more outcomes where errors are larger may improve precision.

The target remains the population prevalence. We first choose fixed sample sizes within groups, then purchase labels from a pool with specified probabilities. These designs require different variance calculations.

Let $G \in \{1, 2\}$ record a known group, such as clear versus cloudy images. Suppose the population group proportions $w_g = \mathbb{P}(G = g)$ are known and positive. Then

$$\mu = w_1\mu_1 + w_2\mu_2, \qquad \mu_g = \mathbb{E}[Y \mid G = g].$$

Collect independent random labeled samples of fixed sizes $n_g$ within each group, and independent unlabeled samples of sizes $N_g$. Using the same frozen predictor in both samples, a stratified estimator is

$$\widehat{\mu}_{\mathrm{strat}} = \sum_{g=1}^{2} w_g(\bar{f}_{U,g} + \overline{Y - f}_{L,g}).$$

The coefficients are the *population* proportions, not the proportions of the labeling budget assigned to the groups. The latter can deliberately be quite different.

For known weights and independent samples across groups, the variance is

$$\sum_{g=1}^{2} w_g^2\left(\frac{\sigma_{f,g}^2}{N_g} + \frac{\sigma_{R,g}^2}{n_g}\right),$$

Here $\sigma_{f,g}^2$ and $\sigma_{R,g}^2$ are the prediction and residual variances within group $g$. With $N_1, N_2$ fixed and equal labeling costs, we choose $n_1, n_2$ to minimize $\sum_g w_g^2 \sigma_{R,g}^2 / n_g$ subject to $n_1 + n_2 = n$. For positive residual standard deviations, Cauchy–Schwarz gives

$$\left( \sum_{g=1}^{2} w_g \sigma_{R,g} \right)^2 \leq n \sum_{g=1}^{2} \frac{w_g^2 \sigma_{R,g}^2}{n_g}.$$

Equality holds at

$$n_g = n \, \frac{w_g \sigma_{R,g}}{w_1 \sigma_{R,1} + w_2 \sigma_{R,2}}.$$

Thus a larger group receives more labels, as does a group whose prediction errors vary more. The formula treats sample sizes as positive real numbers; actual allocations must be rounded.

**Example 13.11: A fixed budget across two groups**

Let $w_1 = w_2 = 1/2$, $\sigma_{R,1} = 0.10$, and $\sigma_{R,2} = 0.30$. Both allocations below use the same unlabeled samples and prediction averages. With 200 labels, equal allocation gives labeled variance

$$\frac{1}{4} \left( \frac{0.10^2}{100} + \frac{0.30^2}{100} \right) = 0.00025.$$

Allocating 50 labels to group 1 and 150 to group 2 instead gives

$$\frac{1}{4} \left( \frac{0.10^2}{50} + \frac{0.30^2}{150} \right) = 0.00020.$$

The labeled variance falls by 20 percent. The total variances are $C_U + 0.00025$ and $C_U + 0.00020$, respectively, where $C_U = \frac{1}{4}(\sigma_{f,1}^2/N_1 + \sigma_{f,2}^2/N_2)$ is the unchanged unlabeled contribution. The calculation assumes the residual standard deviations are known. In practice, a small separate labeled sample can be collected first to estimate them and choose the allocation. This preliminary sample is called a *pilot sample*.

Next, we purchase labels from the prediction pool itself, rather than from a separate independent sample. Consider $N_{\text{pool}}$ i.i.d. underlying pairs $(X_i, Y_i)$ from the target population. All feature vectors are observed, but we purchase only some of the outcomes. Their number is random, unlike the fixed group counts in the preceding example.

For each unit, draw an indicator $A_i$ that equals one if its label is purchased. Conditional on $X_i$, let

$$\mathbb{P}(A_i = 1 \mid X_i, Y_i) = \pi(X_i), \qquad 0 < \pi(X_i) \leq 1.$$

Thus the choice can depend on the observed features but not on the unseen outcome after those features are fixed. Indicators are conditionally independent across units. Write $\pi_i = \pi(X_i)$. The model $f$ and rule $\pi$ are frozen, using a separate pilot sample if needed. We condition on that pilot when discussing the main sample.

The active estimator is

$$\widehat{\mu}_{\text{active}} = \frac{1}{N_{\text{pool}}} \sum_{i=1}^{N_{\text{pool}}} \left[ f(X_i) + \frac{A_i}{\pi_i} \{ Y_i - f(X_i) \} \right]. \tag{42}$$

For an unpurchased outcome, the correction is defined to be zero and does not require knowing $Y_i$. For a purchased outcome, the correction is weighted by the inverse of its probability of being purchased. This is *inverse-probability weighting* (IPW) of the residuals.

**Proposition 13.12: Why selective labeling needs weights**

Under the pool design above, assume the summands in (42) are integrable. Then $\mathbb{E}[\widehat{\mu}_{\text{active}}] = \mu$, even if $f$ is incorrect. If the summands have finite variance, the estimator's variance is

$$\text{Var}(\widehat{\mu}_{\text{active}}) = \frac{1}{N_{\text{pool}}}\left[\text{Var}(Y) + \mathbb{E}\left\{(Y - f(X))^2\left(\frac{1}{\pi(X)} - 1\right)\right\}\right]. \tag{43}$$

**Proof of Proposition 13.12**

Condition on $(X, Y)$. The expected inverse-probability factor is $\mathbb{E}[A/\pi(X) \mid X, Y] = 1$, so the conditional expectation of one summand is $f(X) + Y - f(X) = Y$. Its conditional variance is $(Y - f(X))^2(1/\pi(X) - 1)$, because $\text{Var}(A \mid X, Y) = \pi(X)(1 - \pi(X))$. The law of total variance adds this expected conditional variance to $\text{Var}(Y)$. The independent pool units then give the factor $1/N_{\text{pool}}$. □

The proposition averages over pool sampling and label purchases, with target $\mu = \mathbb{E}[Y]$. Conditional on the complete pool, the target is its mean $N_{\text{pool}}^{-1}\sum_i Y_i$. We use the population target below.

To form a large-sample interval, let $T_i$ denote a summand in (42). For a fixed rule with $\pi(x) \geq \pi_{\min} > 0$, finite second moments, and $\text{Var}(T_i) > 0$, the i.i.d. CLT gives

$$\bar{T} \pm z_{1-\alpha/2}\frac{s_T}{\sqrt{N_{\text{pool}}}}, \qquad s_T^2 = \frac{1}{N_{\text{pool}} - 1}\sum_i (T_i - \bar{T})^2.$$

Next, we minimize variance at a fixed expected labeling cost. Let

$$\text{MSE}_f(x) = \mathbb{E}[(Y - f(X))^2 \mid X = x], \qquad b = \mathbb{E}[\pi(X)]$$

denote the conditional mean squared prediction error and expected labeling fraction. The expected main-sample cost is $bN_{\text{pool}}$ labels. The part of (43) affected by the rule is $\mathbb{E}[\text{MSE}_f(X)/\pi(X)]$. Assume $\text{MSE}_f(X) > 0$ almost surely and $0 < b < 1$. Cauchy–Schwarz gives

$$\left(\mathbb{E}\sqrt{\text{MSE}_f(X)}\right)^2 \leq \mathbb{E}\left[\frac{\text{MSE}_f(X)}{\pi(X)}\right]\mathbb{E}[\pi(X)] = b\,\mathbb{E}\left[\frac{\text{MSE}_f(X)}{\pi(X)}\right].$$

Equality holds when $\pi(x)$ is proportional to $\sqrt{\text{MSE}_f(x)}$, provided these probabilities are at most one. Thus more labels go to feature values with larger squared prediction errors.

Because a probability cannot exceed one, the constrained solution is

$$\pi^\star(x) = \min\{1, c\sqrt{\text{MSE}_f(x)}\},$$

where $c$ is chosen to meet the budget. A positive minimum probability can also be imposed, as in the implementation below.

We next estimate these unknown errors from an independent pilot. Use one purchase probability per group and collect $m_g$ pilot labels in group $g$, using a predictor fixed before the pilot, and compute

$$\widehat{u}_g = \left\{\frac{1}{m_g}\sum_{i \in \mathcal{P}_g}(Y_i - f(X_i))^2\right\}^{1/2},$$

where $\mathcal{P}_g$ indexes its pilot observations. This estimates residual RMS, including bias. It determines where to collect labels; the final standard error will use the corrected values $T_i$.

**Algorithm 13.2: A pilot followed by active labeling**

1. Fix the predictor $f$, known group proportions $w_g$, and pool size $N_{\text{pool}}$. From a total expected budget of $B$ labels, reserve $m = m_1 + m_2$ for an independent pilot. Set $b = (B - m)/N_{\text{pool}}$ and choose $0 < \pi_{\min} \le b < 1$.
2. Collect $m_g$ pilot labels in group $g$. Compute the root mean squared residual $\widehat{u}_g$ defined above.
3. Choose $c$ so that the group probabilities $\widehat{\pi}_g = \max\{\pi_{\min}, \min\{1, c\widehat{u}_g\}\}$ satisfy $\sum_g w_g \widehat{\pi}_g = b$. The left side increases with $c$, so bisection finds the solution. If neither bound applies, use $\widehat{\pi}_g = b\widehat{u}_g / \sum_h w_h \widehat{u}_h$. Use uniform probabilities if a pilot RMS is zero. Freeze the chosen rule.
4. For each unit in a fresh pool, purchase its label independently with probability $\widehat{\pi}_{G_i}$. Record that probability and whether the label was purchased.
5. For every pool unit, set $T_i = f(X_i)$ if its label was not purchased, and $T_i = f(X_i) + \{Y_i - f(X_i)\}/\widehat{\pi}_{G_i}$ if it was. Report $\bar{T} = N_{\text{pool}}^{-1} \sum_i T_i$.
6. Compute the sample variance $s_T^2$ from all $N_{\text{pool}}$ values, including unpurchased units. Use $\bar{T} \pm z_{1-\alpha/2} s_T / \sqrt{N_{\text{pool}}}$ under the population-mean CLT assumptions above. Report both the expected and realized numbers of purchased labels.

To illustrate the gain from active sampling, we now use a continuous outcome. In Figure 13.3, the two groups are equally common and $Y = f_G + \sigma_G \varepsilon$, where $f_1 = 0$, $f_2 = 1$, and $\varepsilon$ is independent standard normal noise. Thus $\mu = 0.5$. The predictor is $f_G$; the residual standard deviations are either $(1, 1)$ or $(0.2, 1)$.

The total expected budget is 1,000 labels. Active sampling uses 40 pilot labels per group and 920 labels on average from a fresh pool of 5,000 units, so $b = 920/5{,}000 = 0.184$. For example, suppose the pilot gives $\widehat{u}_1 = 0.2$ and $\widehat{u}_2 = 1$. With $\pi_{\min} = 0.005$, neither bound applies, and

$$\widehat{\pi}_1 = \frac{0.184(0.2)}{(0.2 + 1)/2} \approx 0.0613, \qquad \widehat{\pi}_2 = \frac{0.184(1)}{(0.2 + 1)/2} \approx 0.3067.$$

More labels go to the second group because its prediction errors vary more. Uniform sampling uses no pilot and spends all 1,000 expected labels on the main sample, with purchase probability 0.2. The pilot labels choose the rule and are excluded from the final mean. Purchases are independent Bernoulli draws, so their count is random even though its expectation is fixed. The earlier fixed-count design uses residual standard deviations; this design uses RMS because random counts also make the average residual contribute to the variance. For example, constant residuals have zero standard deviation but still produce random weighted corrections when labels are purchased at random.

Next, we compare uniform sampling with the learned rule.

When the residual standard deviations are $(0.2, 1)$, the pilot identifies where additional labels are most useful. Mean interval width falls from 0.0935 under uniform sampling to 0.0824 under active sampling, a reduction of about 12 percent after counting the pilot cost. Empirical coverage is 95.2 and 94.5 percent, respectively, close to the nominal 95 percent level.

With equal residual standard deviations, uniform allocation is already appropriate. The pilot then uses labels without improving the allocation: mean width increases from 0.1269 to 0.1325, with empirical coverage of 95.1 and 94.6 percent. Active sampling helps when the differences in prediction-error variability are large enough to repay the cost of learning where to collect labels.

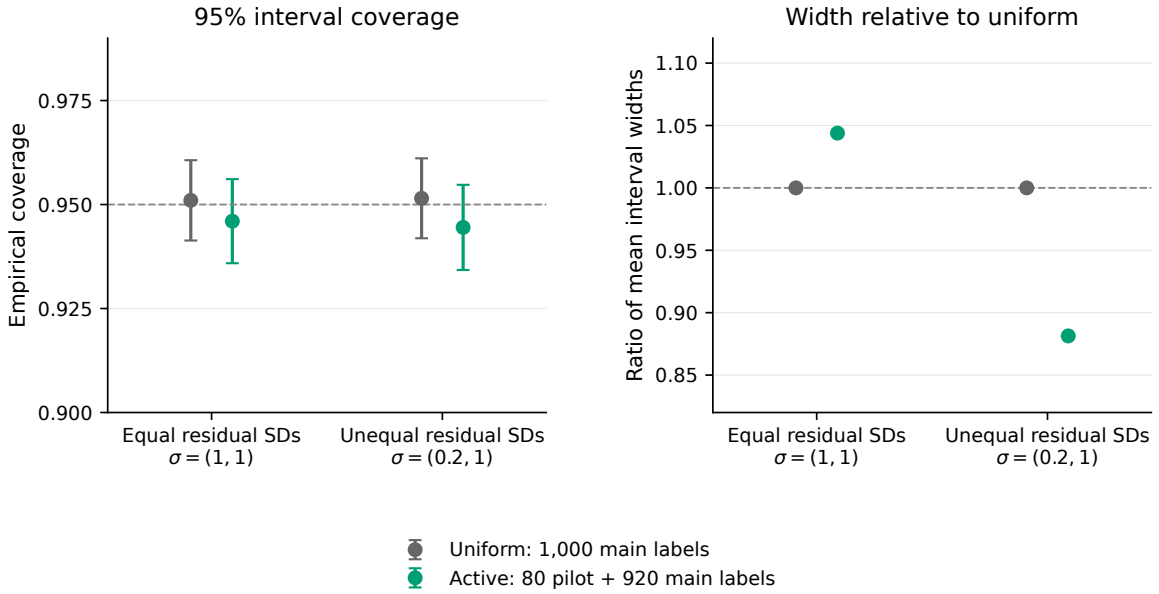



FIGURE 13.3. Uniform and active sampling at the same total expected cost. Both estimate $\mu = 0.5$ under $Y = f_G + \sigma_G\varepsilon$, with equally common groups, $(f_1, f_2) = (0, 1)$, and independent standard normal $\varepsilon$. The two settings differ only in the residual standard deviations. Each of 2,000 repetitions uses a fresh pool of 5,000 units. Uniform sampling uses 1,000 expected labels with no pilot. Active sampling uses 80 independent pilot labels and 920 expected main-sample labels; its probabilities use the pilot RMS with floor 0.005. Both use the prediction-corrected estimator and normal 95 percent intervals based on the sample variance of $T_i$. Error bars show two Monte Carlo SEs; width ratios use paired repetitions.

An inaccurate pilot can assign too few labels to a group with large errors. Dividing its few observed residuals by small purchase probabilities then increases variance. One practical safeguard, already used by Zrnic and Candès [205], mixes a proposed rule with uniform acquisition:

$$\pi_\gamma(x) = (1-\gamma)b + \gamma\pi_{\text{proposal}}(x), \qquad 0 \le \gamma \le 1.$$

If the proposal has mean budget $b$, so does the mixture. The values $\gamma = 0$ and $\gamma = 1$ give uniform sampling and the proposed rule, respectively; $\gamma = 1/2$ averages their probabilities. This choice is made before buying labels. A unit is purchased with probability $\pi_\gamma(X_i)$, and that same probability is used in its inverse-probability correction. Li et al. [124] study data-based choices of $\gamma$.

**Example 13.13: Human ratings with unequal sampling**

Suppose we want the proportion of question–answer pairs that human raters would accept. A pool of 1,000 pairs drawn from the target population contains 500 from each of two topics. An automated judge assigns score 0.80 to every pair.

We purchase human ratings independently with probabilities 0.05 in topic 1 and 0.15 in topic 2, fixed before seeing these ratings. Suppose 25 and 75 ratings are purchased, with 20

and 45 accepted answers, respectively. Equation (42) gives

$$\widehat{\mu}_{\text{active}} = 0.80 + \frac{1}{1000}\left\{\frac{20 - 0.80(25)}{0.05} + \frac{45 - 0.80(75)}{0.15}\right\} = 0.70.$$

The correction is zero in topic 1 and negative in topic 2. Averaging the purchased ratings directly gives $65/100 = 0.65$. That average gives topic 2 three times as much weight because more of its labels were purchased. The inverse-probability correction accounts for this unequal sampling.

## 8. Assumptions in Plain Language

**Route: Core**

We now review the assumptions. For ordinary two-sample PPI, the labels must represent the same target population as the prediction sample. Correction labels from another region or period may have a different average residual. A poor predictor trained elsewhere can still be used if both inference samples are representative of the target population; this differs from transporting a correction learned in another population.

Independent training justifies the basic correction argument. When labels also train, tune, or recalibrate the model, the fitting and correction steps must be separated appropriately. Cross-fitting addresses the data flow, while its validity also depends on properties of the learner and variance estimator.

The displayed variances assume independent sampling units. Spatial blocks, clusters of respondents, or repeated evaluations require uncertainty estimates that respect those units. Large unlabeled samples do not make a small number of independent labeled clusters behave like a large i.i.d. sample. For rare outcomes, observing no errors in a small labeled sample is weak evidence that the residual variance is negligible.

In active inference, the selection mechanism is part of the experiment. A label must have a known positive chance of being acquired when its residual can matter. Unrecorded human discretion about which cases to inspect cannot in general be repaired by plugging in the intended probabilities afterward.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | Document the target and sampling unit; compare feature and stratum distributions across samples; inspect held-out residuals, acquisition probabilities, and the largest weighted corrections. |
| **What these data cannot show** | A feature comparison cannot establish equality of unobserved outcome laws. A small sample with no observed model errors cannot certify perfect prediction. Good coverage in a simulation is not a proof for the application. |
| **Sensitivity checks** | Compare with labeled-only inference; examine tuning and fold sensitivity; compare uniform and targeted allocations; report how results depend on rare, heavily weighted corrections. |

| | |
|---|---|
| **Safer fallback** | Collect additional representative labels, use a simpler fixed prediction rule, or use a variance calculation that accounts for the sampling design. A prediction-only interval is not a fallback for uncertainty about outcomes. |

Next, consider testing at a fixed analysis time. Excluding a prespecified null value from a $1-\alpha$ interval yields a level-$\alpha$ test, with the interval's finite-sample or asymptotic qualification.

If many outcomes, regions, or model comparisons are analyzed, individual coverage does not ensure simultaneous coverage. Similarly, applying BH to many corrected p-values requires the validity and dependence conditions from Chapter 5; sharing correction labels can induce dependence. Choosing the most favorable comparison after inspecting the estimates also requires accounting for selection, as in Chapter 12.

Finally, updating the acquisition rule over time is not the same as allowing arbitrary stopping. Sequential active methods can use predictable rules, meaning that the decision uses information available before the new outcome is revealed. Their fixed-time asymptotic intervals do not automatically become confidence sequences. Optional stopping needs an additional sequential construction, such as the evidence processes discussed in Chapter 8.

## 9. Bibliographic Notes and Further Directions

**Route: Core**

Angelopoulos et al. [3] introduced PPI. Their Amazon forest-loss application motivates the running example, whose numerical summaries here are hypothetical. Angelopoulos et al. [6] developed power tuning and efficient estimation, including regression methods that avoid direct inversion over a parameter grid.

The correction is related to auxiliary information in survey sampling, control variates, and missing-data estimation. Relevant precedents include inverse-probability estimation [87], augmented estimating equations [153], prediction-plus-residual survey estimators [31], and semi-supervised mean inference [198].

Zrnic and Candès [206] address learning predictors from available labels; Ji et al. [99] use recalibration to predict the estimating function for the target parameter. Miao et al. [137] develop another post-prediction framework with adjustments for precision. These methods differ in their targets, use of labeled data, and assumptions.

For bootstrap inference, Zrnic [204] describes a prediction-powered bootstrap, while Kluger et al. [105] treats imputed variables and nonuniform sampling. Resampling must preserve the dependence between outcomes and predictions and reflect the sampling units. Bootstrap validity still requires regularity conditions. These works are cited in their available preprint forms.

For design, Fisch et al. [60] study stratified hybrid evaluation, Zrnic and Candès [205] develop batch and sequential active inference, and Li et al. [124] address unreliable acquisition scores. Skalnes and Oberst [165] studies a fixed label count, a useful alternative when the budget cannot be random. For the same limiting inclusion policy and under the paper's conditions, its conditional Poisson design has no larger asymptotic covariance than independent Bernoulli acquisition.

Mani et al. [132] analyze the finite-sample cost of estimated tuning, even when its asymptotic behavior is favorable. For sequential inference, Csillag et al. [44] construct prediction-powered e-values. Their acquisition probabilities depend on predictable bounds on the e-value factors, preserving nonnegativity and the required expectation bound. An arbitrary residual correction can make an e-value negative and therefore does not provide the same guarantee.

## 10. Exercises

**Route: Core**

### Basic.

**Exercise 13.14: What does the prediction average estimate?**

A model assigns scores with mean 0.24 and standard deviation 0.20 to a population whose binary outcome has mean 0.20. For an independent sample of $N = 10{,}000$ scores, calculate the SE of the score average and its usual 95 percent normal half-width. Identify the parameter covered by that interval. Explain what happens to its coverage of the outcome mean as $N \to \infty$.

**Exercise 13.15: Correction by hand**

Four labeled observations have outcomes $(0, 0, 1, 1)$ and predictions $(0.1, 0.4, 0.3, 0.8)$. An independent unlabeled sample has mean prediction 0.35. Compute the labeled mean, mean labeled prediction, mean residual, and PPI estimate. Repeat after adding 0.2 to every prediction in both samples. Use the offset-cancellation observation following (36) as a check on your calculations. No interval calculation is requested for this small example.

**Exercise 13.16: Two sources of uncertainty**

Suppose $n = 100$, $N = 2{,}500$, $s_R = 0.30$, and $s_{f,U} = 0.50$. Compute the PPI SE. What fraction of its estimated variance comes from the prediction average? Recalculate with $N = 100$, holding the sample variances fixed. Explain why the first calculation does not justify always ignoring prediction-sample uncertainty.

### Intermediate.

**Exercise 13.17: The cost and benefit of a predictor**

Derive (38) directly from independence of the samples. Suppose $\mathrm{Var}(Y) = 1$, $\mathrm{Var}(f(X)) = 1$, $\mathrm{Cov}(Y, f(X)) = 0.4$, and $n/N = 0.1$. Compare the labeled-only, basic PPI, and optimally tuned variances, expressing them as multiples of $1/n$. Which estimator requires an unknown population coefficient in practice?

**Exercise 13.18: Perfect predictions and a degenerate limit**

Suppose $f(X) = Y$ almost surely with positive finite variance. Use the perfect-prediction example in Section 3, where the optimal coefficient is $N/(N+n)$, to calculate the resulting variance. Explain why the nondegeneracy assumption of Proposition 13.4 can fail for basic PPI when $n/N \to 0$, without implying that the estimator is invalid.

**Exercise 13.19: A memorizing predictor**

Let $X$ be continuous and independent of a binary $Y$ with mean 0.3. A learner predicts the observed label at each training feature vector and predicts zero at every other feature vector. Describe what happens if its training labels are reused for the PPI correction and the unlabeled sample is independent. Contrast this with correcting on fresh labeled observations. Which assumption of Proposition 13.3 fails in the first analysis?

**Exercise 13.20: Biased acquisition**

Two equally common groups have binary outcome probabilities 0.1 and 0.5. Labels are acquired independently across units and independently of the outcomes given group, with probabilities 0.1 and 0.3, respectively. Compute the population mean and the limiting unweighted mean among acquired labels. Compare these quantities with the target of the weighted estimator in Proposition 13.12.

**Exercise 13.21: Allocating a fixed number of labels**

For known positive group weights $w_g$ and residual standard deviations $\sigma_{R,g} > 0$, minimize $\sum_g w_g^2 \sigma_{R,g}^2 / n_g$ subject to $\sum_g n_g = n$, first treating the $n_g$ as positive real numbers. Derive the allocation rule using Cauchy–Schwarz or a Lagrange multiplier. Explain the effects of integer rounding and unknown residual variances.

## Computational.

**Exercise 13.22: Coverage as well as width**

Generate independent labeled and unlabeled data from $Y = 1 + X + 0.8\varepsilon$, with independent standard normal $X, Z, \varepsilon$, using $(X, Z)$ as the observed features. Compare predictions $1 + X$, $1.6 + X$, $1.6 + Z$, and $2 - 2X + 2Z$, using $n \in \{50, 200\}$ and $N = 5{,}000$. Over at least 2,000 repetitions, compare labeled-only, prediction-only, PPI, and estimated-weight PPI++ 95 percent normal intervals. Report bias, coverage, mean width, and Monte Carlo uncertainty. Check whether adding an offset changes PPI. Use your estimated coverage and its Monte Carlo SE to identify narrow prediction-only intervals that fail to attain nominal coverage.

**Exercise 13.23: Misleading acquisition scores**

Reproduce Figure 13.3 using equally common groups and $Y = f_G + \sigma_G \varepsilon$, with $(f_1, f_2) = (0, 1)$ and independent standard normal $\varepsilon$. Use residual standard deviations $(1, 1)$ and $(0.2, 1)$, a pool of 5,000, and 2,000 repetitions. Give uniform sampling all 1,000 expected labels in the main sample. For learned allocation, use 40 pilot labels per group and 920 expected main-sample labels, with probability floor 0.005. Compare coverage and mean width using Monte Carlo SEs. Then extend the comparison by swapping the two pilot RMS estimates and mixing that rule equally with uniform sampling at the same main-sample budget. Finally, examine the effect of ignoring the purchase probabilities. Report both pilot and main-sample costs for every method.

## Advanced.

**Exercise 13.24: An oracle acquisition rule**

For the pool design, suppose $\mathrm{MSE}_f(X) = \mathbb{E}[(Y - f(X))^2 \mid X]$ is positive and integrable. Use Cauchy–Schwarz to minimize $\mathbb{E}[\mathrm{MSE}_f(X)/\pi(X)]$ subject to $\mathbb{E}[\pi(X)] = b$, initially ignoring $\pi \le 1$. Derive the proportional-to-RMS rule; the capped and floored forms are given in Section 7. For binary outcomes, let $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. Derive

$$\mathrm{MSE}_f(x) = \eta(x)(1 - \eta(x)) + (\eta(x) - f(x))^2.$$

Give an example in which $f(x)$ is close to 0 or 1 but the conditional mean squared prediction error is large.

CHAPTER 14

# Applications and Research Frontiers

**Route: Research frontier**

*Rapidly evolving and preprint-based material in this chapter is current as of August 2026.*

Genomics and LLM evaluation both involve many statistical questions and strong dependence among the measurements. A genomic experiment may measure expression levels for thousands of genes or test millions of variants. An LLM may be evaluated on many benchmarks, audited for watermarks or contamination, and asked to make factual claims whose reliability must be calibrated.

In both applications, we need to define the null hypothesis carefully, choose the family of claims before inspecting the most interesting cases, account for dependence, and report a statement whose error rate is the one we actually want. Genomics shows how FDR, local FDR, knockoffs, and high-dimensional regression work in a mature scientific workflow. LLM applications show how exchangeability, p-values, multiple testing, and conformal prediction reappear in modern model evaluation and safety auditing.

## 1. A Shared Workflow

**Route: Core**

Figure 14.1 places statistical inference after preprocessing and before scientific follow-up in a genomic study. Quality control (QC) and normalization are allowed to use scientific knowledge, but the error rate should be attached to a family of hypotheses that has been defined before individual discoveries are interpreted.

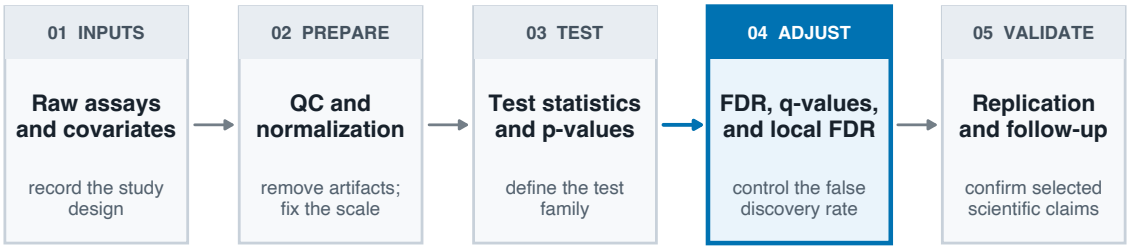



FIGURE 14.1. A genomic discovery pipeline. Multiplicity control follows QC and test-family definition; selected findings then require replication or biological validation.

An LLM evaluation pipeline has the same structure. Raw model outputs are converted into scores, the scores are compared across many tasks or prompts, and the analyst must decide whether a reported improvement, watermark signal, or contamination signal is strong enough to survive the relevant multiplicity. The rest of the chapter uses this analogy repeatedly.

## 2. Genomic Screening

**Route: Core**

Suppose $m$ genes or variants are tested. For each feature $i$ there is a null hypothesis $H_{0i}$, a test statistic $Z_i$, and a p-value $p_i$. In a differential-expression study, $H_{0i}$ might state that gene $i$ has the same mean expression in two conditions. In a GWAS, $H_{0i}$ might state that a variant has no marginal association with a phenotype after adjustment for pre-specified covariates.

With 20,000 genes, reporting only the smallest p-values hides how many opportunities there were to find an apparently interesting result. Apply the BH procedure from Chapter 5 to the prespecified testing family, provided its validity and dependence assumptions are justified [20, 54]. Report the family size, the target level $q$, the BH-adjusted p-values, and the complete discovery list. The selected genes are candidates for follow-up experiments; their statistical selection does not establish a biological mechanism.

Storey's q-values and local FDR can also help prioritize follow-up [57, 58, 168]. Their definitions and assumptions are developed in Chapter 6. In a genomic results table, label these quantities separately from BH-adjusted p-values, name the estimation method, and state the mixture model used for local FDR. This lets a reader distinguish the reported discovery rule from a fitted model's assessment of an individual gene. The next section considers the dependence assumptions needed to justify the discovery list.

## 3. Dependence, Linkage Disequilibrium, and Structured Discoveries

**Route: Core**

Genomic p-values are rarely independent. Nearby variants are correlated by LD, expression measurements share batch and pathway effects, and sample-level covariates can create broad shifts across many features. Dependence does not merely change a proof condition. It changes the shape of the evidence. A region with many correlated variants may produce many small p-values even when there is one underlying signal.

Several responses are possible.

- If the dependence is positive and satisfies the PRDS-type conditions from Chapter 6, BH remains valid.
- If the analyst wants protection under arbitrary dependence, the BY correction is valid but can be conservative.
- If the goal is to identify regions or pathways, the hypotheses should be defined at the region or pathway level rather than pretending that every correlated variant is a separate biological discovery.
- If the target is conditional variable importance, knockoffs or CRTs, as developed in Chapter 9, are more appropriate than marginal p-values.

The last point is particularly important in genetics. Let

$$X \in \mathbb{R}^{n \times p}$$

be a genotype or feature matrix and let $y$ be a phenotype. A marginal test of $X_j$ asks whether $X_j$ is associated with $y$. A conditional null asks whether

$$Y \perp X_j \mid X_{-j}.$$

These are different null hypotheses when features are correlated. Concretely, suppose variant $X_j$ is in tight LD with a causal variant $X_k$. Assume that, given all measured variants, the distribution of $Y$ depends only on $X_k$. Once the other variants, including $X_k$, are known, $X_j$ supplies no additional information about $Y$. Nevertheless, its correlation with $X_k$ can produce a

strong marginal association. Thus $X_j$ can satisfy the conditional null $Y \perp X_j \mid X_{-j}$ targeted by model-X knockoffs while failing the marginal null. Model-X knockoffs construct artificial variables $\widetilde{X}_j$ that serve as engineered null competitors for the original variables $X_j$. Feature statistics $W_j$ are built so that large positive values favor the original variable, large negative values favor its knockoff, and null signs are symmetric. The knockoff threshold estimates how many false positives are entering the selected set [9, 38].

Figure 14.2 separates two consequences of genomic dependence. The upper heatmap displays blocks of correlated variants created by LD, explaining why several marginal associations can reflect one underlying signal. The lower panel switches to the conditional target: knockoff contrasts $W_j$ compare each original feature with its knockoff competitor, and the marked threshold turns those contrasts into an FDR-controlled discovery set.

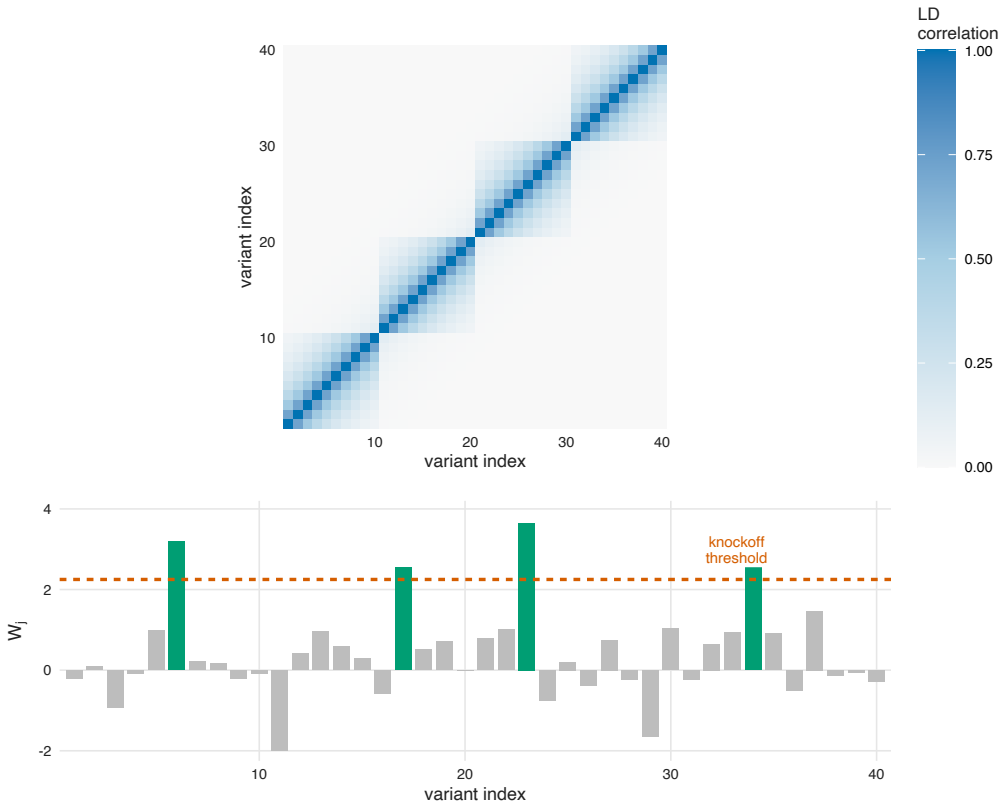



FIGURE 14.2. A schematic GWAS dependence problem. The heatmap displays an LD-like correlation structure among variants, with the color bar reporting LD correlation. The lower panel shows knockoff feature statistics $W_j$, where large positive values indicate evidence for conditional importance and the threshold is chosen to control FDR.

Debiased Lasso intervals from Chapter 11 can be used for pre-specified variants or low-dimensional contrasts. Knockoffs are more natural when the primary objective is discovery among many correlated features. Selective-inference ideas from Chapter 12 enter after a set of genes or variants has already been selected and the analyst wants intervals for selected targets. The methods are complementary because they attach error control to different inferential targets.

## 4. Benchmark Multiplicity for Language Models

**Route: Core**

LLMs are usually compared across many tasks, datasets, metrics, prompts, and decoding settings. Let $\Delta_{ab}$ be the mean performance difference between candidate model $a$ and a fixed baseline on task $b$. A comparison may use paired scores on the same questions, binary success indicators, or another prespecified outcome. If many comparisons are inspected, the largest estimated improvements receive many opportunities to benefit from noise, just as the smallest genomic p-values do.

**What is random?** Before computing an SE, ask what would change if the evaluation were repeated. Would we draw new questions, generate new responses to the same questions, ask different judges, or train a new model? These are different experiments. Table 14.1 connects each one to the performance claim it can support and the repetitions needed to measure uncertainty.

TABLE 14.1. What would change in a repeated benchmark evaluation?

| What could change? | What do we want to learn? | How do we measure uncertainty? |
|---|---|---|
| Nothing: items, responses, and scores are fixed | Average performance on this exact collection | Report the observed average. There is no sampling uncertainty for this fixed collection. |
| The questions or test items | Average performance on new items from the same population | Resample independent items, or independent groups of related items, keeping paired model results together. |
| Random generation on fixed items | Average performance over repeated responses to these same items | Generate independent repetitions within each item and preserve any pairing between models. |
| The annotators or judges | Average score across a specified population of judges | Account for shared judges and shared items; repeated scores from one judge are not independent judges. |
| The trained model | Average performance across independent runs of the training procedure | Repeat training. More test items cannot replace independent training runs. |

If two models answer the same questions, keep each question together with both models' scores when resampling. This preserves the pairing in each comparison. Ten generated answers to one question do not provide ten independent questions. They reveal how much the model's answers vary on that question. A claim about performance on new questions instead needs items that reasonably represent the intended question population and an analysis that respects their dependence. If both questions and generated answers vary, both sources of variation must be included.

**Example 14.1: Worked benchmark evaluation: 30 predeclared comparisons**

Figure 14.3 illustrates five candidate models compared with a fixed baseline on six tasks. To make the calculation explicit, this simulation generates a separate, independent dataset for each of the 30 comparisons. Each dataset contains 60 paired item differences, with three random generation repetitions per item. It does not simulate the dependence created by reusing the same questions and baseline responses across models.

For model $a$, task $b$, item $i$, and repetition $r$, generate the candidate-minus-baseline score difference as

$$d_{abir} = \Delta_{ab} + u_{abi} + \varepsilon_{abir}, \qquad u_{abi} \sim N(0, 0.11^2), \quad \varepsilon_{abir} \sim N(0, 0.07^2).$$

All item effects and errors are independent. The three repetitions on one item share $u_{abi}$, so they are correlated. First average within each item, $\bar{d}_{abi} = 3^{-1}\sum_{r=1}^{3} d_{abir}$. The 60 item averages are independent Gaussian observations. Their sample mean estimates $\Delta_{ab}$, and their sample standard deviation divided by $\sqrt{60}$ is its SE. Test each one-sided null $H_{ab} : \Delta_{ab} \leq 0$ with the upper tail of a $t_{59}$ distribution, evaluated at the estimated mean divided by its SE. Under $\Delta_{ab} = 0$, this is an exact t-test; for $\Delta_{ab} < 0$, its rejection probability is no larger.

The full family of 30 comparisons is declared before the simulation, and BH is applied once at $q = 0.10$. The valid, independent p-values in this simulation satisfy the assumptions of the BH result in Chapter 5. BH identifies 23 comparisons. Each heatmap cell shows an estimated mean difference, and circles mark the selected cells. These tests concern the population generating the item differences. On a fixed collection with fixed responses and scores, the observed averages can instead be reported directly; population uncertainty then requires an additional sampling model.

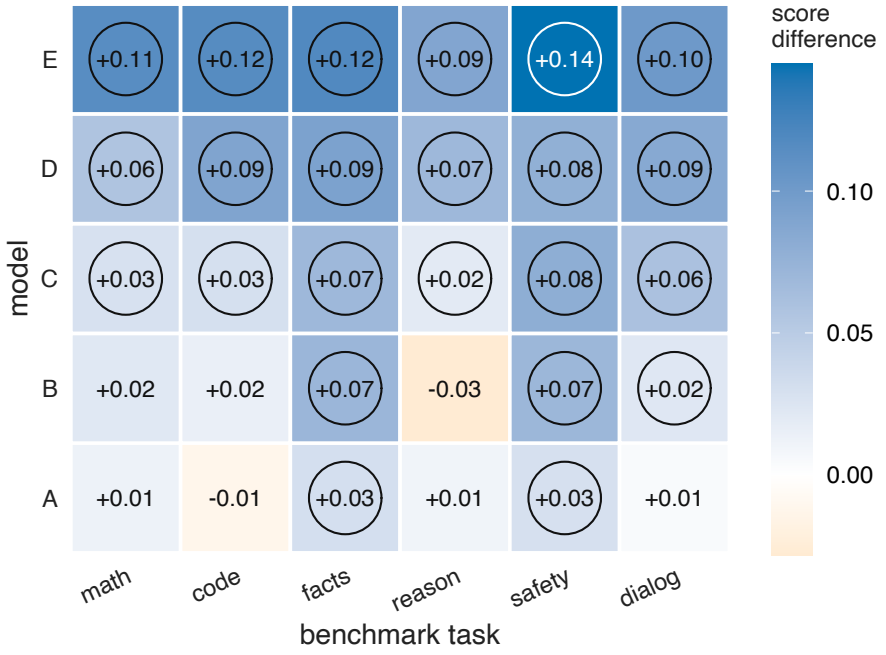



FIGURE 14.3. Benchmark multiplicity for model evaluation. Each cell represents one of 30 predeclared model–task comparisons, using 60 independent item averages with three correlated repetitions per item. The comparison datasets are independent in this simulation. Circles mark the 23 discoveries from BH at $q = 0.10$, applied once to all 30 one-sided t-test p-values. Colors and numbers show estimated mean score differences in one simulated dataset.

There is no universal error rate for leaderboards. If the scientific claim is "model $A$ improves this one pre-specified metric," then an ordinary paired test or confidence interval may be appropriate. If the claim is "model $A$ improves many tasks," the family is the collection of task-level tests. If the claim is "we found the best model among many models and tasks," the

selection process must be included in the uncertainty statement. The same distinction appeared in Chapters 4 and 12: the reported claim determines the family.

Real benchmark comparisons often share questions and baseline responses. To use a bootstrap for performance on new questions, resample the independent questions within each task and carry along every model's responses and all repeated scores for each selected question. Recompute the model comparisons from those resampled records. Here a *cell* means one model–task summary in the heatmap; resampling the 30 cell summaries would discard the paired observations that produced them. If related questions occur in independent groups, such as several questions about one passage, resample those groups. Resampling can estimate uncertainty in each comparison, but multiplicity across comparisons still needs a separate correction with suitable dependence assumptions.

A hierarchical analysis starts by fixing what each level of discovery means. For example, organize the hypotheses as capability domain → task → model–baseline comparison. A task's null says that none of its candidate models improves on the baseline; a domain's null says that none of its tasks contains an improvement. TreeBH in Section 4 of Chapter 7 tests down this prespecified tree, reaching a comparison only through rejected ancestors. It adjusts the testing levels along each selected branch and controls the selected-family FDR described there. The version proved in this book requires independent leaf p-values; reusing questions across comparisons does not automatically satisfy that assumption. When the desired report is simply a flat list of improvements and their dependence is unrestricted, BY supplies the corresponding FDR correction for valid individual p-values.

Benchmark-design work such as Liang et al. [127] helps specify the tasks and metrics that belong in such a report. Governance frameworks such as National Institute of Standards and Technology [139] place that report alongside the intended uses and limitations of the evaluated system.

## 5. LLM-as-a-Judge and Latent Entanglement

**Route: Research frontier**

An LLM can also be used as a judge: it reads another model's answer and assigns a score or a correctness label. This can reduce the cost of evaluation, but the judge's score is itself a measurement that needs checking. A useful judge must respond to the answer it reads. The questions are whether its scores track answer quality, whether it favors particular models, and whether several judges make the same mistakes.

For numerical scores, a simple model separates these sources of variation:

$$J_{irk} = L_{ir} + b_k + u_{ik} + \varepsilon_{irk}, \tag{44}$$

Here $J_{irk}$ is judge $k$'s score for response $r$ to item $i$, $L_{ir}$ is the answer quality we would like to measure, and $b_k$ is the judge's average tendency to score high or low. The term $u_{ik}$ allows that judge to find particular items unusually easy or difficult to assess; $\varepsilon_{irk}$ is the remaining error. Shared training data and similar training methods can make different judges' errors correlated. Averaging more judges need not remove a bias they share.

Kuai et al. [108] study *behavioral entanglement*, meaning dependence in model responses and errors, and report shared failures among the models they evaluate. Agreement can therefore reflect a common mistake rather than several independent checks. Yang et al. [196] compare branches of model reasoning traces instead of relying only on final-answer votes. Their reported accuracy gains motivate further study; they do not establish an error guarantee for a new evaluation.

Before using judges to support scientific claims, compare their decisions with independently checked reference answers. When model identity or answer order could influence a score, conceal the identity and randomize the order in the evaluation design. Keep scores on the same question together when estimating uncertainty. A multiple-testing correction cannot remove a systematic scoring error. The following example shows what multiplicity control can do once a specific audit and target have been defined.

**Example 14.2: Auditing five fixed judging rules**

Suppose five judging rules each label an answer as correct or incorrect. A rule may be a single judge or a fixed combination of judges. Freeze all five rules, including their prompts and combination weights, before drawing 100 answers independently from the same population of interest. Obtain a trusted reference label for each answer and apply every rule to all 100 answers. Any random generation within a rule is repeated independently across audit items. Different rules may make dependent errors on the same item.

Let $X_{ij} = 1$ when rule $j$ disagrees with the reference label on answer $i$, and let $\theta_j$ be its probability of disagreement on a new answer from this population. Then

$$S_j = \sum_{i=1}^{100} X_{ij} \sim \text{Binomial}(100, \theta_j), \qquad j = 1, \ldots, 5.$$

To look for rules with error probability below 10%, test

$$H_j : \theta_j \geq 0.10 \quad \text{against} \quad \theta_j < 0.10, \qquad p_j = \mathbb{P}\{\text{Binomial}(100, 0.10) \leq S_j\}.$$

The lower tail is appropriate because unusually few errors support the alternative. Under $H_j$, the chance of observing so few errors is largest at $\theta_j = 0.10$; increasing $\theta_j$ shifts the error count upward. Thus this binomial-tail p-value is valid throughout the composite null, with possible conservatism because the count is discrete.

For an illustrative audit, suppose the five error counts are

$$(S_1, S_2, S_3, S_4, S_5) = (2, 4, 6, 10, 12).$$

The p-values, rounded for display, are

$$(p_1, p_2, p_3, p_4, p_5) \approx (0.00194,\ 0.02371,\ 0.11716,\ 0.58316,\ 0.80182).$$

Because the rules were evaluated on the same answers, we allow arbitrary dependence among these p-values and apply BY from Chapter 6 at $q = 0.10$. The five BY step-up thresholds are approximately

$$(0.00876,\ 0.01752,\ 0.02628,\ 0.03504,\ 0.04380).$$

Only the first ordered p-value passes its threshold, so only rule 1 is selected. The resulting claim is that this rule's disagreement probability is below 10%. BY controls the FDR among such claims across repeated audits; it does not guarantee that every selected rule satisfies the target. The guarantee concerns agreement with the reference labels on the specified answer population. Incorrect reference labels or a different deployment population can change what this audit tells us about actual answer quality.

This example also connects to selection in Chapter 12. An analyst may choose judges, prompts, or ensemble weights using an initial dataset, then freeze the chosen rule and audit it on independent fresh items. Conditional on the initial data, the chosen rule is fixed for the new audit. If the analyst instead tunes the rule on the audit answers, the binomial calculation above no longer follows from a fixed-rule model. More generally, reporting the best observed judge

among many candidates requires accounting for that search. The target should be the chosen rule's performance on new answers, not merely its performance on the data that made it look best.

## 6. Watermarked Generation

**Route: Research frontier**

A watermark links generated text to a secret random key. The model provider uses the key when choosing tokens, and a detector later checks whether a received document has an unusually strong match to that key. The document may have been edited or mixed with other text before it reaches the detector. The statistical question is whether the observed match could plausibly arise without the key having influenced the text [104, 157].

There are three roles. The *provider* runs the generating model and controls the secret key. The *text submitter* supplies a document to be checked. The *detector* performs the test using authorized access to the key or its required evaluations. The provider may also operate the detector, but the roles need not belong to the same party. Figure 14.4 follows the text from a user's prompt to publication and then to detection.

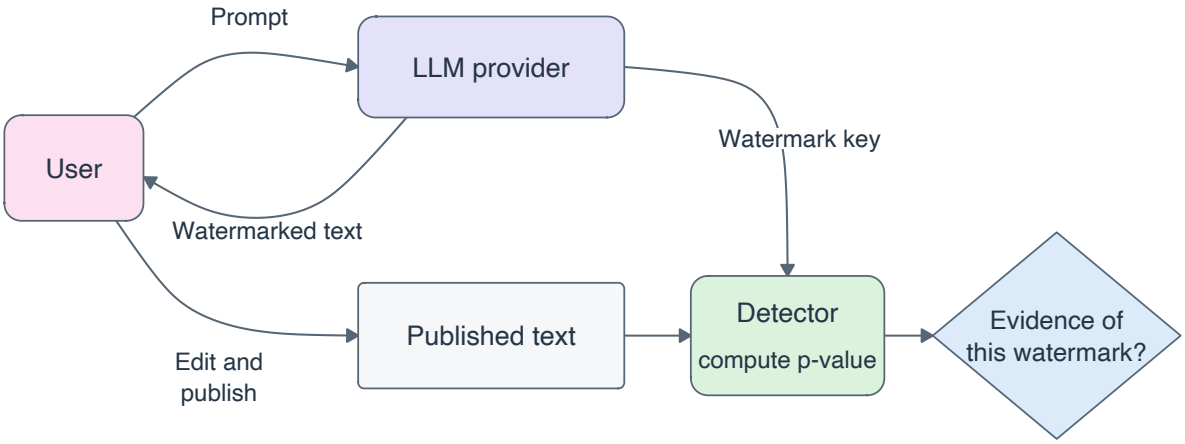



FIGURE 14.4. From a prompt to a watermark test. The provider returns watermarked text, which the user may edit before publishing. The detector receives the published text and authorized access to the watermark key, then computes a p-value by comparison with reference keys. A small p-value is evidence against independence of the text and this key; it is not a general test of whether any LLM generated the text.

Let $\mathcal{A}$ be the vocabulary and $K = |\mathcal{A}|$. A token is an element of this vocabulary, such as a word or a word fragment. Write $Y_{<t}$ for the tokens generated before position $t$, restricted to the model's context window when necessary. Given this history, the model assigns the next-token probabilities

$$p_t(k) = \mathbb{P}(Y_t = k \mid Y_{<t} = y_{<t}), \qquad k \in \mathcal{A}.$$

An ordinary decoder samples from these probabilities. A *keyed decoder* instead uses a random key $\xi_t$ to make that choice:

$$Y_t = \Gamma(\xi_t, p_t).$$

For the model below, draw keys independently from a specified distribution, with the current key independent of the preceding tokens.

In software, a pseudorandom function (PRF) often reconstructs the key material from a master secret and the position. Its outputs are deterministic once that secret is fixed. We first analyze genuinely random keys and return to this implementation distinction after the testing result.

We would like the watermark to preserve the model's next-token probabilities when averaging over the random key. This requirement makes it possible to detect a relation between the text and the key without changing the distribution of the text alone.

**Definition 14.3: Distortion-free keyed generation**

A keyed decoder $\Gamma$ is distortion-free for a key distribution under which $\xi_t$ is independent of $Y_{<t}$ if

$$\mathbb{P}\{\Gamma(\xi_t, p_t) = k \mid Y_{<t} = y_{<t}\} = p_t(k), \qquad k \in \mathcal{A}.$$

If we ignore the key, the next token still has the model's probabilities. The detectable signal lies in how the token choice depends on the key.

For a binary vocabulary $\mathcal{A} = \{0, 1\}$, a simple distortion-free decoder is

$$Y_t = \mathbf{1}\{U_t \le p_t(1)\}, \qquad U_t \sim \mathrm{Unif}[0, 1].$$

Another distortion-free decoder used in practice is exponential minimum sampling. Assume for the moment that $p_1, \dots, p_K > 0$ and $\sum_{k=1}^{K} p_k = 1$. Draw independent

$$U_k \sim \mathrm{Unif}[0, 1], \qquad E_k = \frac{-\log U_k}{p_k}.$$

Then $E_k \sim \mathrm{Exp}(p_k)$, where the parameter is the rate, and the decoder chooses the token with the smallest exponential variable:

$$Y = \arg\min_{1 \le k \le K} E_k.$$

Tokens with $p_k = 0$ can be excluded or assigned $E_k = \infty$.

**Proposition 14.4: Exponential minimum sampling is distortion-free**

With $E_1, \dots, E_K$ as above,

$$\mathbb{P}(Y = k) = p_k, \qquad k = 1, \dots, K.$$

**Proof of Proposition 14.4**

Token $k$ wins when every other exponential variable exceeds $E_k$. Conditioning on the value $t$ of $E_k$ gives

$$\mathbb{P}(Y = k) = \int_0^\infty \mathbb{P}(E_j > t \text{ for all } j \ne k)\, p_k e^{-p_k t}\, dt.$$

By independence, the probability inside the integral is the product of the survival probabilities $e^{-p_j t}$ over $j \ne k$. Since the rates sum to one,

$$\mathbb{P}(Y = k) = \int_0^\infty p_k e^{-t \sum_j p_j}\, dt = p_k.$$

□

## 7. Watermark Detection and Scanning

**Route: Research frontier**

Let $\widetilde{y}_{1:\ell}$ be a received segment of $\ell$ tokens. The null hypothesis used below is that this text is independent of the tested key. Under the alternative, some or all of the text was generated using that key. We need a statistic that measures their agreement. Two examples show how to construct one for exponential minimum sampling.

**Two scores for agreement with the key.** At position $t$, the key consists of the uniforms $U_{t,1}, \dots, U_{t,K}$ used by the decoder. The number $U_{t,\widetilde{y}_t}$ is the entry associated with the token we actually observe. First consider the average score

$$S(\xi, \widetilde{y}_{1:\ell}) = \frac{1}{\ell} \sum_{t=1}^{\ell} \{1 + \log U_{t,\widetilde{y}_t}\}. \tag{45}$$

Under the null, fixing the text does not change the key distribution. Each selected entry remains uniform, and $\mathbb{E}(\log U) = -1$, so each summand has mean zero. The keyed decoder tends to choose a token with a large associated uniform, making this score larger. This is the exponential-minimum score used in Li et al. [125].

A second score also uses the model's probabilities. Suppose the detector can evaluate $p_t$ at each received history, with the prompt and scoring rule fixed. Define $\rho_t = p_t(\widetilde{y}_t)$, and assume $0 < \rho_t \le 1$. A probability-weighted score is

$$S_{\mathrm{wt}}(\xi, \widetilde{y}_{1:\ell}) = \frac{1}{\ell} \sum_{t=1}^{\ell} \frac{1 - \rho_t}{\rho_t} \{1 + \log U_{t,\widetilde{y}_t}\}. \tag{46}$$

Why these weights? Consider one unedited token from the keyed decoder, holding its preceding history fixed. Conditional on $U_{t,k} = u$, token $k$ wins the exponential race precisely when $U_{t,j} < u^{p_t(j)/p_t(k)}$ for every $j \neq k$. Independence gives the probability $u^{(1-p_t(k))/p_t(k)}$ for these inequalities. Dividing by $\mathbb{P}(Y_t = k) = p_t(k)$, the conditional density of the winning uniform is therefore

$$f(u \mid Y_t = k, Y_{<t}) = \frac{1}{p_t(k)} u^{1/p_t(k) - 1}, \qquad 0 < u < 1.$$

Under the null its density is one. Using this selected entry of the key, the token contributes the log likelihood ratio

$$-\log \rho_t + \frac{1 - \rho_t}{\rho_t} \log U_{t,\widetilde{y}_t}.$$

Its null mean is $-\log \rho_t - (1 - \rho_t)/\rho_t$. Subtracting that mean gives the weighted summand in (46). For a fixed text, this subtraction is the same for every key, so ranking keys by their summed log likelihood ratios gives the same ordering as (46). An unlikely token receives a larger weight; a token with probability one receives weight zero because its choice cannot reveal the key. This calculation motivates the likelihood approach of Li et al. [126]. It concerns the ideal, unedited decoder; it does not assert optimal detection after arbitrary edits.

Both scores can be calibrated by the same procedure below. For the weighted score, probabilities must be obtained by a rule that does not use the tested key. Hold those probabilities fixed for all reference keys. If instead a fitting or tuning step uses the key, that step must be repeated for every reference key as part of the statistic. This requirement prevents the tested key from receiving a favorable choice that the reference keys do not receive.

**Compare the observed key with reference keys.** Write $T_{\mathrm{wm}}(\xi, \widetilde{y}_{1:\ell})$ for the chosen statistic, such as $S$ or $S_{\mathrm{wt}}$, with larger values indicating stronger agreement. Before scoring a displayed document, choose one tokenizer and one preprocessing rule and use them in every calculation. This fixes what the tokens $\widetilde{y}_t$ mean; we discuss alternative encodings below.

Draw independent reference keys $\xi^{(1)}, \ldots, \xi^{(B)}$ from the same distribution as the tested key, and score each against the same text. Count how often a reference score is at least as large as the observed score:

$$p_{\mathrm{wm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_{\mathrm{wm}}(\xi^{(b)}, \widetilde{y}_{1:\ell}) \geq T_{\mathrm{wm}}(\xi, \widetilde{y}_{1:\ell})\}}{B+1}. \tag{47}$$

A small value means that few reference keys match the text as well as the tested key. The added 1 counts the tested key itself, and also prevents a p-value of zero when none of the $B$ reference scores exceeds it. If a different statistic treats smaller values as stronger evidence, reverse the comparison in the numerator.

**Proposition 14.5: Key-resampling p-value**

Suppose $\xi, \xi^{(1)}, \ldots, \xi^{(B)}$ are exchangeable—normally iid from the declared key distribution—and, under the null, $\widetilde{y}_{1:\ell}$ is independent of all $B+1$ keys. Then the p-value in (47) is super-uniform:

$$\mathbb{P}(p_{\mathrm{wm}} \leq \alpha) \leq \alpha.$$

**Proof of Proposition 14.5**

Write the observed-key score as $S_0$ and the resampled-key scores as $S_1, \ldots, S_B$. Conditional on the text, these $B+1$ scores are exchangeable under the null. Rank them from largest to smallest, breaking any ties uniformly at random, independently of the scores. By symmetry, the observed score's rank $R_0$ is uniform on $\{1, \ldots, B+1\}$.

The reported p-value counts *every* score tied with $S_0$ as at least as extreme. Its numerator is therefore at least $R_0$, so $p_{\mathrm{wm}} \geq R_0/(B+1)$. Consequently,

$$\mathbb{P}(p_{\mathrm{wm}} \leq \alpha \mid \widetilde{y}_{1:\ell}) \leq \mathbb{P}\{R_0 \leq \alpha(B+1) \mid \widetilde{y}_{1:\ell}\} = \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \leq \alpha.$$

Averaging over the text gives the claimed unconditional bound. □

**Remark 14.6: Random keys and their software implementation**

The proposition treats the tested key and the reference keys as random draws from the same distribution. A PRF instead produces a reproducible sequence from a secret master key. Security assumptions are needed to justify treating that sequence as indistinguishable from random numbers for the computations an adversary can perform. The statistical proof establishes the result for the random-key experiment; it does not establish those security assumptions for a particular implementation.

**Search for a watermarked passage.** A long document may contain only a short watermarked passage. An editor may also move a passage, changing which key positions correspond to which text positions. Fix a window length $L$ and the allowed positions before testing. Let $\mathcal{M}$ be a score for one aligned pair of windows, such as the average in (45). A detector that searches

key starts $a$ and text starts $c$ reports

$$T_{\mathrm{wm,max}} = \max_{a,c} \mathcal{M}(\xi_{a:a+L-1}, \widetilde{y}_{c:c+L-1}).$$

Even under the null, this maximum has many opportunities to be large. The reference calculation must give every reference key the same opportunities:

$$T^{(b)}_{\mathrm{wm,max}} = \max_{a,c} \mathcal{M}(\xi^{(b)}_{a:a+L-1}, \widetilde{y}_{c:c+L-1}), \qquad b = 1, \ldots, B.$$

Substitute these maxima into (47). The rank argument still applies, because the entire search is now the statistic computed for every key. If several window lengths are searched, repeat that additional choice as well. Another valid option is to calibrate each prespecified window separately and use a Bonferroni correction for the number of tests.

Detecting a watermark somewhere and locating its boundaries are different questions. A significant maximum supports the first claim. It does not by itself provide uncertainty statements for the selected endpoints. Li et al. [125] study the second problem by combining window-based randomization tests with change-point methods to separate watermarked and non-watermarked portions of a document.

**Example 14.7: Worked watermark audit: resample the full scan**

Consider a vocabulary of four tokens with probabilities $(0.4, 0.3, 0.2, 0.1)$ at every position. Generate a text of length 120. At positions 41–60 use exponential minimum sampling with the tested key; at all other positions sample independently from the same four-token distribution. Thus the token probabilities are unchanged, but only the middle passage uses the tested key.

For this teaching example, key and text positions are aligned, so a window starting at position $c$ uses key positions $c, \ldots, c+19$. Compute the score $S$ for each of the 101 contiguous windows of length 20, and take their maximum. Draw $B = 999$ independent reference keys, keep the text fixed, and repeat all 101 windows for each key. The resulting p-value is

$$p_{\max} = \frac{1 + \sum_{b=1}^{999} \mathbf{1}\{T^{(b)}_{\mathrm{wm,max}} \geq T_{\mathrm{wm,max}}\}}{1000}.$$

In one reproducible realization, the observed maximum exceeds all 999 reference maxima, giving $p_{\max} = 0.001$. Comparing that maximum with reference scores for only the first window also gives 0.001. The signal is strong enough for both comparisons to reach the smallest possible rank p-value. This example alone therefore cannot show whether either comparison controls false alarms.

To examine calibration, instead hold a 120-token text fixed independently of all keys. In each of 1,000 repetitions, draw a fresh tested key and 999 reference keys. Figure 14.5 compares three rules, each rejecting when its reported p-value is at most 0.05:

- *Full scan versus full scan:* compare the observed maximum over all 101 windows with the corresponding maximum for each reference key.
- *Fixed window versus the same window:* use only tokens 1–20 for both the tested and reference keys, choosing this window before seeing any scores. This valid test concerns that window alone.
- *Full scan versus a fixed window:* compare the observed maximum with reference scores for tokens 1–20. This is invalid because only the tested key receives 101 chances to produce a large score.

The two matched rules each reject in 50 of the 1,000 null repetitions (5.0%). The mismatched rule rejects in 612 repetitions (61.2%). Thus the issue is not whether both rules can detect a strong watermark; it is whether they avoid reporting a watermark when the text is independent of the key. Searching is allowed, provided the reference calculation repeats the same search. The simulation illustrates this point; the rank argument supplies the validity guarantee.

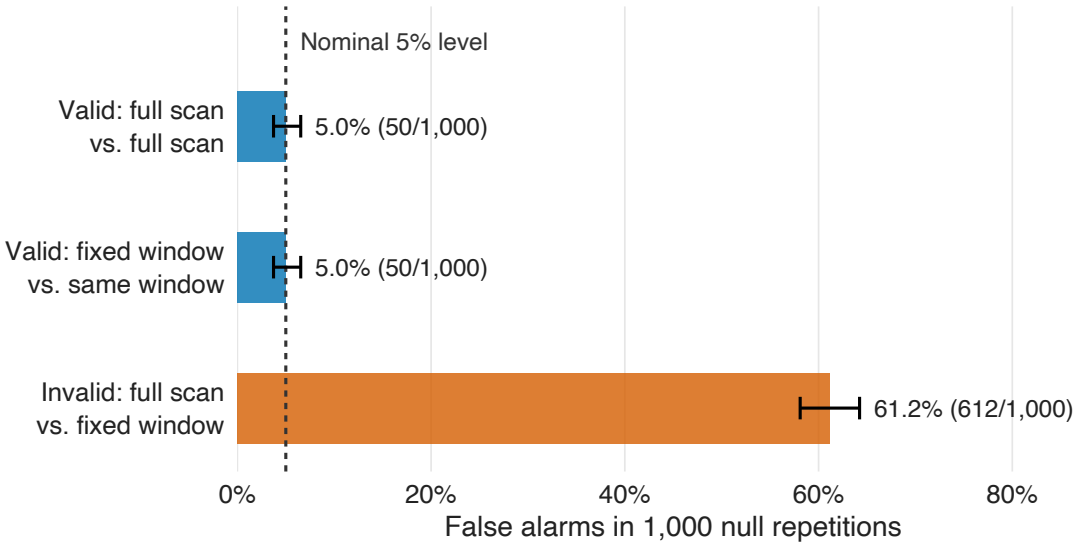



FIGURE 14.5. False alarms when the text is independent of the tested key. Each of 1,000 null repetitions uses a 120-token text, 999 reference keys, and windows of length 20. The first two rules apply the same statistic to every key and attain the nominal 5% level in this experiment. Comparing an observed scan maximum with single-window reference scores gives 61.2% false alarms. The dashed line marks 5%; the whiskers are 95% binomial confidence intervals showing simulation uncertainty in the rejection probabilities.

**Fix the text representation before testing.** A tokenizer splits a displayed string into tokens. For example, a vocabulary could permit both (`water`, `mark`) and (`watermark`) to display the same word. A fixed deterministic tokenizer chooses one encoding. We call that specified choice the *canonical encoding*; using it consistently is *canonicalization*. It fixes the representation of a string, not the meaning of every possible paraphrase.

The issue called *tokenization multiplicity* arises when a protocol accepts several encodings of the same displayed text. Chatzi et al. [40] document how different token counts can lead to different charges under token-based pricing. Their proposal restricts generation to a canonical encoding. For watermark auditing, the related lesson is to fix the encoding before calculating the score.

To see why this matters, distinguish two possible searches. A text submitter trying to conceal a watermark may seek an accepted encoding with a small detection score; this can reduce detection power. An analyst searching for evidence may instead try several encodings and report the largest score; this gives extra opportunities for a false alarm. These are possible uses of a

flexible protocol, not reasons to assume that the model provider wants to hide its own watermark. If the analyst searches several encodings, the same search must be repeated for every reference key, just as for windows.

Privacy is a separate implementation concern. A text submitter may want a document checked without disclosing its content to the key holder. Verifiable oblivious watermark detection (VOW) studies protocols for such private verification [131]. This cryptographic protection does not determine how a detection score should be calibrated; that statistical task still depends on the score and search procedure used in the audit.

## 8. Contamination Auditing by Exchangeability

**Route: Research frontier**

Benchmark contamination asks whether a model had access to evaluation data during training or tuning. String matching can be useful evidence, but it is not a statistical proof by itself. A strong audit states a null hypothesis under which a test statistic has a known or resampling-calibrated reference distribution.

The exchangeability test of Oren et al. [143] asks whether the model gives unusually high probability to the order in which benchmark examples were published. Let

$$X = (X_1, \dots, X_n)$$

be the examples in that order, called the *canonical order*, and let $\mathcal{P}$ be the language model being audited. Join the examples into one text and compute its log probability:

$$T(X) = \log \mathcal{P}(X) = \log \mathcal{P}(X_1) + \sum_{i=2}^{n} \log \mathcal{P}(X_i \mid X_1, \dots, X_{i-1}).$$

When the model scores example $X_i$, the earlier examples are part of its context. For example, in the order $(X_1, X_2, X_3)$, it scores $X_3$ after reading $X_1, X_2$; in the order $(X_1, X_3, X_2)$, it scores $X_3$ after reading only $X_1$. Thus changing the order can change the score. Resetting the context before every example and adding the separate log probabilities would give the same total in every order, leaving this test with no power.

For the reference scores, uniformly shuffle the examples and recompute the log probability of the entire joined text. Use the same separators, formatting, and context-window rule for every ordering. The null assumes that $\mathcal{P}$ is independent of the benchmark dataset and that the benchmark examples are exchangeable under the audit's permutations. Under these assumptions, the canonical order has the same score distribution as a random ordering.

**Proposition 14.8: Permutation p-value for canonical-order contamination**

Let $\pi_1, \dots, \pi_B$ be independent uniform random permutations of $\{1, \dots, n\}$. If $X$ is exchangeable and independent of the audited model under the null, then

$$p_{\text{contam}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T(X_{\pi_b}) \geq T(X)\}}{B + 1}$$

is a valid Monte Carlo permutation p-value when large $T$ is evidence of contamination.

**Proof of Proposition 14.8**

Under the null, the canonical ordering and the randomly permuted orderings are exchangeable. Conditional on the unordered collection of examples, the scores

$$T(X), T(X_{\pi_1}), \ldots, T(X_{\pi_B})$$

therefore have an exchangeable joint distribution. The rank argument used in Proposition 14.5 applies verbatim. □

Figure 14.6 illustrates the comparison schematically. The line represents a canonical-order score, and the shaded curve represents reference scores. In an actual audit, these reference scores must come from permuting the same benchmark examples and recomputing the same statistic. The proposition then counts how many reference scores are at least as large as the canonical-order score, including the plus-one correction.

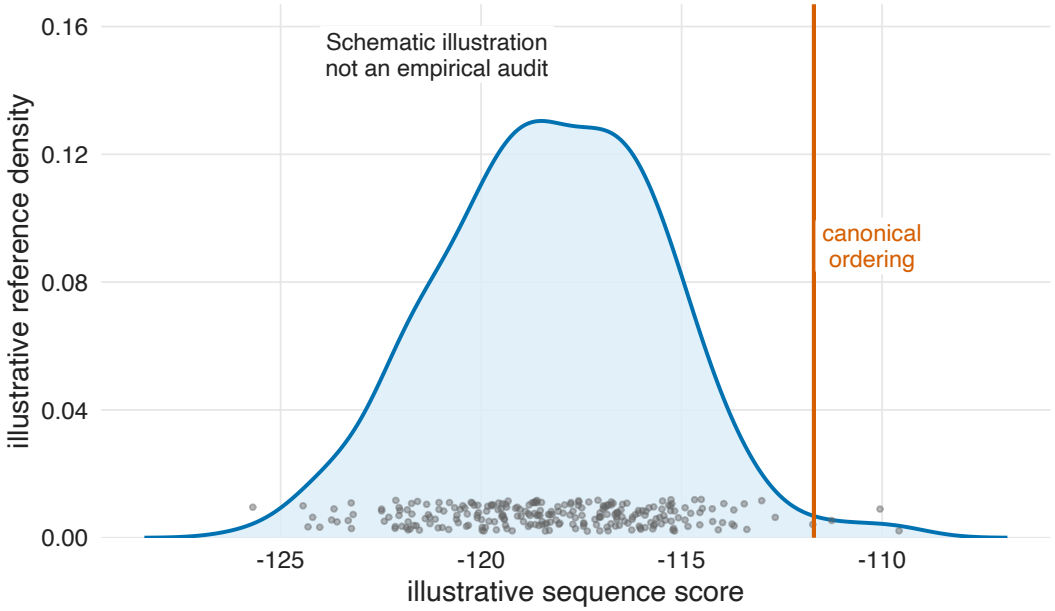



FIGURE 14.6. Schematic comparison for a contamination audit. The line marks an illustrative canonical-order score and the shaded curve shows illustrative reference scores. The plotted values are simulated to explain the comparison; they are not results from an audited language model.

The canonical ordering matters. If the benchmark has a natural order that could have appeared in training, a contaminated model may assign high probability to the transitions between consecutive examples. Shuffling the examples breaks those transitions. Conversely, if examples are not exchangeable because the benchmark is deliberately ordered by difficulty or topic, the null reference distribution needs to preserve that structure.

Watermark-based contamination tests are another route [157]. If a benchmark or synthetic dataset is released with a known watermark, then a later model can be audited for the statistical signature of that watermark. The logic is again a hypothesis test. The null must specify how the watermark statistic behaves when the model was not trained on the watermarked data, and the p-value must be calibrated with the same statistic that will be reported.

When many datasets, models, or watermark keys are audited, the tests form a family. Declare that family before selecting its most suspicious results. For a fixed family, use FWER control to bound the probability of any false finding, or FDR control to bound the expected fraction of false findings in the reported list. If the audits supply valid e-values, e-BH from Chapter 8 provides an FDR procedure.

An audit that continues over time raises two different questions. First, suppose we keep gathering evidence about one fixed model–dataset pair and want to stop as soon as the evidence is strong. This calls for an e-process: a sequence whose value remains a valid e-value at any stopping time. As in Section 5 of Chapter 8, crossing $1/\alpha$ then gives a level-$\alpha$ test for that one null. To build such a process by multiplying new evidence, each factor must have null expectation at most one conditional on all information available before that factor is computed. Multiplying separately valid e-values, or repeatedly reanalyzing the same benchmark, does not by itself meet this requirement.

Second, each newly released model can create a new model–dataset hypothesis. An online multiple-testing rule such as e-LOND allocates testing levels as these hypotheses arrive [194]. Its FDR target concerns the list of discoveries accumulated across audits. The version proved in Theorem A.17 requires each true-null e-value to have conditional expectation at most one given the past audits; it then permits stopping the overall audit at an almost-surely finite stopping time, meaning a rule that eventually stops with probability one and uses only information already observed. Shared benchmarks and model choices based on earlier results must be accounted for when establishing that condition. Applying a separate fixed-family procedure to each model release would instead give a guarantee for each reported family, not automatically for their accumulated union.

## 9. Reference-Relative Factuality

**Route: Research frontier**

A generated answer may contain some well-supported claims and some details that the available reference does not support. Reference-relative factuality addresses this problem by removing detail or making the answer less specific. Calibration examples determine how far to simplify future answers. The aim is a probability guarantee that the remaining answer is supported by its reference. The construction below follows Mohri and Hashimoto [138] and uses the split conformal method from Chapter 10.

Let $x \in \mathcal{X}$ be a user input and let a language model produce a base answer $A_0(x)$. Let $y^\star \in \mathcal{Y}$ denote a reference answer or reference knowledge object. We write

$$y^\star \Rightarrow a$$

to mean that the reference supports all claims in the answer $a$, or *entails* it. For an answer $a$, define its entailment set

$$E(a) = \{y \in \mathcal{Y} : y \Rightarrow a\}.$$

The answer $a$ is factual for reference $y^\star$ exactly when

$$y^\star \in E(a).$$

The challenge is that the base answer may be too specific. Fix a finite threshold grid $\mathcal{T} = \{t_1 < \cdots < t_L\}$. A back-off family $\{F_t : t \in \mathcal{T}\}$ turns the base answer into less specific answers as $t$ increases. For example, $F_t(x)$ may remove claims whose verifier scores are below a threshold, replace a precise numerical claim by a range, or eventually abstain from making any factual claim. The induced entailment sets should be nested:

$$E(F_t(x)) \subseteq E(F_{t'}(x)) \qquad \text{for } t \le t'.$$

Larger $t$ means a safer but less informative answer. Require the terminal output $F_{t_L}(x)$ to abstain from factual claims, so that $E(F_{t_L}(x)) = \mathcal{Y}$ for every $x$.

Freeze the generator, threshold grid, back-off rules, and entailment verifier before calibration. They may be learned from independent training data; in that case the argument conditions on the training data. The same fixed pipeline must score each calibration pair and process the future input. For a randomized generator, include its independently generated random seed in each input so that the resulting augmented pairs remain exchangeable.

For a calibration pair $(X_i, Y_i^\star)$, define the strict-safe score

$$A_{\text{safe}}(X_i, Y_i^\star) = \min\{t \in \mathcal{T} : Y_i^\star \in E(F_s(X_i)) \text{ for every } s \in \mathcal{T},\ s \geq t\}. \tag{48}$$

The score is the first grid threshold after which the reference entails every remaining answer. The terminal abstention makes the set nonempty, and finiteness ensures that the minimum is attained. Under nested entailment sets this is simply the first safe threshold. The stronger "every remaining answer" definition also protects the argument if a concrete back-off path has an occasional nonmonotone step: a single safe answer is then insufficient.

Figure 14.7 links the nested back-off path to the calibration step. Each calibration pair contributes the first threshold after which every more conservative answer remains safe; ordering those strict-safe scores produces $\widehat{t}_\alpha$. Applying $F_{\widehat{t}_\alpha}$ to a new prompt then makes the answer less specific in order to meet the coverage guarantee stated below.

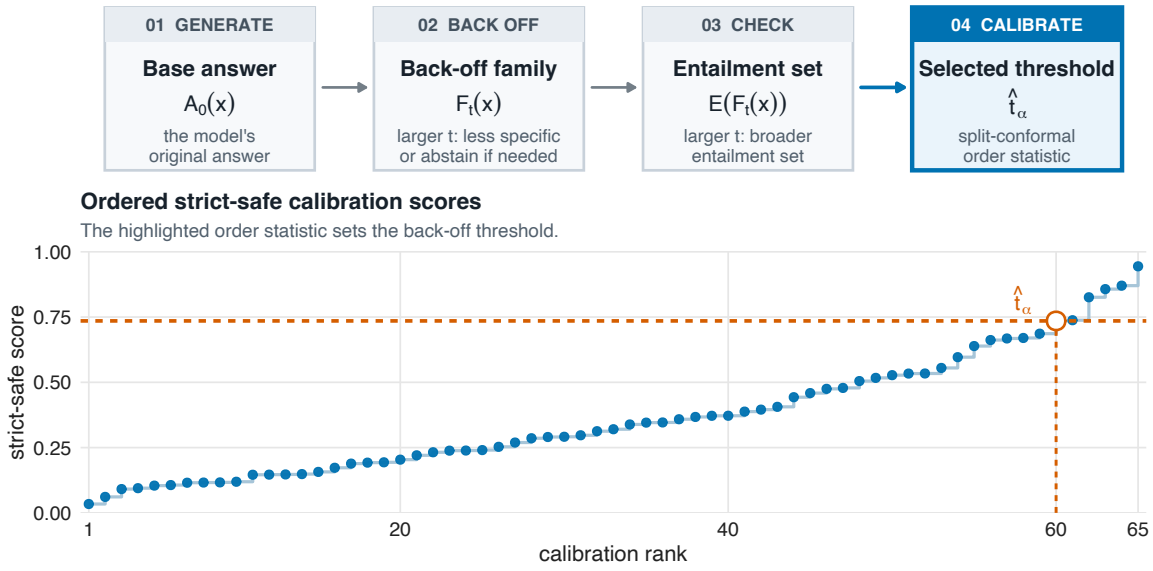



FIGURE 14.7. Reference-relative factuality. Increasing $t$ trades specificity for a larger entailment set; the $k$th ordered strict-safe score supplies the calibrated threshold $\widehat{t}_\alpha$ when $k \leq n$.

Let $(X_1, Y_1^\star), \ldots, (X_n, Y_n^\star)$ be calibration examples and let $A_{\text{safe},i} = A_{\text{safe}}(X_i, Y_i^\star)$. For $\alpha \in (0, 1)$, set

$$k = \lceil (n+1)(1-\alpha) \rceil$$

and let $\widehat{t}_\alpha$ be the $k$th order statistic of $A_{\text{safe},1}, \ldots, A_{\text{safe},n}$ if $k \leq n$. If $k = n+1$, use $\widehat{t}_\alpha = t_L$, the terminal abstention. This last case occurs when the requested error level is smaller than

$1/(n+1)$. The conformalized output is

$$\widetilde{\mathcal{P}}(x) = F_{\widehat{t}_\alpha}(x).$$

**Theorem 14.9: Split conformal reference-relative factuality**

Assume

$$(X_1, Y_1^\star), \ldots, (X_n, Y_n^\star), (X_{n+1}, Y_{n+1}^\star)$$

are exchangeable, conditionally on any independent training data used to construct the pipeline. Assume the pipeline is then fixed, the threshold grid is finite, and $E(F_{t_L}(x)) = \mathcal{Y}$ for every $x$. Use the strict-safe score (48) and the threshold rule above. Then

$$\mathbb{P}\left\{Y_{n+1}^\star \in E(\widetilde{\mathcal{P}}(X_{n+1}))\right\} \geq 1 - \alpha.$$

**Proof of Theorem 14.9**

The scores

$$A_{\text{safe},1}, \ldots, A_{\text{safe},n}, A_{\text{safe},n+1} = A_{\text{safe}}(X_{n+1}, Y_{n+1}^\star)$$

are exchangeable because the same fixed function is applied to each pair. If $k = n+1$, the output abstains and the entailment probability is one. Otherwise, by the split conformal rank argument from Chapter 10,

$$\mathbb{P}(A_{\text{safe},n+1} \leq \widehat{t}_\alpha) \geq 1 - \alpha.$$

On the event $A_{\text{safe},n+1} \leq \widehat{t}_\alpha$, the attained minimum and the "every remaining threshold" condition imply that $\widehat{t}_\alpha$ itself is safe, so

$$Y_{n+1}^\star \in E(F_{\widehat{t}_\alpha}(X_{n+1})).$$

This is exactly the desired reference-relative entailment event. □

**Remark 14.10: What the guarantee means**

Imagine repeating the calibration study and then drawing a new prompt with its reference, under the exchangeability assumption in the theorem. The probability that the resulting answer is supported by that reference is at least $1 - \alpha$. This averaging over the calibration sample and the new prompt is called *marginal coverage*. The theorem does not give the same probability for every individual prompt or for every realized calibration sample.

The meaning of "supported" also matters. The theorem uses the relation $y^\star \Rightarrow a$ used to construct the calibration scores. If that relation is determined by human labels or an automated verifier, the guarantee concerns those assessments. A wrong reference or a verifier that accepts an unsupported claim can therefore make an incorrect answer count as covered. Reliable references and entailment assessments are needed to interpret coverage as factual correctness.

## 10. What the Assumptions Are Doing

**Route: Core**

Each application needs assumptions that match the claim being reported. It helps to check them one application at a time.

Genomics. Measurement artifacts and population structure can distort the p-values before any multiplicity correction is applied. The hypotheses must also match the scientific question. A marginal association is different from conditional importance, and selecting a variant need not identify the causal variant. An FDR guarantee concerns the selected list; it does not assign a posterior null probability to each selected gene.

Benchmarks and judges. Uncertainty calculations must follow the source of randomness: new questions, new generated responses, new judges, or new training runs. Repeated scores on the same question are not independent new questions. A judge must also measure the quality we intend to compare. Multiplicity correction cannot remove a systematic preference for one model family.

Watermarking. The null calibration must use the statistic and search procedure that produced the reported result. If the analyst searches over substrings or keys, every reference draw must repeat that search. The random-key assumption supports the statistical calculation; the security of a deployed key generator is a separate implementation requirement.

Contamination. The permutation test needs exchangeable examples under the stated null and a score that can detect changes in their order. A benchmark deliberately ordered by topic or difficulty needs a reference scheme that preserves the relevant structure. A nonsignificant result means this audit did not detect contamination; contamination that leaves little trace in the ordering may remain undetected.

Reference-relative factuality. The calibration examples and future prompt–reference pairs must satisfy the exchangeability assumption. The reference answers and entailment assessments determine what counts as coverage. A change in the population of prompts, or inaccurate assessments of support, can undermine the intended factuality interpretation.

**Assumptions in practice**

| | |
|---|---|
| **What you can check** | For the application at hand, record the claim, the source of randomness, and the assumptions behind the reported error guarantee. Check that the actual analysis uses the stated sampling, scoring, selection, and calibration rules. |
| **What these data cannot show** | The error guarantee relies on these assumptions. A small p-value or a calculated coverage level does not establish that the assumptions hold. |

## 11. Bibliographic Notes

**Route: Core**

For genomic multiple testing, see Dudoit and van der Laan [54] for a broad treatment of multiple testing in genomics and Efron [58] for empirical Bayes and local FDR. Knockoffs were introduced for fixed designs by Barber and Candès [9] and extended to the model-X setting by Candès et al. [38]; hidden Markov model knockoffs for GWAS are developed by Sesia et al. [159]. Standard data resources for human genetic association studies include the GWAS Catalog of the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute

(EBI) [166], and the Genotype-Tissue Expression (GTEx) tissue-expression atlas [75]. These provide vetted summary statistics and annotation links underlying many of the multiple-testing analyses in this chapter.

For statistical significance testing in natural-language-processing benchmark evaluation, see the survey of Dror et al. [53]. The Holistic Evaluation of Language Models framework of Liang et al. [127] provides the operational template for benchmarking AI systems across many metrics; the National Institute of Standards and Technology (NIST) AI Risk Management Framework [139] gives the policy framing for how these evaluations should be reported. Green-token watermarking for LLM outputs is due to Kirchenbauer et al. [104]; distortion-free watermarking is developed by Kuditipudi et al. [109]. Li et al. [125] study detection and segmentation of partially watermarked text, and Li et al. [126] use next-token probabilities to improve detection. Benchmark-level watermarking is studied by Sander et al. [157]. Tokenization multiplicity and its pricing consequences are studied by Chatzi et al. [40]; the representation-selection issue above applies that lesson to watermark auditing. The VOW protocol and its cryptographic privacy assumptions are from Luan et al. [131]. The exchangeability-based contamination test is developed by Oren et al. [143]. The reference-relative factuality framework is due to Mohri and Hashimoto [138].

LLM-as-a-Judge latent entanglement is analyzed by Kuai et al. [108], who propose a cumulative-information-gain metric to quantify behavioral correlation between judges. Multi-agent reasoning-tree auditing as a corrective procedure is from Yang et al. [196]. Replicability across multiple studies, which is increasingly relevant for AI-evaluation reproducibility, is treated in Bogomolov and Heller [29] and Heller and Bogomolov [81].

## 12. Exercises

**Route: Core**

### Basic.

**Exercise 14.11: Benchmark families**

A team compares four language models on ten tasks. For each model-task pair, it tests whether the model improves over a fixed baseline. To report all improved cells, the family would contain all 40 comparisons, as in Example 14.1. Suppose instead that the team pre-specified one task as primary and treats all others as exploratory. Identify the confirmatory family and the remaining exploratory comparisons.

### Advanced.

**Exercise 14.12: Multiple contamination audits**

An organization audits one model against 1,000 benchmark datasets and obtains one Monte Carlo permutation p-value per dataset. Propose an analysis plan for identifying suspicious datasets while controlling FDR. State your target level and justify the number of permutations per dataset for your chosen multiple-testing rule. What dependence concerns arise?

Now suppose a new model version arrives each month and the organization wants one accumulated list of suspicious model–dataset pairs. Use the discussion of ongoing audits above to extend your plan: specify the hypotheses, their testing order, a concrete multiple-testing rule, and how its testing levels are assigned. Explain how you would obtain evidence satisfying that rule's assumptions when the same datasets are reused. If the available permutation p-values are insufficient, specify what additional data, randomization, or change in the audit design you would need. State the error guarantee for the accumulated list and justify its applicability to your proposed design.

**Exercise 14.13: Design a reference-relative factuality study**

Design a calibration study for reference-relative factuality in a medical question answering system. Specify the population of prompts, the reference answers, the entailment rule, the back-off family, the target value of $\alpha$, and the number of calibration examples. State clearly what the resulting guarantee does and does not say.

**Exercise 14.14: Knockoff interpretation**

In a genotype study with strong LD, a model-X knockoff procedure selects one variant from a correlated block. Chapter 9 defines the conditional null that this procedure targets relative to the measured variants. Explain why the selection need not identify the unique causal variant, for example when the causal variant is not genotyped. Then propose additional analyses or study designs that would strengthen the biological interpretation of this selected variant.

APPENDIX A

# Advanced Safe and Sequential Inference

**Route: Advanced**

Chapter 8 introduces e-values and e-BH. This appendix develops the supporting theory. The first section asks which e-value is best against a specified alternative. The remaining sections study evidence collected over time, FDR control, approximate e-values, and confidence sequences. We use conditional expectation and basic convergence theorems from graduate probability; specialized definitions are recalled where they enter the arguments.

## 1. Numeraires and reverse information projection

**Route: Advanced**

We retain the notation of Section 3 of Chapter 8: $\mathcal{H}_0$ is a set of null distributions and $Q$ is a specified alternative. We compare e-values by their expected log ratio under $Q$. An e-value $E^*$ is log-optimal if $\mathbb{E}_Q[\log(E/E^*)] \le 0$ for every competing e-value $E$. Using the log ratio avoids subtracting two possibly infinite expected logarithms. This is an average comparison under $Q$; a competing e-value may still be larger on a particular dataset. We retain the ratio conventions used there: $x/\infty = 0$ and $\infty/x = \infty$ for finite $x \ge 0$, and $\infty/\infty = 1$. These conventions include alternatives that assign positive probability to events impossible under every null law. A numeraire satisfies a stronger ratio bound.

**Definition A.1: Numeraire**

An e-value $E^*$ for $\mathcal{H}_0$ with $E^* > 0$ $Q$-almost surely is a *numeraire* for the pair $(\mathcal{H}_0, Q)$ if

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] \le 1 \qquad \text{for every e-value } E \text{ for } \mathcal{H}_0.$$

The numeraire condition implies log-optimality: by Jensen's inequality applied to the concave logarithm,

$$\mathbb{E}_Q\left[\log \frac{E}{E^*}\right] \le \log \mathbb{E}_Q\left[\frac{E}{E^*}\right] \le \log 1 = 0.$$

We use three general facts whose proofs are beyond the scope of this appendix [110, 147]. First, for every null model and every alternative $Q$, a numeraire exists and is $Q$-almost surely unique; no regularity conditions on $\mathcal{H}_0$ are needed. Second, the numeraire is log-optimal, and it is the only log-optimal e-value. Third, the numeraire is a likelihood ratio in a generalized sense: the measure $P^*$ defined by $dP^*/dQ = 1/E^*$ has total mass at most one and is called the *reverse information projection* of $Q$ onto the null. A measure with total mass at most one is called a sub-probability measure. When the null is simple and $P_0$ shares the support of $Q$, $P^* = P_0$ and the numeraire is the ordinary likelihood ratio $dQ/dP_0$. When the projection is attained at a genuine null mixture $Q_{G^*}$ with finite KL divergence, it solves the minimization

$$D(Q \,\|\, Q_{G^*}) = \min_G D(Q \,\|\, Q_G),$$

where $G$ ranges over priors on $\Theta_0$ and $D(Q \,\|\, P) = \mathbb{E}_Q[\log\{dQ/dP\}]$ is the KL divergence: among null mixtures, choose the one closest to the alternative. In general, however, the projection can be a sub-probability measure lying outside the set of null mixtures, which is why the modern theory defines it through the numeraire rather than through the minimization.

The attained case has a short proof that also explains why the minimization must range over null mixtures. Let $\mathcal{C}_0$ be a convex family of null distributions. For a parametric null, one may take $\mathcal{C}_0 = \{Q_G : G \text{ is a prior on } \Theta_0\}$, the family of all null mixtures. Here $Q_G$ is the distribution obtained by first drawing $\theta \sim G$ and then drawing the data from $P_\theta$. This family is convex because mixing two priors produces another prior. An e-value is valid for the original null family if and only if it is valid for all these mixtures. Point-mass priors recover the original null distributions, while averaging the original expectation bounds over any prior $G$ proves the other direction.

**Proposition A.2: Attained reverse information projection**

Suppose $Q$ and every $P \in \mathcal{C}_0$ have densities $g$ and $p$ with respect to a common measure $\nu$. If $P^* \in \mathcal{C}_0$, with density $p^*$, attains

$$D(Q \,\|\, P^*) = \inf_{P \in \mathcal{C}_0} D(Q \,\|\, P) < \infty,$$

then $P^*$ is a null mixture closest to $Q$ in KL divergence. The finite-divergence assumption implies $Q \ll P^*$: every set assigned probability zero by $P^*$ also has probability zero under $Q$. Thus the following ratio is defined $Q$-almost surely; set it equal to zero where $p^* = 0$:

$$E^*(X) = \frac{g(X)}{p^*(X)}$$

is an e-value for $\mathcal{C}_0$. Moreover, $E^*$ is the numeraire for $(\mathcal{C}_0, Q)$, and hence is log-optimal against $Q$.

**Proof of Proposition A.2**

The proof has two steps. First, we turn the KL optimality of $P^*$ into the expectation bound required of an e-value. Second, we use the fact that $P^*$ is itself a null mixture to verify the numeraire condition.

*Step 1: KL optimality implies e-validity.* Fix any $P \in \mathcal{C}_0$, and let $p$ be its density. On $\{p^* > 0\}$, define

$$r = \frac{p}{p^*}.$$

This density ratio measures how $P$ differs from the minimizing distribution $P^*$. For $t \in (0, 1/2]$, let

$$P_t = (1-t)P^* + tP.$$

The distribution $P_t$ moves a fraction $t$ of the way from $P^*$ toward $P$. Convexity ensures that $P_t \in \mathcal{C}_0$. Define $\phi(t) = D(Q \,\|\, P_t)$. Since $P^*$ minimizes the divergence, $\phi(t) \geq \phi(0)$, so every right difference quotient $\{\phi(t) - \phi(0)\}/t$ is nonnegative. The calculation below shows that these quotients have a limit; consequently, the right derivative exists and satisfies $\phi'(0+) \geq 0$. Moreover,

$$\phi(t) - \phi(0) = -\mathbb{E}_Q[\log\{1 + t(r-1)\}]\,.$$

To compute the derivative without imposing an extra integrability condition, set

$$h_t(r) = \frac{\log\{1 + t(r-1)\}}{t}.$$

For every $r \geq 0$, elementary calculus gives $h_t(r) \uparrow r - 1$ as $t \downarrow 0$, and

$$h_t(r) \geq \frac{\log(1-t)}{t} \geq -2, \qquad 0 < t \leq \frac{1}{2}.$$

Along any sequence $t_k \downarrow 0$, the nonnegative variables $h_{t_k}(r) + 2$ increase pointwise to $r + 1$. The monotone convergence theorem therefore justifies passing the limit through the expectation:

$$\phi'(0+) = -\lim_{k\to\infty} \mathbb{E}_Q[h_{t_k}(r)] = 1 - \mathbb{E}_Q[r].$$

Because $\phi'(0+) \geq 0$, we obtain $\mathbb{E}_Q[r] \leq 1$. This is exactly the e-value bound under $P$, written after a change of density:

$$\mathbb{E}_P[E^*] = \int_{\{p^*>0\}} p\,\frac{g}{p^*}\,d\nu = \mathbb{E}_Q[r] \leq 1.$$

Since $P$ was arbitrary, $E^*$ is an e-value for $\mathcal{C}_0$.

*Step 2: e-validity under $P^*$ implies the numeraire property.* Let $E$ be any e-value for $\mathcal{C}_0$. Because $P^* \in \mathcal{C}_0$, we have $\mathbb{E}_{P^*}[E] \leq 1$. The ratio $E^*$ is positive $Q$-almost surely, and changing density once more gives

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] = \int_{\{g>0\}} Ep^*\,d\nu \leq \mathbb{E}_{P^*}[E] \leq 1.$$

The integral is restricted to $\{g > 0\}$; adding the omitted nonnegative part can only increase it, which explains the inequality. The final display is exactly the numeraire condition, and Jensen's inequality then gives log-optimality. □

Convexity is essential to this argument because the path $(1-t)P^* + tP$ must remain feasible. Thus minimizing $D(Q \,\|\, P_\theta)$ only over parameter values $\theta \in \Theta_0$ does not by itself prove that $g/p_{\theta^*}$ is an e-value. One must either minimize over the full null-mixture family, as in the proposition, or verify the null expectation bound directly.

In examples, we can often propose a null mixture and check the resulting likelihood ratio directly. The next proposition shows that checking its expectation under each original null distribution is enough to prove optimality; there is no need to solve the KL minimization problem first.

**Proposition A.3: Guess and verify**

Suppose densities exist, let $G$ be a prior on $\Theta_0$ with mixture density $p_G$, and suppose $p_G(X) > 0$ and $g(X) > 0$ almost surely under $Q$ and under every null distribution. If the candidate

$$E^* = \frac{g(X)}{p_G(X)} \qquad \text{satisfies} \qquad \mathbb{E}_{P_\theta}[E^*] \leq 1 \quad \text{for every } \theta \in \Theta_0,$$

then $E^*$ is the numeraire for $(\mathcal{H}_0, Q)$; in particular $E^*$ is log-optimal against $Q$.

**Proof of Proposition A.3**

The displayed condition says $E^*$ is an e-value. Let $E$ be any e-value for $\mathcal{H}_0$. Then

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] = \int E(x)\,\frac{p_G(x)}{g(x)}\,g(x)\,dx = \int E(x)\,p_G(x)\,dx = \int_{\Theta_0} \mathbb{E}_{P_\theta}[E]\,dG(\theta) \le 1,$$

where the interchange of the two integrals uses Tonelli's theorem for nonnegative integrands, and the final inequality holds because $\mathbb{E}_{P_\theta}[E] \le 1$ for every $\theta \in \Theta_0$. □

Thus we propose a null mixture, often a single boundary distribution, and check the expectation inequality over the whole null parameter set. If it holds, $E^* = g/p_G$ is valid and satisfies $\mathbb{E}_Q[\log(E/E^*)] \le 0$ for every competing e-value $E$. The following Gaussian example carries out this check.

**Example A.4: A one-sided Gaussian composite null**

Consider one observation $X \sim N(\mu, 1)$, with

$$H_0 : \mu \le 0 \qquad \text{and} \qquad Q = N(\mu_1, 1), \quad \mu_1 > 0.$$

The null distribution closest to $Q$ in KL divergence is the boundary distribution $N(0, 1)$, because for normals with common variance

$$D\{N(\mu_1, 1) \,\|\, N(\mu, 1)\} = \frac{(\mu_1 - \mu)^2}{2},$$

which is minimized over $\mu \le 0$ at $\mu = 0$. The resulting e-value is

$$E(X) = \exp\{\mu_1 X - \mu_1^2/2\}.$$

For every $\mu \le 0$,

$$\mathbb{E}_{P_\mu}[E(X)] = \exp\{\mu_1 \mu\} \le 1.$$

Thus the same likelihood ratio is valid for the whole half-line null, not only for the boundary point. Proposition A.3, applied with $G$ equal to the point mass at $\mu = 0$, also proves optimality: the display establishes that $E(X) = \exp\{\mu_1 X - \mu_1^2/2\}$ is the numeraire for the half-line null against $N(\mu_1, 1)$. No e-value for $H_0 : \mu \le 0$ has a larger expected logarithm under the alternative. The same computation, with the boundary point as the guess, works for one-parameter exponential families with one-sided nulls [147].

## 2. Optional Continuation and E-Processes

**Route: Advanced**

Optional continuation means that future data collection may depend on current evidence. A study may continue only if the interim evidence is promising, or a second group may decide to add data after seeing an inconclusive first result. The key requirement is that the next evidence multiplier must be valid after conditioning on the information already seen.

**Definition A.5: Conditional e-value**

Let $(\mathcal{F}_t)_{t\ge 0}$ be a filtration. A nonnegative $\mathcal{F}_t$-measurable random variable $E_t$ is a conditional e-value at time $t$ for $\mathcal{H}_0$ if

$$\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \le 1 \qquad \text{for every } P \in \mathcal{H}_0.$$

Recall the standard vocabulary of stochastic processes. An adapted sequence $(M_t)_{t\geq 0}$ is a *supermartingale* under $P$ if $\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}$ for every $t \geq 1$; it is a *martingale* if equality holds. A supermartingale cannot drift upward in conditional expectation.

**Proposition A.6: Products of conditional e-values**

If $E_t$ is a conditional e-value at every time $t$, then

$$M_t = \prod_{s=1}^{t} E_s, \qquad M_0 = 1,$$

is a nonnegative supermartingale under every $P \in \mathcal{H}_0$.

**Proof of Proposition A.6**

The product $M_{t-1}$ is determined by information available before time $t$, so it is $\mathcal{F}_{t-1}$-measurable. Therefore

$$\begin{aligned}\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}_P[M_{t-1}E_t \mid \mathcal{F}_{t-1}] \\ &= M_{t-1}\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}.\end{aligned}$$

Nonnegativity is inherited from the factors $E_t$. This is exactly the definition of a nonnegative supermartingale. □

A nonnegative supermartingale with $M_0 = 1$, used as accumulating evidence against $\mathcal{H}_0$, is called a *test supermartingale*, and products of conditional e-values provide one way to build one. In the literature, conditional e-values are also called *sequential e-values*; the two names describe the same object.

A conditional e-value is one new factor in the product; the test supermartingale is the accumulated product. More generally, we can require the accumulated evidence to be valid whenever we stop, without requiring a one-step conditional expectation bound for the process itself. This gives the definition of an e-process.

**Definition A.7: E-process**

As introduced in Section 5 of Chapter 8, an adapted sequence of nonnegative random variables $(M_t)_{t\geq 0}$ with $M_0 \leq 1$ is an *e-process* for $\mathcal{H}_0$ if

$$\mathbb{E}_P[M_\tau] \leq 1 \qquad \text{for every stopping time } \tau \text{ and every } P \in \mathcal{H}_0,$$

where on the event $\{\tau = \infty\}$ the stopped value is understood as $\liminf_{t\to\infty} M_t$.

In words, an e-process is a process whose value is a legitimate e-value at every stopping time, whichever stopping rule is used. Every test supermartingale is an e-process. The argument is inside the proof of Ville's inequality below: for a bounded horizon, $\mathbb{E}_P[M_{\tau\wedge K}] \leq 1$, and letting $K \to \infty$ with Fatou's lemma gives $\mathbb{E}_P[M_\tau] \leq 1$. Notice that no conditions on the stopping time beyond its being a stopping time for the stated filtration are needed. Nonnegativity permits the use of Fatou's lemma even when the stopping time is unbounded.

The converse fails: an e-process need not be a supermartingale under any null distribution. The equivalent characterization, whose proof is beyond this appendix, is that $(M_t)$ is an e-process for $\mathcal{H}_0$ if and only if for every $P \in \mathcal{H}_0$ there exists a test supermartingale $(M_t^P)$ under $P$, a

different one for each null distribution, with $M_t \le M_t^P$ $P$-almost surely for all $t$ [88, 147, 149]. For example, consider sequential universal inference. Under each null parameter $\theta$, suppose the observations are i.i.d. with density $p_\theta$. At time $j-1$, fit a probability density $\widehat{g}_{j-1}$ using only the observations seen so far, and define

$$M_t = \frac{\prod_{j\le t} \widehat{g}_{j-1}(X_j)}{\sup_{\theta\in\Theta_0} \prod_{j\le t} p_\theta(X_j)}, \qquad M_0 = 1.$$

Assume the displayed likelihoods and supremum are measurable, and set the ratio to zero if both numerator and denominator vanish. For each fixed null $\theta$, the denominator is at least its likelihood, so

$$M_t \le L_t^\theta := \prod_{j\le t} \frac{\widehat{g}_{j-1}(X_j)}{p_\theta(X_j)} \qquad P_\theta\text{-almost surely.}$$

Given the past, each factor in $L_t^\theta$ has null expectation at most one, by the same support-restricted density integral used for split likelihood ratios in Chapter 8. Hence $L_t^\theta$ is a test supermartingale, and its stopped-value bound also bounds $M_\tau$. Thus $M_t$ is an e-process. It need not itself be a supermartingale: its validity follows from a potentially different dominating process for each null parameter, rather than from one conditional-drift inequality for $M_t$ under every null. This distinction also matters for sequential exchangeability testing, where useful e-processes can exist even when the class of test supermartingales relative to the full-data filtration is too restrictive [147, 150]. A martingale defined using a coarser filtration, such as the conformal p-value filtration of Appendix B, need not remain a martingale after all past raw data are revealed.

**Theorem A.8: Ville's inequality**

If $(M_t)_{t\ge0}$ is a nonnegative supermartingale under a probability measure $\mathbb{P}$, with $M_0 = 1$, then for every $\alpha \in (0,1)$,

$$\mathbb{P}\left(\sup_{t\ge0} M_t \ge \frac{1}{\alpha}\right) \le \alpha.$$

**Proof of Theorem A.8**

Let

$$\tau = \inf\{t : M_t \ge 1/\alpha\},$$

with $\inf \varnothing = \infty$. Fix a deterministic horizon $K$. The stopped time $\tau \wedge K$ is bounded. For a bounded stopping time $\rho \le K$, the stopped process $M_{t\wedge\rho}$, $0 \le t \le K$, is still a supermartingale; iterating the one-step inequalities gives $\mathbb{E}[M_\rho] \le \mathbb{E}[M_0]$. Applying this with $\rho = \tau \wedge K$ gives

$$\mathbb{E}[M_{\tau\wedge K}] \le \mathbb{E}[M_0] = 1.$$

On the event $\{\tau \le K\}$, the stopped value is at least $1/\alpha$. On the complementary event it is nonnegative. Hence

$$1 \ge \mathbb{E}[M_{\tau\wedge K}] \ge \frac{1}{\alpha}\mathbb{P}(\tau \le K).$$

Thus $\mathbb{P}(\tau \le K) \le \alpha$. The events $\{\tau \le K\}$ increase to $\{\tau < \infty\}$ as $K \to \infty$, so continuity from below bounds the probability of the crossing event $\{\exists\, t : M_t \ge 1/\alpha\}$ by $\alpha$. Finally, on the event $\{\sup_t M_t \ge 1/\alpha\}$, the process exceeds $1/\alpha'$ at some finite time for every $\alpha' > \alpha$; applying the crossing bound at level $\alpha'$ and letting $\alpha'$ decrease to $\alpha$ gives the claim as stated. □

Ville's inequality extends beyond supermartingales to every e-process, with a proof that is now immediate. If $(M_t)$ is an e-process for $\mathcal{H}_0$, fix $P \in \mathcal{H}_0$ and let $\tau = \inf\{t : M_t \geq 1/\alpha\}$. The crossing time $\tau$ is a stopping time, $M_\tau \geq 1/\alpha$ on the event $\{\tau < \infty\}$, and $\mathbb{E}_P[M_\tau] \leq 1$ by the definition of an e-process, so Markov's inequality gives

$$\mathbb{P}_P\left(\exists\, t \geq 0 : M_t \geq \frac{1}{\alpha}\right) = \mathbb{P}_P(\tau < \infty) \leq \alpha\, \mathbb{E}_P[M_\tau] \leq \alpha,$$

and the supremum form follows by the same limiting argument as in the proof of Theorem A.8. Conversely, every valid sequential rejection rule can be represented by thresholding an e-process, as the next proposition shows.

**Proposition A.9: Sequential tests are thresholded e-processes**

Call a nondecreasing binary process $(\phi_t)_{t\geq 1}$ adapted to $(\mathcal{F}_t)$, with $\phi_0 = 0$, a level-$\alpha$ sequential test for $\mathcal{H}_0$ if $\phi_t = 1$ means that the null has been rejected at or before time $t$, and

$$\sup_{P \in \mathcal{H}_0} \mathbb{P}_P(\exists\, t \geq 1 : \phi_t = 1) \leq \alpha.$$

Then $M_t = \phi_t/\alpha$ is an e-process for $\mathcal{H}_0$, and the test rejects by time $t$ exactly when $M_t \geq 1/\alpha$. For a stopping time that never stops, $\phi_\tau$ is read as the limit $\lim_t \phi_t$, which exists by monotonicity.

**Proof of Proposition A.9**

The process $M_t$ takes only the values 0 and $1/\alpha$. For any stopping time $\tau$ and any $P \in \mathcal{H}_0$,

$$\mathbb{E}_P[M_\tau] = \frac{\mathbb{P}_P(\phi_\tau = 1)}{\alpha} \leq \frac{\mathbb{P}_P(\exists\, t : \phi_t = 1)}{\alpha} \leq 1.$$

□

Together, the two statements give a correspondence: every e-process yields a level-$\alpha$ sequential test by thresholding at $1/\alpha$, and every level-$\alpha$ sequential test arises this way.

**Optional stopping versus optional continuation.** There are two ways to adapt data collection. *Optional stopping* concerns one monitored study: the analyst may stop at any stopping time $\tau$ and report $M_\tau$. *Optional continuation* concerns a chain of studies: after seeing the current evidence, someone, possibly a different laboratory, decides whether to collect more data, and how much. Both are covered, the first by Ville's inequality, the second by the product rule for conditional e-values, because each factor is only required to be valid conditionally on everything already seen, including the decision to continue. Combining p-values across such a chain is much harder: whether the second study exists at all depends on the first p-value, so even the number of p-values to be combined is random, and the standard combination rules do not apply. Conditional e-values can be multiplied under the stated condition, even when the decision to collect another dataset depends on earlier results.

The next example shows why validity at every stopping time is useful.

**Example A.10: The design does not need to be known**

A collaborator reports eight coin tosses,

$$1,\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,$$

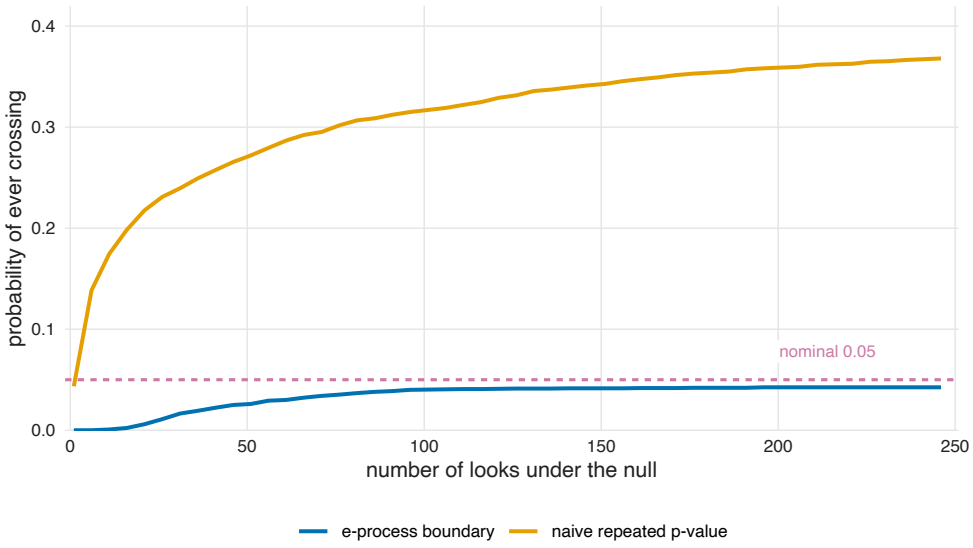



FIGURE A.1. A null simulation with 4,000 independent runs and 250 looks, adding one independent $N(0,1)$ observation at each look. The likelihood-ratio e-process uses the fixed alternative mean $\mu_1 = 0.35$. Repeatedly checking a fixed-sample one-sided Gaussian p-value inflates the probability of ever crossing 0.05. The likelihood-ratio e-process boundary stays controlled because Ville's inequality applies to the whole monitored path.

seven successes in all, and asks whether the coin favors success. Test $H_0$ : the success probability is $1/2$. A conventional fixed-design p-value requires a sampling and stopping design, which we were not told. If the design was "toss exactly $n = 8$ times," the one-sided binomial p-value is

$$\mathbb{P}_0(\text{at least 7 successes in 8 tosses}) = \frac{\binom{8}{7} + \binom{8}{8}}{2^8} = \frac{9}{256} \approx 0.035.$$

If the design was "toss until five successes in a row appear," which produces exactly these data as well, the relevant p-value is the null probability of stopping within eight tosses. The first run of five successes can be completed at toss 5, 6, 7, or 8, with null probabilities $2^{-5}$, $2^{-6}$, $2^{-6}$, $2^{-6}$, so

$$\mathbb{P}_0(\text{stop by toss 8}) = \frac{1}{32} + \frac{3}{64} = \frac{5}{64} \approx 0.078.$$

The same observations therefore lead to different decisions at level 0.05, depending on the sampling rule. Now compute the likelihood-ratio e-process against a prespecified alternative with success probability $\theta_1 > 1/2$:

$$M_8 = \prod_{j=1}^{8} \frac{\theta_1^{X_j}(1-\theta_1)^{1-X_j}}{1/2} = 2^8\,\theta_1^7(1-\theta_1),$$

which is approximately 12.2 when $\theta_1 = 0.9$. Since $(M_t)$ is a test supermartingale under the null, it gives a valid e-value at any stopping time of the recorded sequence, including a fixed sample size, stopping on a streak, or stopping when evidence exceeds a threshold.

If unsuccessful runs of the whole experiment are discarded and only the best run is reported, neither calculation accounts for that selection.

A zero factor sends a plain product to zero permanently. The predictable hedge $1-\lambda_s+\lambda_s E_s$ from Section 5 of Chapter 8 avoids this when $\lambda_s < 1$, while retaining conditional validity. Its affine form has a further justification: among coordinatewise nondecreasing functions that transform every e-value into another e-value, every such function is dominated by some $e \mapsto 1-\lambda+\lambda e$, $\lambda \in [0,1]$ [147, 188]. For example, $\sqrt{e} \le (1+e)/2$, so the hedge with $\lambda = 1/2$ reports at least as much evidence as taking a square root, and unlike the square root it stays positive at $e = 0$. Data-driven choices of $\lambda_s$, tuned on the past to improve growth, are developed by Waudby-Smith and Ramdas [193].

## 3. Universality of e-BH

**Route: Advanced**

The relationship between FDR control and compound e-values runs in both directions. The reverse direction gives a useful structural representation: every FDR-controlling procedure, whatever its construction, can be expressed as e-BH applied to a suitable vector of compound e-values.

**Proposition A.11: Any FDR procedure yields compound e-values**

Let $\mathcal{D}$ be any multiple testing procedure with rejection set $\mathcal{R}_\mathcal{D} \subseteq \{1,\dots,m\}$, and suppose it controls FDR at level $\alpha$ under every distribution in the model. Define

$$E_i = \frac{m}{\alpha} \cdot \frac{\mathbf{1}\{i \in \mathcal{R}_\mathcal{D}\}}{R_\mathcal{D} \vee 1}, \qquad i = 1,\dots,m,$$

where $R_\mathcal{D} = |\mathcal{R}_\mathcal{D}|$. Then $E_1,\dots,E_m$ are compound e-values.

**Proof of Proposition A.11**

Under any distribution in the model,

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] = \frac{m}{\alpha}\,\mathbb{E}\left[\frac{|\mathcal{R}_\mathcal{D} \cap \mathcal{I}_0|}{R_\mathcal{D} \vee 1}\right] = \frac{m}{\alpha}\,\mathrm{FDR}_\mathcal{D} \le \frac{m}{\alpha}\,\alpha = m.$$

□

**Theorem A.12: Universality of e-BH**

With $E_1,\dots,E_m$ as in Proposition A.11, the e-BH procedure at level $\alpha$ applied to $E_1,\dots,E_m$ rejects exactly $\mathcal{R}_\mathcal{D}$. Consequently, every FDR-controlling procedure coincides with e-BH applied to some compound e-values.

**Proof of Theorem A.12**

Write $r = R_\mathcal{D}$. If $r = 0$, all induced e-values are zero and both procedures reject nothing. If $r > 0$, the $r$ rejected hypotheses have $E_i = m/(\alpha r)$ and all others have $E_i = 0$, so the ordered values are $E_{(k)} = m/(\alpha r)$ for $k \le r$ and $E_{(k)} = 0$ for $k > r$. The e-BH condition

$E_{(k)} \geq m/(\alpha k)$ holds at $k = r$, with equality, and fails for every $k > r$, where the ordered e-value is zero. Hence the e-BH rank is $\widehat{k} = r$, and the rejected hypotheses are the $r$ hypotheses with positive induced evidence, namely $\mathcal{R}_{\mathcal{D}}$. □

The theorem gives a representation of an existing FDR procedure: its FDR bound establishes condition (A) for its induced evidence vector. The threshold template in Section 7 of Chapter 8 is the constructive version of this statement, deriving the induced e-values in closed form for threshold-based rules. The representation also allows us to combine several FDR procedures into one rejection list: convert each procedure $\ell$ into compound e-values via Proposition A.11, form a convex combination $\sum_\ell w_\ell E_i^{(\ell)}$ with fixed nonnegative weights summing to one, which is again compound by linearity of expectation, and run e-BH once on the combined vector. This is a special case of the combining results developed in Section 8 of Chapter 8. Averaging over repeated runs of a randomized procedure, such as the derandomization of knockoff or data-splitting methods in Chapter 9, is a leading application [92, 152]. A further result concerns admissibility. If the original procedure is admissible among FDR procedures, meaning that no other level-$\alpha$ FDR procedure rejects a superset of its rejections on every dataset with strict containment on some, the induced compound e-values can be rescaled so that their aggregate null budget is exactly $m$; see Ignatiadis et al. [92] and Chapter 9 of Ramdas and Wang [147].

## 4. Asymptotic E-Values

**Route: Advanced**

Several methods in this book use statistics that are only asymptotically normal, including the debiased lasso statistics of Chapter 11. A CLT gives convergence of distribution functions at continuity points of the limiting distribution. It does not by itself give convergence of the expectation of an unbounded exponential transform. Finite variance therefore does not justify simply using the normal moment-generating function as an exact calibration for a finite-sample statistic. We need a relaxed notion of e-value and an explicit statement of the resulting testing error bound.

**Definition A.13: Approximate and asymptotic e-values**

Let $\varepsilon : \mathcal{H}_0 \to [0, \infty)$ and $\delta : \mathcal{H}_0 \to [0, 1]$. A nonnegative statistic $E$ is an $(\varepsilon, \delta)$-*approximate e-value* for $\mathcal{H}_0$ if

$$\mathbb{E}_P[\min\{E, t\}] \leq 1 + \varepsilon(P) + \delta(P)t \qquad \text{for every } t \geq 0 \text{ and every } P \in \mathcal{H}_0.$$

When $\varepsilon$ and $\delta$ are constants, the same bound applies uniformly to every $P \in \mathcal{H}_0$.
A sequence $E^{(n)}$ is a sequence of *pointwise asymptotic e-values* if each $E^{(n)}$ is $(\varepsilon_n, \delta_n)$-approximate and, for every fixed $P \in \mathcal{H}_0$,

$$\varepsilon_n(P) \to 0 \qquad \text{and} \qquad \delta_n(P) \to 0.$$

It is a sequence of *uniformly asymptotic e-values* if, in addition,

$$\sup_{P \in \mathcal{H}_0} \varepsilon_n(P) \to 0 \qquad \text{and} \qquad \sup_{P \in \mathcal{H}_0} \delta_n(P) \to 0.$$

Unless uniformity is stated explicitly, *asymptotic* means pointwise.

The two approximation parameters play different roles. The following sufficient condition helps interpret them. For a fixed $P$, suppose $\mathbb{P}_P(A_P) \leq \delta(P)$ and $\mathbb{E}_P[E\,\mathbf{1}_{A_P^c}] \leq 1 + \varepsilon(P)$ for

some event $A_P$. Then for every $t$,

$$\mathbb{E}_P[\min\{E,t\}] \le \mathbb{E}_P[E\,\mathbf{1}_{A_P^c}] + t\,\mathbb{P}_P(A_P) \le 1 + \varepsilon(P) + \delta(P)t,$$

so $E$ is $(\varepsilon,\delta)$-approximate. Here $\varepsilon(P)$ bounds the excess expectation outside $A_P$, while $\delta(P)$ bounds the probability of $A_P$, on which $E$ may be arbitrarily large. The definition itself requires only the truncated expectation bound, not an explicit choice of such an event. The truncation at $t$ in the definition is what makes the relaxation compatible with e-BH, as the proposition at the end of this section shows: e-BH never uses the value of an e-value beyond $m/\alpha$.

The CLT supplies asymptotic e-values directly.

**Proposition A.14: E-values from asymptotically normal statistics**

Let $Z^{(n)}$ be a test statistic with $Z^{(n)} \Rightarrow N(0,1)$ under every $P \in \mathcal{H}_0$, where $\Rightarrow$ denotes convergence in distribution. Fix $\lambda > 0$ and set

$$E^{(n)} = \exp\left\{\lambda Z^{(n)} - \frac{\lambda^2}{2}\right\}.$$

Then $E^{(n)}$ is a sequence of pointwise asymptotic e-values. Moreover, for every $t \ge 0$ and every $P \in \mathcal{H}_0$,

$$\limsup_{n\to\infty} \mathbb{E}_P[\min\{E^{(n)},t\}] \le 1.$$

**Proof of Proposition A.14**

For $Z \sim N(0,1)$, the random variable $\exp\{\lambda Z - \lambda^2/2\}$ has mean one; it is the likelihood ratio of $N(\lambda,1)$ against $N(0,1)$. The continuous mapping theorem therefore gives convergence in distribution of $E^{(n)}$, under each fixed $P$, to this exact e-value. The distributional criterion for asymptotic e-values, Proposition 10.7(i) of Ramdas and Wang [147], gives the pointwise asymptotic conclusion. In addition, the function $z \mapsto \min\{\exp(\lambda z - \lambda^2/2), t\}$ is bounded and continuous, so convergence in distribution gives

$$\mathbb{E}_P[\min\{E^{(n)},t\}] \to \mathbb{E}[\min\{\mathrm{e}^{\lambda Z-\lambda^2/2},t\}] \le \mathbb{E}[\mathrm{e}^{\lambda Z-\lambda^2/2}] = 1.$$

□

Pointwise convergence does not supply a uniform guarantee over $P \in \mathcal{H}_0$. If the normal approximation is uniform over a specified null class in the sense required by the distributional criterion, then the construction is uniformly asymptotic. This additional uniformity must be verified for that class. Uniform integrability for each fixed $P$ can strengthen the pointwise expectation conclusion, but does not by itself make the convergence uniform over $\mathcal{H}_0$; see Chapter 10 of Ramdas and Wang [147].

**Example A.15: A universal t-test e-value**

Test the nonparametric null $\mathbb{E}[X] = 0$ with unknown finite nonzero variance, from i.i.d. observations $X_1,\dots,X_n$. Take the self-normalized statistic

$$Z^{(n)} = \frac{\sqrt{n}\,\bar{X}_n}{S_n}, \qquad S_n^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2,$$

Set $Z^{(n)} = 0$ when $S_n = 0$. Since the variance is positive, this event has probability tending to zero. The statistic converges in distribution to $N(0,1)$ under every null distribution, by the CLT and Slutsky's lemma. Then

$$E^{(n)} = \exp\left\{\lambda Z^{(n)} - \frac{\lambda^2}{2}\right\}$$

satisfies the pointwise guarantee of Proposition A.14 under every null distribution. Thus it is a pointwise asymptotic e-value; the argument does not claim an exact finite-sample expectation bound. If the null is restricted to a class over which an appropriate self-normalized CLT has been proved uniformly, the same construction is uniformly asymptotic. Under an alternative $Q$ with $0 < \mathbb{E}_Q[X] < \infty$ and $0 < \mathbb{E}_Q[X^2] < \infty$, the exponent diverges at order $\sqrt{n}$, so the evidence diverges and the test that rejects when $E^{(n)} \geq 1/\alpha$ is consistent.

The approximate expectation bound gives an explicit bound on the increase in FDR.

**Proposition A.16: e-BH with approximate e-values**

Let $E_1, \ldots, E_m$ be nonnegative, and suppose each true null $i \in \mathcal{I}_0$ satisfies the $(\varepsilon, \delta)$-approximate condition for fixed constants $\varepsilon \geq 0$ and $\delta \in [0,1]$. Then e-BH at level $\alpha$ satisfies

$$\mathrm{FDR} \leq \frac{m_0}{m}(1+\varepsilon)\alpha + m_0\delta \leq (1+\varepsilon)\alpha + m\delta.$$

**Proof of Proposition A.16**

Let $\mathcal{R}$ be the e-BH rejection set and $R = |\mathcal{R}| \geq 1$ without loss of generality. Every rejected $i$ satisfies $E_i \geq m/(\alpha R)$, and since $R \leq m$, also $\min\{E_i, m/\alpha\} \geq m/(\alpha R)$. Therefore, path by path,

$$\frac{\mathbf{1}\{i \in \mathcal{R}\}}{R} \leq \frac{\alpha}{m} \min\left\{E_i, \frac{m}{\alpha}\right\}.$$

Summing over true nulls and taking expectations,

$$\mathrm{FDR} \leq \frac{\alpha}{m} \sum_{i \in \mathcal{I}_0} \mathbb{E}\left[\min\left\{E_i, \frac{m}{\alpha}\right\}\right] \leq \frac{\alpha}{m} m_0 \left(1 + \varepsilon + \delta\,\frac{m}{\alpha}\right) = \frac{m_0}{m}(1+\varepsilon)\alpha + m_0\delta.$$

□

The contribution of $\delta$ is multiplied by the number of true nulls. Thus an approximation error that is small for one test may matter when many tests are performed. Consequently, if all true-null coordinates obey a common $(\varepsilon_n, \delta_n)$ bound, e-BH has asymptotic FDR control at level $\alpha$ provided $\varepsilon_n \to 0$ and $m\,\delta_n \to 0$. When $m$ grows with $n$, we therefore need $\delta_n = o(1/m)$, together with a bound that holds for all true-null coordinates under consideration. A separate pointwise CLT for each coordinate does not establish this common bound. These are the additional conditions needed to use CLT-based e-values with e-BH.

## 5. Online FDR Control

**Route: Research frontier**

The procedures above handle a fixed family of $m$ hypotheses. Online testing is different. Hypotheses arrive one at a time, and the decision on $H_t$ must be made before seeing future hypotheses. Examples include monitoring many experiments, testing safety signals as they appear, or screening a long sequence of scientific questions.

Let $\mathcal{F}_t$ be the information available after the first $t$ tests. At time $t$, the procedure may use $\mathcal{F}_{t-1}$ to choose the testing level for $H_t$, but it may not use the current p-value or e-value before choosing the level. Let

$$I_t = \mathbf{1}\{H_t \text{ is rejected}\}, \qquad R_t = \sum_{s=1}^{t} I_s, \qquad V_t = \sum_{s \le t : s \in \mathcal{I}_0} I_s.$$

Here $\mathcal{I}_0 \subseteq \{1, 2, \ldots\}$ denotes the true null times. For a deterministic finite horizon $K$, the online FDR is

$$\mathrm{FDR}(K) = \mathbb{E}\left[\frac{V_K}{R_K \vee 1}\right].$$

For an almost-surely finite data-dependent monitoring time $\tau$, the stopped online FDR is

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right].$$

A finite-horizon guarantee asks for $\mathrm{FDR}(K) \le \alpha$ for each fixed $K$. A stopping-time guarantee asks for the same bound when the time at which we look can depend on the past.

Before seeing the current evidence, choose a predictable testing level $\alpha_t$. For a p-value, reject when $p_t \le \alpha_t$. For an e-value, reject when $E_t \ge 1/\alpha_t$, with the convention that no rejection is made if $\alpha_t = 0$. The problem is to choose the sequence $\alpha_t$ so that early false discoveries do not spend the entire error budget, while genuine discoveries allow the procedure to keep testing with useful power.

**P-value online rules: alpha wealth, LORD, and SAFFRON.** This subsection explains how these rules allocate testing levels. The wealth updates and LORD formula below are illustrative; they are not a complete specification of a procedure with an FDR guarantee. Such a guarantee also requires the reward constraints and dependence assumptions of the particular rule. Complete specifications are given by Javanmard and Montanari [98] for LORD and Ramdas et al. [148] for SAFFRON. The e-LOND theorem in the next subsection gives a fully specified rule and proof.

For p-values, the basic null condition is conditional super-uniformity: for each true null $t$,

$$\mathbb{P}(p_t \le u \mid \mathcal{F}_{t-1}) \le u, \qquad 0 \le u \le 1.$$

Here predictable means that $\alpha_t$ is chosen from the past information $\mathcal{F}_{t-1}$: it may depend on previous p-values, e-values, rejections, and covariates already revealed, but not on the current p-value $p_t$ before $H_t$ is tested. If $H_t$ is tested at a predictable level $\alpha_t$, then

$$\mathbb{P}(I_t = 1 \mid \mathcal{F}_{t-1}) \le \alpha_t$$

for a true null. This is the basic one-step inequality. Online FDR rules turn it into a many-step guarantee by deciding how much level can be spent at each time.

Alpha wealth is bookkeeping for this spending. The wealth $W_t$ is not a new error rate; it is an internal balance that limits future testing levels. Start with wealth $W_0 \le \alpha$. Testing at

level $\alpha_t$ spends some wealth. A rejection earns a reward, which permits more testing later. A simplified update is

$$W_t = W_{t-1} - \alpha_t + b_t I_t,$$

where $b_t \geq 0$ is a predictable reward. The statistical idea is: spend a small amount of Type I error budget now, and allow larger future levels only after past discoveries have paid for them. Different online FDR rules differ mainly in how they choose $\alpha_t$ and the rewards.

LORD is the first main example. It chooses testing levels from two sources: initial wealth and rewards from past discoveries. Choose nonnegative numbers $\gamma_1, \gamma_2, \ldots$ with $\sum_{j\geq 1} \gamma_j = 1$. Think of $\gamma_j$ as a schedule that spreads one unit of wealth over future tests. If $D_1 < D_2 < \cdots$ denote the discovery times revealed so far, a LORD-style level has the following simplified form:

$$\alpha_t = W_0 \gamma_t + \sum_{D_j < t} b_j \gamma_{t-D_j}.$$

The full LORD rule imposes additional constraints on the rewards $b_j$ so the wealth account never overspends. This displayed formula is predictable because it depends only on previous rejections. Operationally, each discovery releases a new stream of future testing levels.

The accounting argument for LORD-type rules. For true nulls,

$$\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] \leq \alpha_t.$$

Therefore, for any deterministic horizon $K$,

$$\mathbb{E}[V_K] \leq \mathbb{E}\left[\sum_{t\leq K : t\in\mathcal{I}_0} \alpha_t\right] \leq \mathbb{E}\left[\sum_{t\leq K} \alpha_t\right].$$

The wealth constraints limit cumulative spending in terms of the initial wealth and rewards from $R_K$ discoveries. To bound FDR, we must also account for the random denominator $R_K \vee 1$, which depends on the same p-values as $V_K$. The standard guarantee for monotone LORD rules assumes that true-null p-values are independent of each other and of the non-null p-values [98]. Here monotonicity means that changing a previous nonrejection to a rejection cannot reduce the current testing level. Conditional super-uniformity gives the one-step inequality above, but does not by itself complete this FDR argument.

SAFFRON is the next idea after LORD. LORD treats every tested hypothesis as a place where alpha might have been wasted. SAFFRON tries to estimate that waste online. Before seeing $p_t$, choose a predictable candidate threshold $\lambda_t$ with $\alpha_t \leq \lambda_t \leq 1$, and define

$$A_t = \mathbf{1}\{p_t \leq \lambda_t\}.$$

A candidate is not necessarily rejected; it is a hypothesis whose p-value is at most the chosen threshold $\lambda_t$. Very large p-values are not candidates. SAFFRON uses the number of candidates and discoveries to estimate how much alpha was effectively spent on nulls, and then reallocates the saved budget to later tests. This is the online analogue of Storey's offline idea: estimate how many hypotheses look null, then use the saved budget for power.

To describe the rule, track four objects at every time: the predictable test level $\alpha_t$, the predictable candidate threshold $\lambda_t$, the candidate indicator $A_t$, and the rejection indicator $I_t = \mathbf{1}\{p_t \leq \alpha_t\}$. The full SAFFRON formula specifies how candidates reduce the estimate of wasted alpha and how discoveries replenish wealth. Its proof follows the LORD template, but replaces raw spending $\sum \alpha_t$ by an online estimate of alpha spent on true nulls.

**E-value online rules: e-LOND.** The following version of e-LOND uses conditional e-values. Its proof applies the conditional expectation bound to each false rejection. This assumption allows dependence on past observations, but it is stronger than requiring only $\mathbb{E}[E_t] \le 1$ at each null time. The testing level is proportional to the number of discoveries made so far.

Choose a deterministic spending schedule $\gamma_t$ with total mass at most one. At time $t$, enlarge the level in proportion to the number of previous discoveries:

$$\alpha_t = \alpha\gamma_t(R_{t-1} + 1).$$

Then reject large e-values. More past discoveries permit a larger current level, but the factor $\gamma_t$ keeps the total evidence budget summable.

**Theorem A.17: A simple e-LOND bound**

Let $E_t$ be nonnegative and $\mathcal{F}_t$-measurable. For every true null $t$, assume

$$\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le 1.$$

Let $\gamma_t \ge 0$ be deterministic with $\sum_{t\ge1} \gamma_t \le 1$. Define the predictable testing level

$$\alpha_t = \alpha\,\gamma_t\,(R_{t-1} + 1).$$

Reject $H_t$ when $E_t \ge 1/\alpha_t$, with no rejection if $\alpha_t = 0$. Then for every almost-surely finite stopping time $\tau$,

$$\mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right] \le \alpha.$$

**Proof of Theorem A.17**

Let $I_t = \mathbf{1}\{E_t \ge 1/\alpha_t\}$. If $t \le \tau$ and $I_t = 1$, then the final number of rejections by time $\tau$ is at least $R_{t-1} + 1$. Therefore,

$$\frac{V_\tau}{R_\tau \vee 1} \le \sum_{t\in\mathcal{I}_0} \frac{\mathbf{1}\{t \le \tau\}I_t}{R_{t-1} + 1}.$$

The rejection event implies $E_t \ge 1/\alpha_t$, so, path by path,

$$I_t \le \alpha_t E_t$$

when $\alpha_t > 0$; when $\alpha_t = 0$, both sides are zero by convention. The factor

$$\frac{\mathbf{1}\{t \le \tau\}\alpha_t}{R_{t-1} + 1} = \mathbf{1}\{t \le \tau\}\alpha\gamma_t$$

is $\mathcal{F}_{t-1}$-measurable. Indeed, for a stopping time, $\{t \le \tau\} = \{\tau \ge t\}$ is determined by whether the procedure had already stopped before time $t$, and $\alpha_t$, $R_{t-1}$, and $\gamma_t$ are predictable. Using Tonelli's theorem for the nonnegative sum, followed by the tower property and conditional e-value validity, gives

$$\begin{aligned}
\mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right] &\le \sum_{t\in\mathcal{I}_0} \mathbb{E}\left[\mathbf{1}\{t \le \tau\}\frac{\alpha_t}{R_{t-1} + 1}E_t\right] \\
&= \sum_{t\in\mathcal{I}_0} \mathbb{E}\left[\mathbf{1}\{t \le \tau\}\alpha\gamma_t\mathbb{E}(E_t \mid \mathcal{F}_{t-1})\right] \\
&\le \sum_{t\in\mathcal{I}_0} \mathbb{E}[\mathbf{1}\{t \le \tau\}\alpha\gamma_t] \le \alpha\sum_{t\ge1}\gamma_t \le \alpha.
\end{aligned}$$

□

## 6. Confidence Sequences

**Route: Advanced**

Testing asks whether a null value should be rejected. Estimation asks which parameter values remain plausible. In a monitored experiment, the interval should be valid no matter when we look.

**Definition A.18: Confidence sequence**

Let $\theta$ be a target parameter and $(\mathcal{F}_t)_{t\geq 0}$ a filtration. A sequence of random sets $(C_t)_{t\geq 1}$, adapted in the sense that $\{\vartheta \in C_t\} \in \mathcal{F}_t$ for every candidate value $\vartheta$, is a $(1-\alpha)$ confidence sequence if

$$\mathbb{P}_\theta\{\theta \in C_t \text{ for every } t \geq 1\} \geq 1-\alpha.$$

The universal quantifier over $t$ is the key difference from ordinary fixed-time confidence intervals. A 95% interval at each fixed time does not imply 95% simultaneous coverage over all times.

An equivalent formulation replaces the universal quantifier by stopping times: $(C_t)_{t\geq 1}$ is a $(1-\alpha)$ confidence sequence if and only if

$$\mathbb{P}_\theta(\theta \in C_\tau) \geq 1-\alpha \qquad \text{for every stopping time } \tau.$$

One direction is immediate, since simultaneous coverage implies coverage at any random time. For the converse, apply the hypothesis to the stopping time $\tau = \inf\{t : \theta \notin C_t\}$, with $C_\infty$ read as the whole parameter space. Adaptedness makes this first-exclusion time a stopping time, and coverage at this $\tau$ forces $\mathbb{P}_\theta(\tau = \infty) \geq 1-\alpha$, which is simultaneous coverage. The stopping-time form is how confidence sequences are used in practice: an experiment monitored until an arbitrary data-dependent moment still reports a valid interval.

**Theorem A.19: Confidence sequence by inverting e-processes**

For each candidate value $\theta_0$, suppose $(M_t^{\theta_0})_{t\geq 0}$ is an e-process under every distribution in the null model $H_0 : \theta = \theta_0$. Define

$$C_t = \{\theta_0 : M_t^{\theta_0} < 1/\alpha\}.$$

Then $(C_t)_{t\geq 1}$ is a $(1-\alpha)$ confidence sequence.

**Proof of Theorem A.19**

The true value $\theta$ is excluded at some time if and only if

$$M_t^\theta \geq 1/\alpha$$

for at least one $t$. Since $M_t^\theta$ is an e-process under the true distribution, Ville's inequality bounds the probability of this event:

$$\mathbb{P}_\theta\left(\sup_{t\geq 1} M_t^\theta \geq 1/\alpha\right) \leq \alpha. \qquad \square$$

**Example A.20: A simple bounded-mean confidence sequence**

Let $X_1, X_2, \ldots \in [0,1]$ be independent with common mean $\theta$. The fixed-time form of Hoeffding's inequality says that, for each $t$ and each $\eta_t \in (0,1)$,

$$\mathbb{P}_\theta\left(|\bar{X}_t - \theta| > \sqrt{\frac{\log\{2/\eta_t\}}{2t}}\right) \le \eta_t.$$

Choose $\eta_t = 6\alpha/(\pi^2 t^2)$, so that $\sum_{t=1}^{\infty} \eta_t = \alpha$. By the union bound,

$$C_t = \left[\bar{X}_t - \sqrt{\frac{\log\{\pi^2 t^2/(3\alpha)\}}{2t}}, \bar{X}_t + \sqrt{\frac{\log\{\pi^2 t^2/(3\alpha)\}}{2t}}\right] \cap [0,1]$$

is a $(1-\alpha)$ confidence sequence. This construction is conservative, but it is fully self-contained and shows the simultaneous-coverage idea.

Betting-style confidence sequences use the e-process inversion theorem more directly. Suppose $X_s \in [0,1]$ and $\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = \theta$ for every $s$; independent observations with common mean $\theta$ satisfy this condition when the filtration contains only past observations and independent randomization. Fix a candidate value $\theta_0 \in (0,1)$ and choose predictable bets

$$\lambda_t \in \left[-\frac{1}{1-\theta_0}, \frac{1}{\theta_0}\right].$$

Define

$$M_t^{\theta_0} = \prod_{s=1}^{t} \{1 + \lambda_s(X_s - \theta_0)\}.$$

The bounds on $\lambda_s$ keep each factor nonnegative for every $X_s \in [0,1]$. Under $\theta = \theta_0$,

$$\mathbb{E}[1 + \lambda_s(X_s - \theta_0) \mid \mathcal{F}_{s-1}] = 1,$$

so $M_t^{\theta_0}$ is a nonnegative martingale. Inverting these processes over $\theta_0$ gives a confidence sequence. More refined choices of the bets adapt to the observed variance and often give much shorter intervals than the simple union-bound construction.

These factors also follow from the conditional e-values $X_s/\theta_0$ and $(1-X_s)/(1-\theta_0)$, both with conditional expectation one under $\theta = \theta_0$. Shrinking the first toward one gives the factors with $\lambda_s \in [0, 1/\theta_0]$; shrinking the second produces those with $\lambda_s \in [-1/(1-\theta_0), 0]$. Together the two shrinkage families give the displayed range of nonnegative linear bets. The earlier domination result justifies affine shrinkage among nondecreasing transformations valid for every e-value; it does not say that all valid betting factors for bounded observations must be linear.

For the connection to selected intervals in Chapter 12, consider a family of e-values $E(\theta_0)$. It defines an *e-confidence set* $\{\theta_0 : E(\theta_0) < 1/\alpha\}$; it is an interval when this set is connected. Inverted e-processes retain this form at a stopping time. For a selected set $S$, the expected fraction of intervals that miss their targets is $\mathrm{FCR} = \mathbb{E}[|\{i \in S : \theta_i \notin C_i\}|/(|S| \vee 1)]$. Reporting each e-confidence interval at level $\alpha_i = \delta|S|/m$, out of $m$ candidates, controls FCR at level $\delta$ under arbitrary selection rules and arbitrary dependence, by the same indicator bound and expectation argument used in Theorem 8.15 [195].

APPENDIX B

# Advanced Conformal Theory

**Route: Advanced**

Chapter 10 contains the core full- and split-conformal algorithms and their coverage statements. This appendix gives the proofs and extensions that can be deferred on a first reading. The sections can also be read separately. Conditional and selected-point coverage use the rank argument and the symmetric-conditioning lemma from that chapter. The PRDS proof connects to Chapter 6, and online betting uses the conditional e-value products of Chapter 8. The final LTT construction uses FWER control from Chapter 4. Each extension below states the additional sampling or modeling assumptions it needs.

## 1. What Conditional Coverage Means

**Route: Advanced**

Section 5 of Chapter 10 distinguished several conditioning targets. We now state their guarantees precisely. The important question is which observations remain random in each probability statement.

Table B.1 compares the conditioning statements used below. The rows describe different targets; a guarantee for one row does not automatically establish a guarantee for another. In particular, the term "training-conditional" includes the calibration data, whereas conditioning on the proper training set alone does not.

TABLE B.1. What is held fixed in different coverage statements. The proper training set is $D_1$, and the calibration set is $D_2$.

| Coverage target | Conditioning information | What is still random |
|---|---|---|
| Marginal | None | Training data, calibration data, and the test pair |
| Given proper training | $D_1$ | Calibration data and the test pair |
| Training-conditional | $D_1, D_2$ | The test pair |
| Test-conditional | $X_{n+1} = x$ | Observed data and the test response at $x$ |
| Selection-conditional | The test point is selected | Observed data and the test pair, conditional on selection |

First, split conformal gives a coverage bound conditional on the proper training set $D_1$. Once $D_1$ is fixed, the fitted predictor is fixed, and the calibration scores and the future test score are exchangeable. Thus Theorem 10.3 actually proves

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) \mid D_1\} \geq 1 - \alpha.$$

This is still an average over the random calibration set and the random test point.

**Training-conditional coverage.** Second, one can condition on the realized calibration set as well. Then the coverage need not be at least $1-\alpha$ for every realized calibration sample. Before that sample is drawn, it is a random quantity determined by the calibration quantile, and the useful question is how this random coverage is distributed across draws of the calibration sample. For i.i.d. observations and continuous score distributions, the answer is a classical Beta law; ties lead to a conservative bound. In this terminology, *training-conditional* means conditional on all observed data $D_1, D_2$, including the calibration set. Recall that the $\mathrm{Beta}(a,b)$ distribution is the distribution on $[0,1]$ with density proportional to $t^{a-1}(1-t)^{b-1}$; its mean is $a/(a+b)$. One further piece of vocabulary is needed: a real random variable $T$ *stochastically dominates* a random variable $T'$ if $\mathbb{P}(T \le t) \le \mathbb{P}(T' \le t)$ for every $t \in \mathbb{R}$, that is, if $T$ is at least as large as $T'$ in distribution.

**Theorem B.1: Training-conditional coverage of split conformal [184]**

Assume $Z_1, \dots, Z_{n+1}$ are i.i.d. and consider the split conformal set from Section 4 of Chapter 10, built from a nonconformity score $S(x,y) = S(x,y;\widehat{f}_{n_1})$ trained on $D_1$, with calibration scores $R_i = S(X_i, Y_i)$ for $i \in D_2$, calibration quantile $\widehat{c} = R_{(k)}$, and

$$k = \lceil (1-\alpha)(n_2+1) \rceil \le n_2.$$

Let

$$G(t) = \mathbb{P}\{S(X,Y;\widehat{f}_{n_1}) \le t \mid D_1\}$$

be the conditional distribution function of the score of a fresh draw, so that

$$T = G(\widehat{c}) = \mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) \mid D_1, D_2\}$$

is the realized coverage of the fitted prediction set. Then, conditional on $D_1$:

(a) $T$ stochastically dominates a $\mathrm{Beta}(k,\, n_2+1-k)$ random variable;
(b) if $G$ is continuous, then $T \mid D_1 \sim \mathrm{Beta}(k,\, n_2+1-k)$ exactly;
(c) for every $\Delta \ge 0$, $\mathbb{P}(T \le 1-\alpha-\Delta \mid D_1) \le \exp(-2n_2\Delta^2)$.

**Proof of Theorem B.1**

Condition on $D_1$ throughout; the score function is then fixed, and the calibration scores $R_i = S(X_i, Y_i)$, $i \in D_2$, are i.i.d. with distribution function $G$, independent of the test pair. Set $U_i = G(R_i)$.

The first observation is that $T$ can be written as an order statistic. The function $G$ is nondecreasing, so applying $G$ preserves the ordering of the scores, and $\widehat{c} = R_{(k)}$ gives

$$T = G(R_{(k)}) = U_{(k)},$$

the $k$th smallest of $U_1, \dots, U_{n_2}$.

*Step 1: the transformed scores are super-uniform.* For every $\tau \in [0,1]$,

$$\mathbb{P}(U_1 \le \tau \mid D_1) \le \tau.$$

The inequality is trivial when $\tau = 1$, so take $\tau < 1$ and let $t_\tau = \sup\{t : G(t) \le \tau\}$; the supremum is finite or $-\infty$ because $G(t) \to 1 > \tau$ as $t \to \infty$, and if the set is empty then $G(R_1) > \tau$ always and the probability is zero. Since $G$ is nondecreasing, $G(s) > \tau$ for every $s > t_\tau$ and $G(t) \le \tau$ for every $t < t_\tau$. If $G(t_\tau) \le \tau$, then $\{G(R_1) \le \tau\} \subseteq \{R_1 \le t_\tau\}$, so $\mathbb{P}(U_1 \le \tau \mid D_1) \le G(t_\tau) \le \tau$. If $G(t_\tau) > \tau$, then $\{G(R_1) \le \tau\} \subseteq \{R_1 < t_\tau\}$, whose

probability is $\lim_{t\uparrow t_\tau} G(t) \le \tau$. When $G$ is continuous, $U_1$ is exactly uniform: for $\tau \in (0,1)$, the intermediate value theorem supplies $t$ with $G(t) = \tau$, whence $\mathbb{P}(U_1 \le \tau \mid D_1) \ge \mathbb{P}(R_1 \le t \mid D_1) = \tau$, and combining with super-uniformity gives $\mathbb{P}(U_1 \le \tau \mid D_1) = \tau$.

*Step 2: a binomial representation.* For $\tau \in [0,1]$, the event $\{U_{(k)} \le \tau\}$ states that at least $k$ of the $n_2$ variables $U_i$ fall at or below $\tau$. The number of such variables is a binomial count, $\mathrm{Bin}(n_2, p_\tau)$ with $p_\tau = \mathbb{P}(U_1 \le \tau \mid D_1)$, so

$$\mathbb{P}(U_{(k)} \le \tau \mid D_1) = \mathbb{P}\{\mathrm{Bin}(n_2, p_\tau) \ge k\}.$$

A binomial count is stochastically increasing in its success probability: realizing the $i$th Bernoulli variable as $\mathbf{1}\{V_i \le p\}$ for i.i.d. uniform $V_1, \dots, V_{n_2}$ couples the counts at two success probabilities so that the count with the larger $p$ is pointwise larger. Since $p_\tau \le \tau$ by Step 1,

$$\mathbb{P}(U_{(k)} \le \tau \mid D_1) \le \mathbb{P}\{\mathrm{Bin}(n_2, \tau) \ge k\}.$$

*Step 3: uniform order statistics are Beta.* Let $U_1^*, \dots, U_{n_2}^*$ be i.i.d. uniform on $(0,1)$. The same counting identity gives, for $t \in [0,1]$,

$$\mathbb{P}\{U_{(k)}^* \le t\} = \mathbb{P}\{\mathrm{Bin}(n_2, t) \ge k\} = \sum_{j=k}^{n_2} \binom{n_2}{j} t^j (1-t)^{n_2-j}.$$

Differentiating term by term and using the identities $j\binom{n_2}{j} = n_2\binom{n_2-1}{j-1}$ and $(n_2-j)\binom{n_2}{j} = n_2\binom{n_2-1}{j}$, the sum telescopes:

$$\frac{d}{dt}\,\mathbb{P}\{U_{(k)}^* \le t\} = n_2 \binom{n_2-1}{k-1} t^{k-1}(1-t)^{n_2-k} = \frac{n_2!}{(k-1)!\,(n_2-k)!}\, t^{k-1}(1-t)^{n_2-k},$$

which is the $\mathrm{Beta}(k, n_2+1-k)$ density. Hence $U_{(k)}^* \sim \mathrm{Beta}(k, n_2+1-k)$, and combining with Step 2,

$$\mathbb{P}(T \le \tau \mid D_1) \le \mathbb{P}\{U_{(k)}^* \le \tau\} \qquad \text{for every } \tau,$$

which is part (a). When $G$ is continuous, Step 1 gives $p_\tau = \tau$, every inequality above becomes an equality, and $T \mid D_1$ has exactly the $\mathrm{Beta}(k, n_2+1-k)$ distribution, which is part (b).

*Step 4: the tail bound.* Fix $\Delta \ge 0$. The bound is trivial when $1-\alpha-\Delta < 0$, so assume $1-\alpha-\Delta \in [0,1]$. By Steps 2 and 3 with $\tau = 1-\alpha-\Delta$,

$$\mathbb{P}(T \le 1-\alpha-\Delta \mid D_1) \le \mathbb{P}\{\mathrm{Bin}(n_2, 1-\alpha-\Delta) \ge k\}.$$

The binomial count has mean $n_2(1-\alpha-\Delta)$, while

$$k \ge (1-\alpha)(n_2+1) \ge n_2(1-\alpha) = n_2(1-\alpha-\Delta) + n_2\Delta,$$

so the event requires the count to exceed its mean by at least $n_2\Delta$. Hoeffding's inequality for a sum of $n_2$ independent variables with values in $[0,1]$ states that $\mathbb{P}\{\mathrm{Bin}(n_2, p) \ge n_2 p + s\} \le \exp(-2s^2/n_2)$ for every $s \ge 0$; the same inequality reappears in the LTT calibration of Section 7. Taking $s = n_2\Delta$ gives $\exp(-2n_2\Delta^2)$, which is part (c). □

When the score distribution $G$ is continuous, the theorem gives the exact distribution of the realized coverage across calibration samples. It depends on the calibration size and the chosen quantile index, rather than on the shape of $G$. The $\mathrm{Beta}(k, n_2+1-k)$ law has mean $k/(n_2+1) \in [1-\alpha,\, 1-\alpha+1/(n_2+1))$ and standard deviation of order $n_2^{-1/2}$. The realized coverage therefore concentrates near $1-\alpha$ as $n_2$ grows, but a particular calibration sample can give coverage below $1-\alpha$. If $G$ has jumps, the theorem instead gives a lower bound on coverage

in distribution; the coverage can remain conservative. In both cases, the marginal guarantee still holds at every calibration size [5, 184].

High-probability coverage. Part (c) can be read as a recipe when the requested confidence is feasible. Suppose the goal is that the realized coverage exceed $1-\alpha$ with probability at least $1-\delta$ over the draw of the calibration set, and assume

$$\sqrt{\frac{\log(1/\delta)}{2n_2}} < \alpha.$$

Run split conformal at a smaller working level $\alpha' \in (0, \alpha)$, so that the quantile index becomes $k' = \lceil (1-\alpha')(n_2+1) \rceil$. If $k' = n_2 + 1$, the prescribed set is the full response space and the desired coverage statement is trivial. Otherwise, applying part (c) at level $\alpha'$ with $\Delta = \alpha - \alpha'$ gives $\mathbb{P}(T \le 1-\alpha \mid D_1) \le \exp\{-2n_2(\alpha-\alpha')^2\}$, which is at most $\delta$ as soon as

$$\alpha - \alpha' \ge \sqrt{\frac{\log(1/\delta)}{2n_2}}.$$

Thus the high-probability guarantee follows by reducing the working miscoverage level by an amount of order $\sqrt{\log(1/\delta)/n_2}$, which vanishes as the calibration set grows [184]. If the displayed feasibility condition fails, this tail bound does not supply a nontrivial positive working level $\alpha'$; one must accept a weaker confidence statement or use the trivial full prediction set.

Scope of the guarantee. Theorem B.1 is a statement about split conformal prediction with i.i.d. data and a score function fixed by the proper training set. For full conformal prediction, where the score of each point depends on the entire augmented sample, there exist symmetric score functions for which training-conditional coverage fails badly, with realized coverage equal to zero with probability close to $\alpha$, even though marginal coverage continues to hold [26]. The training-conditional result above uses the independence of the calibration scores given the proper training data. Exchangeability alone does not give that conclusion.

**Limits of test-conditional coverage.** Another target is coverage conditional on the exact covariate value:

$$\mathbb{P}\{Y_{n+1} \in C_n(x) \mid X_{n+1} = x\} \ge 1-\alpha \qquad \text{for } P_X\text{-almost every } x.$$

Here the observed data remain random. Thus this conditions on different information from the training-conditional statement above. The almost-everywhere qualification avoids assigning meaning to an arbitrary version of a conditional probability at a covariate value of probability zero. This is the guarantee a practitioner usually wants: whatever covariate the next case presents, the reported set should cover its response with probability $1-\alpha$. In continuous feature spaces, no procedure can deliver it distribution-free unless the reported sets are essentially uninformative. Two definitions are needed to make this precise. An *atom* of a probability distribution is a point that carries positive probability, and a distribution is *nonatomic* if it has no atoms; for a real-valued covariate, nonatomic means a continuous distribution function. The *total variation distance* between two probability distributions $P$ and $Q$ on the same space is $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$, the supremum running over all measurable sets $A$; it is the largest change in the probability of any single event when $P$ is replaced by $Q$.

The result applies to an arbitrary set-valued procedure, not only to conformal prediction: $C_n$ may be any measurable rule that maps a training sample of size $n$ and a covariate value to a subset of $\mathcal{Y}$.

**Theorem B.2: Impossibility of test-conditional coverage [13, 115]**

Suppose the procedure $C_n$ satisfies distribution-free test-conditional coverage: for every distribution $P$ on $\mathcal{X} \times \mathcal{Y}$, when $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are i.i.d. draws from $P$,

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) \mid X_{n+1}\} \geq 1 - \alpha \qquad \text{almost surely.}$$

Then for every distribution $P$ whose covariate marginal $P_X$ is nonatomic, every $x \in \mathcal{X}$, and every $y \in \mathcal{Y}$,

$$\mathbb{P}\{y \in C_n(x)\} \geq 1 - \alpha,$$

where the probability is over the training sample drawn i.i.d. from $P$.

The conclusion says that at any fixed covariate value, the set must contain every candidate response with probability at least $1 - \alpha$, no matter how implausible that response is under $P$. A set with this property carries essentially no information about where the response actually lies.

**Proof of Theorem B.2**

Fix $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\epsilon \in (0, 1)$. Define the perturbed distribution

$$P' = (1 - \epsilon)P + \epsilon\, \delta_{(x,y)},$$

where $\delta_{(x,y)}$ is the point mass at $(x, y)$: a draw from $P'$ comes from $P$ with probability $1 - \epsilon$ and equals $(x, y)$ with probability $\epsilon$.

*Step 1: coverage under the perturbed distribution.* The hypothesis applies to every distribution, in particular to $P'$. Under $P'$, the event $\{X_{n+1} = x\}$ has probability at least $\epsilon > 0$, so the almost-sure conditional bound applies at $X_{n+1} = x$:

$$\mathbb{P}_{P'}\{Y_{n+1} \in C_n(X_{n+1}) \mid X_{n+1} = x\} \geq 1 - \alpha,$$

where $\mathbb{P}_{P'}$ indicates that all $n + 1$ observations are i.i.d. from $P'$. Because $P_X$ is nonatomic, $P$ assigns probability zero to $\{X = x\}$, so under $P'$ the event $\{X_{n+1} = x\}$ can occur only through the point-mass component; conditional on $X_{n+1} = x$, therefore, $Y_{n+1} = y$ almost surely. The display becomes

$$\mathbb{P}_{P'}\{y \in C_n(x)\} \geq 1 - \alpha.$$

The event $\{y \in C_n(x)\}$ no longer involves the test point at all; the probability refers only to the training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn i.i.d. from $P'$.

*Step 2: transfer back to $P$.* Two facts about total variation are needed. First, $d_{\mathrm{TV}}(P', P) \leq \epsilon$, because $P'(A) - P(A) = \epsilon\{\delta_{(x,y)}(A) - P(A)\}$ for every event $A$. Second, the laws of i.i.d. samples of size $n$ satisfy

$$d_{\mathrm{TV}}\big((P')^{\otimes n}, P^{\otimes n}\big) \leq n\, d_{\mathrm{TV}}(P', P)\,:$$

write the difference of the two product measures as a telescoping sum of $n$ terms, where the $i$th term changes the $i$th coordinate from $P$ to $P'$ while the first $i - 1$ coordinates follow $P'$ and the last $n - i$ follow $P$; each term changes any event's probability by at most $d_{\mathrm{TV}}(P', P)$. Applying the definition of total variation to the event $\{y \in C_n(x)\}$,

$$\mathbb{P}_P\{y \in C_n(x)\} \geq \mathbb{P}_{P'}\{y \in C_n(x)\} - d_{\mathrm{TV}}\big((P')^{\otimes n}, P^{\otimes n}\big) \geq 1 - \alpha - n\epsilon.$$

*Step 3: remove the perturbation.* The sample size $n$ is fixed. Letting $\epsilon \downarrow 0$ gives $\mathbb{P}_P\{y \in C_n(x)\} \geq 1 - \alpha$. □

**Corollary B.3: Infinite length**

In the setting of Theorem B.2, suppose in addition that $\mathcal{Y} = \mathbb{R}$. Then for every distribution $P$ with nonatomic $P_X$ and every $x \in \mathcal{X}$,

$$\mathbb{P}\{\mathrm{Leb}(C_n(x)) = \infty\} \geq 1 - \alpha,$$

where Leb denotes Lebesgue measure on $\mathbb{R}$. In particular, $\mathbb{E}[\mathrm{Leb}(C_n(x))] = \infty$ whenever $\alpha < 1$.

**Proof of Corollary B.3**

Fix finite constants $a, b > 0$. Deterministically,

$$\mathrm{Leb}(C_n(x)) \geq \int_0^{a+b} \mathbf{1}\{y \in C_n(x)\}\, dy,$$

so on the event $\{\mathrm{Leb}(C_n(x)) \leq a\}$ the set $C_n(x)$ occupies at most $a$ of the interval $[0, a+b]$, and hence

$$\int_0^{a+b} \mathbf{1}\{y \notin C_n(x)\}\, dy \geq b.$$

By Markov's inequality applied to this nonnegative integral, followed by the Fubini–Tonelli theorem to exchange expectation and integration,

$$\mathbb{P}\{\mathrm{Leb}(C_n(x)) \leq a\} \leq \frac{1}{b}\,\mathbb{E}\left[\int_0^{a+b} \mathbf{1}\{y \notin C_n(x)\}\, dy\right] = \frac{1}{b}\int_0^{a+b} \mathbb{P}\{y \notin C_n(x)\}\, dy \leq \frac{(a+b)\alpha}{b},$$

where the last step uses Theorem B.2, which gives $\mathbb{P}\{y \notin C_n(x)\} \leq \alpha$ for every $y$. Send $b \to \infty$ with $a$ fixed: $\mathbb{P}\{\mathrm{Leb}(C_n(x)) \leq a\} \leq \alpha$ for every finite $a$. Now send $a \to \infty$; the events $\{\mathrm{Leb}(C_n(x)) \leq a\}$ increase to $\{\mathrm{Leb}(C_n(x)) < \infty\}$, so $\mathbb{P}\{\mathrm{Leb}(C_n(x)) < \infty\} \leq \alpha$, which is the first claim. When $\alpha < 1$, the length is infinite with positive probability, so the expected length is infinite. □

Sharpness. The bound is attained. The trivial randomized procedure that ignores the data and returns $C_n(x) = \mathcal{Y}$ with probability $1 - \alpha$ and the empty set otherwise satisfies the test-conditional guarantee for every distribution, and it has $\mathbb{P}\{y \in C_n(x)\} = 1 - \alpha$ for every $x$ and $y$. Theorem B.2 is therefore sharp. Any procedure with the stated distribution-free guarantee must include each fixed candidate response with at least this probability. This does not say that its prediction sets have the same distribution as the trivial procedure. When $\mathcal{Y} = \mathbb{R}$, the infinite-length corollary gives the restriction on their size. That restriction holds under every distribution with a nonatomic covariate marginal.

A reusable proof device. The proof combines two moves: perturb the data-generating distribution by a small point-mass mixture, and transfer conclusions between the original and perturbed distributions at total cost $n\epsilon$ through the product total variation bound. Any guarantee required to hold for all distributions must survive the perturbation, and letting $\epsilon \to 0$ makes the probability difference vanish. This argument is useful for distribution-free impossibility results; the same device underlies the hardness of testing conditional independence without assumptions on the joint law, which motivates the model-X viewpoint of Chapter 9.

**Achievable relaxations.** Useful conditional coverage guarantees are possible under weaker requirements. For example, coverage within pre-specified groups of covariates is achievable: partition the feature space into bins fixed before calibration, calibrate a separate quantile within

each bin, and the rank argument applies bin by bin. Group-wise calibration of this kind is known as Mondrian conformal prediction [189]. Coverage conditional on membership in a selected subset of test points is achievable as well. Both guarantees are instances of the selection-conditional theorem of Section 3, since membership in a fixed covariate group is a symmetric, feature-based selection rule. Finally, localized conformal methods can give approximate test-conditional statements under explicit smoothness and localization assumptions; this is different from the density-ratio weighting developed later in this appendix, which gives marginal coverage under the target distribution [13, 115].

## 2. The conformal PRDS argument

**Route: Optional proof**

Recall the property to be proved: for any increasing set $D$, the probability that the p-value vector belongs to $D$, conditional on a true-null p-value equaling $x$, must be nondecreasing in $x$. We will hold the pooled scores fixed and show that assigning a larger rank to that null test point can only increase the other p-values.

### Proof of Theorem 10.6

First take the score function as fixed. A score trained on independent data is handled by conditioning on those data; the final paragraph explains how to remove that conditioning. Consider first $p_1$ and assume $Z_{n+1} \sim P$. Let

$$S_{(1)} \leq \cdots \leq S_{(n+1)}$$

be the order statistics of

$$S(Z_1), \ldots, S(Z_n), S(Z_{n+1}).$$

The unordered multiset of these $n+1$ scores, equivalently the vector of order statistics, is a permutation-symmetric function of the exchangeable vector $(Z_1, \ldots, Z_{n+1})$. Lemma 10.1 therefore applies: conditional on the order statistics, the vector $(Z_1, \ldots, Z_{n+1})$ remains exchangeable, and conditioning further on the scores of the remaining test points $Z_{n+2}, \ldots, Z_{n+m}$ changes nothing, because those points are independent of $(Z_1, \ldots, Z_{n+1})$. Under this conditioning, since the scores are continuous and hence distinct almost surely, the rank of $S(Z_{n+1})$ among the $n+1$ scores is uniform on $\{1, \ldots, n+1\}$. If $p_1 = k/(n+1)$, then $S(Z_{n+1}) = S_{(k)}$, and the reference scores used to compute $p_2, \ldots, p_m$ are the order statistics with $S_{(k)}$ removed. Increasing $k$, equivalently increasing $p_1$, removes a larger reference score. This cannot decrease any of the other conformal p-values, because each $p_j$ counts how many reference scores are less than or equal to $S(Z_{n+j})$.

In symbols, fix another test score $r = S(Z_{n+j})$ and remove $S_{(k)}$ from the ordered list to form the reference sample used for $p_j$. The resulting count is

$$N_j(k) = \#\{\ell : S_{(\ell)} \leq r\} - \mathbf{1}\{S_{(k)} \leq r\}.$$

If $k$ is increased, the indicator $\mathbf{1}\{S_{(k)} \leq r\}$ can only decrease, so $N_j(k)$ can only increase. Hence $p_j(k) = \{1 + N_j(k)\}/(n+1)$ is nondecreasing in $k$, and therefore nondecreasing in $p_1 = k/(n+1)$.

To make the final conditioning step explicit, let $W$ collect the order statistics $(S_{(1)}, \ldots, S_{(n+1)})$ and the remaining test scores. Thus $W$ records all the numerical score information except which ordered position belongs to the first test point. For a supported value $x$, let $G(x, W)$ be the full vector $(p_1, \ldots, p_m)$ reconstructed by assigning the first coordinate the value

$p_1 = x$ and recomputing the remaining p-values from $W$. The preceding argument shows that $G(x, W)$ is coordinatewise nondecreasing in $x$ for every fixed $W$.
Under the null, the rank of $S(Z_{n+1})$, and hence $p_1$, is uniform on its grid and independent of $W$. In particular, conditioning on $p_1 = x$ does not change the distribution of $W$. Therefore, for every increasing set $D$ and every $x \in \{1/(n+1), \dots, 1\}$,

$$\begin{aligned}\mathbb{P}\{(p_1, \dots, p_m) \in D \mid p_1 = x\} &= \mathbb{P}\{G(p_1, W) \in D \mid p_1 = x\} \\ &= \mathbb{E}\big[\mathbf{1}\{G(x, W) \in D\}\big].\end{aligned}$$

The second equality is precisely where independence of $p_1$ and $W$ is used. The integrand is nondecreasing in $x$, so the conditional probability is nondecreasing as well. The same argument applies to any true-null p-value, which is the PRDS property.
For a learned score, denote the independent training data by $D_{\text{fit}}$. The argument gives PRDS conditional on $D_{\text{fit}}$. Each true-null p-value has the same uniform grid law for almost every $D_{\text{fit}}$, so it is independent of $D_{\text{fit}}$. Therefore conditioning on a supported value of that p-value does not change the distribution of $D_{\text{fit}}$. Averaging the conditional PRDS inequalities over this common distribution preserves their monotonicity and proves unconditional PRDS as well. □

## 3. Coverage After Selection

**Route: Advanced**

Marginal coverage can also fail after we select the test points that look most interesting. For example, a scientist may fit a model, choose only compounds whose predicted response exceeds a threshold, and then ask for prediction intervals only for those compounds. The original conformal guarantee averages over all future compounds; it does not automatically guarantee coverage within the selected subset.

Coverage after selection can be recovered when the selection rule treats the calibration and test points symmetrically. For the version studied here, condition on the proper training set $D_1$, so the score function and any fitted predictor are fixed. Let

$$I = I(X_i : i \in D_2;\, X_{n+1}) \subseteq D_2 \cup \{n+1\}$$

be a feature-based selection rule that is invariant to relabeling the calibration points and the test point. For instance, $I$ might select points with $\widehat{f}(X_i) \geq c$. Independent per-point random tie breakers can be included as additional features, as described below. If $n+1 \in I$, let

$$r_I = |I \cap D_2|, \qquad k_I = \lceil (1-\alpha)(r_I + 1) \rceil,$$

and let $\widehat{q}_I$ be the $k_I$th order statistic of the selected calibration scores $\{S(X_i, Y_i) : i \in I \cap D_2\}$, with $\widehat{q}_I = \infty$ if $k_I = r_I + 1$. For a selected test point, report

$$C_I(x) = \{y : S(x, y) \leq \widehat{q}_I\}.$$

**Theorem B.4: Selection-conditional conformal coverage [101]**

Under exchangeability of the calibration and test observations conditional on $D_1$, and for a symmetric selection rule $I$ as above,

$$\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I\} \geq 1 - \alpha,$$

provided the conditioning event has positive probability.

The proof runs the rank argument of Proposition 10.2 conditionally on the selection event. Two preliminary results make this precise: a lemma showing that conditioning on the outcome of a symmetric selection rule preserves exchangeability among the selected points, and a corollary converting that exchangeability into a conformal p-value computed from the selected points only. Throughout the argument, condition on the proper training set $D_1$, so the fitted predictor, the score function $S$, and the selection rule are fixed. To lighten notation, this conditioning is suppressed in the probabilities of the lemma and the corollary; it is restored at the end of the proof of the theorem. The main point is that, after fixing which points were selected, their labels must still be interchangeable. We can then apply the rank bound within that selected set and average over the possible sets.

The setup above described symmetry of the selection rule informally, as invariance to relabeling. Here is the formal condition. Write $\mathcal{N} = D_2 \cup \{n+1\}$ for the calibration and test indices, and view the selection rule as a function of the data vector $Z = (Z_i)_{i\in\mathcal{N}}$ that depends on $Z$ only through the features $(X_i)_{i\in\mathcal{N}}$ and returns a subset $I(Z) \subseteq \mathcal{N}$. For a permutation $\sigma$ of $\mathcal{N}$, let $Z_\sigma = (Z_{\sigma(i)})_{i\in\mathcal{N}}$ denote the relabeled data vector. The rule is *symmetric* if

$$i \in I(Z_\sigma) \quad \Longleftrightarrow \quad \sigma(i) \in I(Z)$$

for every permutation $\sigma$ and every index $i \in \mathcal{N}$; equivalently, $I(Z_\sigma) = \sigma^{-1}\{I(Z)\}$. In words, relabeling the points relabels the selection in the same way, so the order in which the points are listed does not affect which points are selected. The threshold rule $I = \{i : \widehat{f}(X_i) \geq c\}$ is symmetric. Selecting all points at or above the $k$th largest predicted response is also symmetric, but it may select more than $k$ points when predictions tie. To select exactly $k$ points, draw independent $U_i \sim \mathrm{Unif}(0,1)$, independent of all observations, and order the pairs $(\widehat{f}(X_i), U_i)$, using $U_i$ only to break prediction ties. Apply the symmetry condition to these augmented features. The augmented observations remain exchangeable, so the same proof applies. A fixed preference for smaller observation indices would not preserve this symmetry.

**Lemma B.5: Conditional exchangeability after symmetric selection**

Assume the observations $\{Z_i : i \in \mathcal{N}\}$ are exchangeable conditional on $D_1$, and let $I$ be a symmetric feature-based selection rule as above, including the augmented observations when random tie breakers are used. Fix a subset $I_0 \subseteq \mathcal{N}$ with $n+1 \in I_0$ and $\mathbb{P}\{I = I_0\} > 0$. Then the observations $(Z_i)_{i\in I_0}$ are exchangeable conditional on the event $\{I = I_0\}$.

**Proof of Lemma B.5**

Fix a permutation $\sigma$ of $I_0$, and extend it to a permutation $\tilde{\sigma}$ of $\mathcal{N}$ by

$$\tilde{\sigma}(i) = \begin{cases} \sigma(i), & i \in I_0, \\ i, & i \notin I_0. \end{cases}$$

Then $\tilde{\sigma}$ maps $I_0$ onto itself and fixes every index outside $I_0$. Fix a measurable set $A \subseteq (\mathcal{X} \times \mathcal{Y})^{|I_0|}$, with coordinates listed in a fixed enumeration of $I_0$. The chain of equalities is

$$\begin{aligned} \mathbb{P}\{(Z_i)_{i\in I_0} \in A,\ I(Z) = I_0\} &= \mathbb{P}\{(Z_{\tilde{\sigma}(i)})_{i\in I_0} \in A,\ I(Z_{\tilde{\sigma}}) = I_0\} \\ &= \mathbb{P}\{(Z_{\tilde{\sigma}(i)})_{i\in I_0} \in A,\ I(Z) = I_0\} \\ &= \mathbb{P}\{(Z_{\sigma(i)})_{i\in I_0} \in A,\ I(Z) = I_0\}. \end{aligned}$$

The first equality is exchangeability: the two lines apply one and the same measurable functional to the vectors $Z$ and $Z_{\tilde{\sigma}}$, and these two vectors have the same law conditional on $D_1$. The second equality is the symmetry of the selection rule: $I(Z_{\tilde{\sigma}}) = \tilde{\sigma}^{-1}\{I(Z)\}$, so the event $\{I(Z_{\tilde{\sigma}}) = I_0\}$ is the event $\{I(Z) = \tilde{\sigma}(I_0)\}$, and $\tilde{\sigma}(I_0) = I_0$. The third equality holds because $\tilde{\sigma}$ agrees with $\sigma$ on $I_0$. Dividing the two ends of the chain by $\mathbb{P}\{I = I_0\} > 0$ gives

$$\mathbb{P}\{(Z_i)_{i\in I_0} \in A \mid I = I_0\} = \mathbb{P}\{(Z_{\sigma(i)})_{i\in I_0} \in A \mid I = I_0\}.$$

Since $A$ and $\sigma$ are arbitrary, the conditional law of $(Z_i)_{i\in I_0}$ given $\{I = I_0\}$ is invariant under every permutation of $I_0$, which is the claimed exchangeability. □

The lemma does not say that selection has no effect. Conditioning on $\{I = I_0\}$ changes the law of the selected observations; under a threshold rule, every selected point has a large predicted response. It changes the law *symmetrically*, so the selected calibration points remain interchangeable with the selected test point, and that is all the rank argument needs. The proof never uses $n + 1 \in I_0$; the restriction appears in the statement only because it is the case needed below.

The corollary converts this exchangeability into a p-value. Because this section works with nonconformity scores, where larger values are less conforming, the p-value counts the selected calibration scores that are at least as large as the test score; this is the mirror image of the conformity convention used in Section 7 of Chapter 10.

**Corollary B.6: Selective conformal p-value**

In the setting of Lemma B.5, write $S_i = S(X_i, Y_i)$ for $i \in \mathcal{N}$. On the event $\{n + 1 \in I\}$, define

$$p = \frac{1 + \#\{i \in I \cap D_2 : S_i \geq S_{n+1}\}}{1 + |I \cap D_2|}.$$

If $\mathbb{P}\{n + 1 \in I\} > 0$, then

$$\mathbb{P}\{p \leq \alpha \mid n + 1 \in I\} \leq \alpha \qquad \text{for every } \alpha \in [0, 1].$$

**Proof of Corollary B.6**

Fix $I_0 \subseteq \mathcal{N}$ with $n + 1 \in I_0$ and $\mathbb{P}\{I = I_0\} > 0$, and abbreviate

$$r = |I_0 \cap D_2|, \qquad k = \lceil (1 - \alpha)(r + 1) \rceil, \qquad N = \#\{i \in I_0 \cap D_2 : S_i \geq S_{n+1}\}.$$

On the event $\{I = I_0\}$, the p-value is $p = (1 + N)/(1 + r)$. When $\alpha = 1$ the claimed bound is trivial, so assume $\alpha < 1$; then $k \geq 1$.

Two deterministic equivalences come first. The inequality $p > \alpha$ is $N > \alpha(r + 1) - 1$, and since $N$ is an integer this is the same as $N \geq \lfloor \alpha(r + 1) \rfloor$; checking the cases where $\alpha(r + 1)$ is and is not an integer verifies the equivalence. Moreover, since $m - \lceil x \rceil = \lfloor m - x \rfloor$ for every integer $m$ and real $x$,

$$r + 1 - k = r + 1 - \lceil (1 - \alpha)(r + 1) \rceil = \lfloor \alpha(r + 1) \rfloor,$$

so $\{p > \alpha\} = \{N \geq r + 1 - k\}$. Next, let $S_{(1)} \leq \cdots \leq S_{(r)}$ be the ordered selected calibration scores, with the convention $S_{(r+1)} = \infty$. The event $\{N \geq r + 1 - k\}$ is exactly $\{S_{n+1} \leq S_{(k)}\}$. Indeed, if $S_{n+1} \leq S_{(k)}$, then the $r + 1 - k$ scores $S_{(k)}, \ldots, S_{(r)}$ are all at least $S_{n+1}$, so $N \geq r + 1 - k$; if $S_{n+1} > S_{(k)}$, then the $k$ scores $S_{(1)}, \ldots, S_{(k)}$ are all strictly smaller than

$S_{n+1}$, so $N \le r-k$. When $k = r+1$ both events hold trivially. Note that $S_{(k)}$ is exactly the threshold $\widehat{q}_{I_0}$ of the selective prediction set on the event $\{I = I_0\}$, so the two equivalences together identify $\{p > \alpha\}$ with $\{S_{n+1} \le \widehat{q}_{I_0}\}$.

Now the probabilistic step. By Lemma B.5, the observations $(Z_i)_{i \in I_0}$ are exchangeable conditional on $\{I = I_0\}$. Given $D_1$, the score function is a fixed function applied to each observation separately, so the $r+1$ scores $\{S_i : i \in I_0\}$ are also exchangeable conditional on $\{I = I_0\}$. The event $\{S_{n+1} \le S_{(k)}\}$ says that the test score is among the $k$ smallest of these $r+1$ exchangeable scores, so Proposition 10.2 gives

$$\mathbb{P}\{S_{n+1} \le S_{(k)} \mid I = I_0\} \ge \frac{k}{r+1} \ge 1 - \alpha,$$

equivalently $\mathbb{P}\{p \le \alpha \mid I = I_0\} \le \alpha$.

Finally, average over the selection events. The event $\{n+1 \in I\}$ is the disjoint union of the events $\{I = I_0\}$ over the finitely many subsets $I_0 \subseteq \mathcal{N}$ that contain $n+1$ and have positive probability. Summing,

$$\mathbb{P}\{p \le \alpha,\ n+1 \in I\} = \sum_{I_0 \ni n+1} \mathbb{P}\{I = I_0\}\, \mathbb{P}\{p \le \alpha \mid I = I_0\} \le \alpha\, \mathbb{P}\{n+1 \in I\},$$

and dividing by $\mathbb{P}\{n+1 \in I\}$ completes the proof.

□

In words, $p$ remains a valid p-value even after conditioning on the test point having been selected: an analyst who screens points with a symmetric feature-based rule and then tests only the selected ones can use $p$ at the stated level. The bound holds for each $\alpha$, and no continuity of the scores is required; ties only make the p-value larger, hence more conservative. To obtain the prediction-set theorem, we now identify its coverage event with $\{p > \alpha\}$.

**Proof of Theorem B.4**

Work conditionally on $D_1$ and on the event $\{n+1 \in I\}$. Because the selection rule is feature-based, the selected set $I$, the count $r_I$, and the threshold $\widehat{q}_I$ do not depend on the candidate response $y$, so the definition of the prediction set gives

$$\{Y_{n+1} \in C_I(X_{n+1})\} = \{S(X_{n+1}, Y_{n+1}) \le \widehat{q}_I\}.$$

On each selection event $\{I = I_0\}$ with $n+1 \in I_0$, the deterministic equivalences in the proof of Corollary B.6 identify $\{S_{n+1} \le \widehat{q}_{I_0}\}$ with $\{p > \alpha\}$, where $p$ is the selective conformal p-value. Hence, on $\{n+1 \in I\}$, the coverage event is exactly $\{p > \alpha\}$, and Corollary B.6 gives

$$\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I,\, D_1\} = \mathbb{P}\{p > \alpha \mid n+1 \in I,\, D_1\} \ge 1 - \alpha.$$

This is the conditional-on-$D_1$ form of the theorem. For the unconditional form, note the joint bound $\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}),\, n+1 \in I \mid D_1\} \ge (1-\alpha)\, \mathbb{P}\{n+1 \in I \mid D_1\}$, which holds for every realization of $D_1$, trivially so when the conditional selection probability is zero. Taking expectations over $D_1$ and dividing by $\mathbb{P}\{n+1 \in I\} > 0$ gives $\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I\} \ge 1 - \alpha$. □

The lemma-and-corollary route follows the textbook treatment of selective coverage in Angelopoulos et al. [5]; the result and its extensions to more general selection rules are developed by Jin and Ren [101].

This result should not be confused with the $X = x$ conditional coverage that is impossible without assumptions. It is a conditional guarantee for a specified selection rule. If the rule depends on the unknown candidate response $y$, the prediction set must be built by the full conformal inversion logic, checking which candidate values would have made the test point selected. Modern selective conformal methods develop these ideas for more complex selection rules, covariate shift, and false-coverage-statement rates [8, 100, 101].

Group-conditional coverage as a special case. Let $g : \mathcal{X} \to \{1, \dots, G\}$ be a fixed grouping function, chosen before the data are seen or fitted using $D_1$ only, so that $g$ is fixed once we condition on $D_1$. For each group $k$, the rule

$$I = \{i \in D_2 \cup \{n+1\} : g(X_i) = k\}$$

is a symmetric feature-based selection rule, and the event $\{n+1 \in I\}$ is exactly $\{g(X_{n+1}) = k\}$. Theorem B.4 therefore gives, for every group $k$ with positive probability,

$$\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid g(X_{n+1}) = k\} \geq 1 - \alpha.$$

On the conditioning event, the selected calibration points are exactly the calibration points in the same group as the test point, so the procedure reads: calibrate the conformal quantile within the group of the test point, and report the resulting set. Calibrating within the bins of a fixed partition of the feature space is the basic instance. In classification, the same idea is applied with groups defined by the class label; the label of the test point is unknown, so each candidate label $y$ is checked in turn against the calibration scores from class $y$, an instance of the full conformal inversion logic noted above, because the grouping then depends on the response. The general scheme, in which the sample is partitioned by a fixed grouping of the data, the conformal quantile is calibrated within each cell, and coverage holds conditional on the cell of the test point, is known as Mondrian conformal prediction [189]. This makes precise the group-wise relaxation promised in Section 1: coverage conditional on $X_{n+1} = x$ cannot be guaranteed distribution-free in nonatomic feature spaces, but coverage conditional on a fixed finite grouping is available, at the price that only the within-group calibration points are used, so a group with far fewer than $1/\alpha$ calibration points yields an uninformative set. The guarantee is also a distribution-free instance of conditioning on selection: the coverage statement holds given the event that determined which calibration points were used. Chapter 12 studies the related problem of confidence intervals for parameters chosen after selection. Here the target is a future response, and the guarantee concerns one selected test point. It does not by itself give simultaneous coverage of a selected collection or control FCR across that collection.

## 4. The jackknife+ tournament argument

**Route: Optional proof**

The tournament argument. The proof compares observations in pairs. It first bounds how many observations can have unusually large residuals in most of their pairwise comparisons, then uses exchangeability to bound the probability that the test point is one of them. We begin with the counting argument. Imagine $N$ players in a round-robin competition, with one game for each pair of players. A *tournament matrix* is a 0/1 array $A \in \{0,1\}^{N \times N}$ in which $A_{ij} = 1$ records that player $i$ loses its game against player $j$, subject to the constraint

$$A_{ij} + A_{ji} \leq 1 \qquad \text{for all } i, j \in \{1, \dots, N\},$$

so that at most one of the two ordered comparisons between a pair is recorded as a loss. Ties are allowed: a game may produce no loss at all. The row sum $\sum_{j=1}^{N} A_{ij}$ counts the games that player $i$ loses. The following deterministic lemma says that not too many players can lose nearly all of their games [5].

**Lemma B.7: Tournament lemma**

Let $N \geq 1$ and let $A \in \{0,1\}^{N\times N}$ satisfy $A_{ij} + A_{ji} \leq 1$ for all $i, j \in \{1, \ldots, N\}$. Then, for any $t \in [0, 1]$,

$$\sum_{i=1}^{N} \mathbf{1}\Big\{\sum_{j=1}^{N} A_{ij} \geq (1-t)N\Big\} \leq 2tN.$$

**Proof of Lemma B.7**

Let

$$J = \Big\{i \in \{1, \ldots, N\} : \sum_{j=1}^{N} A_{ij} \geq (1-t)N\Big\}$$

be the set of players with at least $(1-t)N$ losses; the goal is to bound $|J|$. Count the losses of the players in $J$ in two ways. By the definition of $J$,

$$|J|(1-t)N \leq \sum_{i\in J}\sum_{j=1}^{N} A_{ij} = \sum_{i\in J}\sum_{j\in J} A_{ij} + \sum_{i\in J}\sum_{j\notin J} A_{ij}.$$

The first double sum runs over ordered pairs inside $J$, so it can be symmetrized and then bounded by the tournament constraint:

$$\sum_{i\in J}\sum_{j\in J} A_{ij} = \frac{1}{2}\sum_{i\in J}\sum_{j\in J}(A_{ij} + A_{ji}) \leq \frac{|J|^2}{2}.$$

The second double sum has $|J|(N - |J|)$ terms, each at most one. Combining the last two displays with the first,

$$|J|(1-t)N \leq \frac{|J|^2}{2} + |J|(N - |J|) = |J|N - \frac{|J|^2}{2}.$$

If $J$ is empty there is nothing to prove. Otherwise, dividing by $|J|$ gives $(1-t)N \leq N - |J|/2$, and rearranging gives $|J| \leq 2tN$. □

The lemma counts the same losses in two ways. Each player in $J$ has at least $(1-t)N$ losses, but at most $N - |J|$ of them can come from games against players outside $J$, and the games inside $J$ supply at most one loss per pair, at most $|J|(|J|-1)/2$ losses in total. Comparing these bounds limits the number of players in $J$.

**Proof of Theorem 10.8**

Algorithmic symmetry means that the fitting rule returns the same fitted function when its input data are presented in a different order, so every fit below depends only on the unordered set of points supplied to it.

*Step 1: a tournament on the augmented data.* Augment the data with the test point, giving $Z_1, \ldots, Z_{n+1}$ with $Z_i = (X_i, Y_i)$. For each pair $i \neq j$ in $\{1, \ldots, n+1\}$, let $\widehat{f}_{-(i,j)}$ be the fit obtained by applying the algorithm to the augmented data with *both* points $i$ and $j$ removed; note that $\widehat{f}_{-(i,j)} = \widehat{f}_{-(j,i)}$, so the two players in each game are judged by a shared fit. Define the residual of point $i$ in its game against point $j$ as

$$R_{i,j} = |Y_i - \widehat{f}_{-(i,j)}(X_i)|, \qquad R_{j,i} = |Y_j - \widehat{f}_{-(i,j)}(X_j)|,$$

and declare that $i$ loses to $j$ when its residual is strictly larger,

$$A_{ij} = \mathbf{1}\{R_{i,j} > R_{j,i}\}, \qquad A_{ii} = 0,$$

so a loss means that the point conforms strictly less well than its opponent under their shared fit. The strict inequalities $R_{i,j} > R_{j,i}$ and $R_{j,i} > R_{i,j}$ cannot hold simultaneously, so $A_{ij} + A_{ji} \leq 1$ and $A$ is a tournament matrix over $N = n+1$ players; when the two residuals tie, neither player loses that game.

The games against the test point recover the jackknife+ ingredients. For $i \leq n$, removing both point $i$ and point $n+1$ from the augmented data leaves exactly the original $n$ observations without point $i$, so $\widehat{f}_{-(i,n+1)} = \widehat{f}_{-i}$ and

$$R_{i,n+1} = |Y_i - \widehat{f}_{-i}(X_i)| = R_i^{\mathrm{LOO}}, \qquad R_{n+1,i} = |Y_{n+1} - \widehat{f}_{-i}(X_{n+1})|.$$

*Step 2: miscoverage forces at least* $(1-\alpha)(n+1)$ *losses.* Suppose first that $Y_{n+1} > q_{\mathrm{up}}$. This forces $q_{\mathrm{up}} < \infty$, so $q_{\mathrm{up}}$ is the $\lceil (1-\alpha)(n+1) \rceil$th smallest of $U_1, \ldots, U_n$, and $Y_{n+1}$ exceeds at least $\lceil (1-\alpha)(n+1) \rceil$ of the values $U_i = \widehat{f}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}$. For each such index $i$,

$$Y_{n+1} - \widehat{f}_{-i}(X_{n+1}) > R_i^{\mathrm{LOO}} \geq 0,$$

hence $R_{n+1,i} = |Y_{n+1} - \widehat{f}_{-i}(X_{n+1})| > R_i^{\mathrm{LOO}} = R_{i,n+1}$, which is exactly $A_{n+1,i} = 1$. Suppose instead that $Y_{n+1} < q_{\mathrm{low}}$. This forces $q_{\mathrm{low}} > -\infty$, so $q_{\mathrm{low}}$ is the $\lfloor \alpha(n+1) \rfloor$th smallest of $L_1, \ldots, L_n$, and $Y_{n+1}$ falls below at least $n - \lfloor \alpha(n+1) \rfloor + 1 \geq (1-\alpha)(n+1)$ of the values $L_i = \widehat{f}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}$. Each such index gives $\widehat{f}_{-i}(X_{n+1}) - Y_{n+1} > R_i^{\mathrm{LOO}} \geq 0$, hence again $A_{n+1,i} = 1$. In both tails the conclusion is the same:

$$Y_{n+1} \notin [q_{\mathrm{low}}, q_{\mathrm{up}}] \implies \sum_{j=1}^{n+1} A_{n+1,j} \geq (1-\alpha)(n+1).$$

No union bound is needed: each tail separately forces player $n+1$ to lose at least $(1-\alpha)(n+1)$ of its games.

*Step 3: every row has the same distribution of losses.* The matrix $A$ is a fixed function of the augmented data. Write $A(z_1, \ldots, z_{n+1})$ for the matrix computed from a generic list of data points, so $A = A(Z_1, \ldots, Z_{n+1})$. By algorithmic symmetry, the fit $\widehat{f}_{-(i,j)}$ depends only on the unordered set of the remaining $n-1$ points, so relabeling the data by a permutation $\sigma$ of $\{1, \ldots, n+1\}$ merely relabels the players:

$$\big(A(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})\big)_{ij} = \big(A(Z_1, \ldots, Z_{n+1})\big)_{\sigma(i)\sigma(j)}.$$

Exchangeability says that $(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})$ has the same joint law as $(Z_1, \ldots, Z_{n+1})$, so the left side has the same distribution as $A$ itself. Combining the two facts, the relabeled matrix $(A_{\sigma(i)\sigma(j)})_{i,j}$ has the same distribution as $(A_{ij})_{i,j}$ for every permutation $\sigma$. Fix $i \in \{1, \ldots, n+1\}$ and choose $\sigma$ with $\sigma(n+1) = i$. The $(n+1)$th row sum of the relabeled matrix is $\sum_j A_{i,\sigma(j)} = \sum_j A_{ij}$, so

$$\mathbb{P}\Big\{\sum_{j=1}^{n+1} A_{n+1,j} \geq (1-\alpha)(n+1)\Big\} = \mathbb{P}\Big\{\sum_{j=1}^{n+1} A_{ij} \geq (1-\alpha)(n+1)\Big\} \qquad \text{for every } i.$$

*Step 4: average and apply the tournament lemma.* By Step 2, the miscoverage probability is at most the probability that row $n+1$ has at least $(1-\alpha)(n+1)$ losses. Averaging

the display of Step 3 over $i = 1, \dots, n+1$ and writing the average of probabilities as the expectation of an average of indicators,

$$\mathbb{P}\{Y_{n+1} \notin [q_{\text{low}}, q_{\text{up}}]\} \le \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\Big\{\sum_{j=1}^{n+1} A_{ij} \ge (1-\alpha)(n+1)\Big\}$$
$$= \mathbb{E}\Big[\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\Big\{\sum_{j=1}^{n+1} A_{ij} \ge (1-\alpha)(n+1)\Big\}\Big].$$

For every realization of the data, Lemma B.7 applied with $N = n+1$ and $t = \alpha$ bounds the indicator sum by $2\alpha(n+1)$. The expectation is therefore at most $2\alpha(n+1)/(n+1) = 2\alpha$, which proves $\mathbb{P}\{Y_{n+1} \in [q_{\text{low}}, q_{\text{up}}]\} \ge 1 - 2\alpha$. □

The factor $2\alpha$ is the price of the tournament bound, not of a union bound over the two tails. The tournament route follows Angelopoulos et al. [5]; the original proof of Barber et al. [12] bounds the same event by a closely related combinatorial argument. Stronger statements are possible under additional stability assumptions. Under the assumptions used here, jackknife+ has a finite-sample coverage guarantee of $1 - 2\alpha$, which the ordinary jackknife does not have in general.

Why the factor of two. It is natural to ask whether the constant $2\alpha$ could be improved by a sharper argument using only the tournament constraint. The generic tournament bound cannot be improved without additional structure. Full conformal prediction ranks all $n+1$ points by a single list of scores computed under one shared fit on the augmented data. Comparisons from a single score list are transitive: if player $i$ loses to player $j$ and player $j$ loses to player $r$, then player $i$ loses to player $r$. In such a tournament, a player loses at least $(1-\alpha)(n+1)$ games only if its score is among the largest $\alpha(n+1)$ scores. Thus at most $\alpha(n+1)$ players can have that many losses; ties can only reduce the number. This is the rank argument of Proposition 10.2 again, and it is why full conformal achieves $1 - \alpha$ with no factor of two.

Jackknife+ compares each pair of points under a *different* fit, $\widehat{f}_{-(i,j)}$. The comparisons need not be consistent with any single ordering, and the tournament can contain cycles: $i$ loses to $j$, $j$ loses to $r$, and $r$ loses to $i$. Cycles let losses concentrate. There is an explicit cyclic construction that nearly attains the bound of Lemma B.7: arrange $2m+1$ players in a circle, where $m = \alpha(n+1) - 1$ is assumed to be an integer satisfying $0 \le m \le n/2$, and let each circle player lose to the $m$ players that follow it in cyclic order and to all $n - 2m$ players outside the circle. Each circle player then loses exactly $m + (n+1) - (2m+1) = (1-\alpha)(n+1)$ games, so the tournament has $2m + 1 = 2\alpha(n+1) - 1$ players with that many losses, nearly twice the transitive count [5]. The $1 - 2\alpha$ guarantee is therefore essentially tight over all tournaments; it is not an artifact of a loose union bound. What the worst case does not say is that jackknife+ typically loses coverage: for stable algorithms its empirical coverage is usually close to $1 - \alpha$.

Averaging p-values. Jackknife+ has a $K$-fold analogue: split the data into $K$ folds, replace each leave-one-out fit by the fit with one fold held out, and build the interval from the held-out residuals; the resulting interval is called CV+, and the prediction set defined directly by the rank comparisons is called cross-conformal prediction [5, 12]. For a hypothesized test response, membership in the cross-conformal set amounts to thresholding the *average* of $K$ split-conformal p-values, one per fold, each valid on its own but all dependent because they share the data. An average $\bar{p}$ of dependent valid p-values is not in general a valid p-value, but $\min\{2\bar{p}, 1\}$ always is [156, 187], and this doubling is exactly where the $1 - 2\alpha$ level arises in the p-value route to the cross-conformal guarantee [5]. By contrast, the average of e-values is always a valid

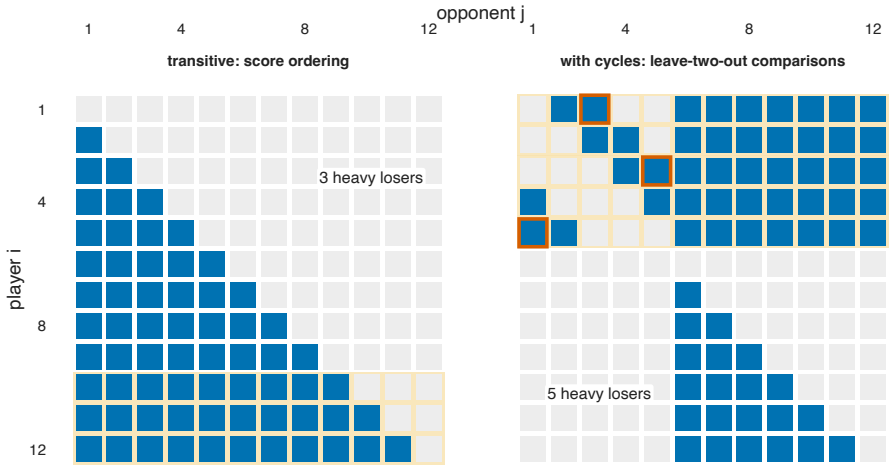



FIGURE B.1. Transitive versus cyclic tournaments. Left: a single score list, as in full conformal prediction, gives transitive comparisons. Players are displayed in score order, best conformity first, so losses fill a staircase, and a player can lose at least $(1-\alpha)(n+1)$ games only by having one of the $\alpha(n+1)$ worst scores; these are the bottom rows, so at most $\alpha(n+1)$ players have that many losses. Right: the pairwise leave-two-out comparisons of jackknife+ can produce cycles, and a cyclic arrangement makes nearly $2\alpha(n+1)$ players lose at least $(1-\alpha)(n+1)$ games each. The tournament lemma shows that this factor of two is the worst case.

e-value (Chapter 8), which is one motivation for the conformal e-prediction methods discussed in Chapter 10.

## 5. Online Conformal P-Values and Testing Exchangeability

**Route: Research frontier**

The basic constructions use a fixed calibration sample. In an online problem, observations arrive one at a time. For example, a hospital may predict a response for each new patient and use earlier patients' outcomes when updating the predictor. Formally, let

$$Z_t = (X_t, Y_t), \qquad t = 1, 2, \ldots,$$

be a sequence of observations. At time $t$, the analyst has observed $Z_1, \ldots, Z_{t-1}$ and the covariate $X_t$, reports a prediction set $C_t(X_t)$ for $Y_t$, then observes $Y_t$ and folds $Z_t$ into the training data before moving on to time $t+1$.

Online full conformal prediction runs the full conformal construction of Chapter 10 at every step:

$$C_t(X_t) = \{y : p_t(y) > \alpha\},$$

where $p_t(y)$ is the full conformal p-value computed from the training data $Z_1, \ldots, Z_{t-1}$ augmented with the candidate pair $(X_t, y)$. At time $t = 1$ there is no training data, and we set $C_1(X_1) = \mathcal{Y}$. For each fixed $t$, exchangeability of $Z_1, \ldots, Z_t$ and the rank argument of Proposition 10.2 give the marginal guarantee $\mathbb{P}\{Y_t \in C_t(X_t)\} \geq 1 - \alpha$. We now consider the joint behavior over time. Every observation is reused as training data at later steps. For online full conformal prediction with a symmetric score and almost surely distinct scores, the conformal p-values at different times

are exactly independent. Fresh independent smoothing can replace the distinct-score assumption, as discussed after the theorem.

To state this precisely, fix a *score algorithm* $S$ that maps a data set $\mathcal{D}$ and a point $z$ to a nonconformity score $S(z;\mathcal{D})$, where smaller values mean better conformity, as in the prediction-set constructions. The algorithm is *symmetric* if the scores it produces do not depend on the order in which the data set is presented: $S(z;\mathcal{D}) = S(z;\mathcal{D}_\sigma)$ for every point $z$ and every permutation $\sigma$ of the entries of $\mathcal{D}$. Symmetry is the same requirement made of the fitting step in full conformal prediction. Write $\mathcal{D}_t = (Z_1,\dots,Z_t)$ for the data observed through time $t$, and compute all $t$ scores at time $t$:

$$R_{t,i} = S(Z_i;\mathcal{D}_t), \qquad i = 1,\dots,t.$$

The *online conformal p-value* at time $t$ is the rank-based p-value of the $t$th score among the $t$ scores available at time $t$:

$$p_t = \frac{1 + \#\{1 \le i \le t-1 : R_{t,i} \ge R_{t,t}\}}{t} = \frac{\#\{1 \le i \le t : R_{t,i} \ge R_{t,t}\}}{t},$$

where the second form holds because the term $i = t$ always contributes one to the count. In particular $p_1 = 1$, and $p_t$ takes values on the grid $\{1/t, 2/t, \dots, 1\}$. The quantity $p_t$ is the full conformal p-value $p_t(y)$ evaluated at the realized response $y = Y_t$, so the miscoverage event at time $t$ is exactly a small p-value:

$$Y_t \notin C_t(X_t) \quad \Longleftrightarrow \quad p_t \le \alpha.$$

**Theorem B.8: Independence of online conformal p-values**

Fix $T \ge 1$. Suppose $Z_1,\dots,Z_T$ are exchangeable, the score algorithm $S$ is symmetric, and for each $t \le T$ the scores $R_{t,1},\dots,R_{t,t}$ are distinct almost surely. Then each $p_t$ is uniformly distributed on $\{1/t, 2/t, \dots, 1\}$, and $p_1,\dots,p_T$ are mutually independent [183, 190].

**Proof of Theorem B.8**

For $t \le T$, let $\widehat{P}_t$ denote the empirical distribution of $Z_1,\dots,Z_t$, that is, the unordered multiset of the first $t$ observations, and let

$$\mathcal{G}_t = \sigma(\widehat{P}_t, Z_{t+1}, \dots, Z_T)$$

be the information consisting of this multiset together with all later observations. This is information used to prove the result, not information available to the analyst at time $t$. We condition backward in time because the later p-values are determined by this information, while the rank of the observation in position $t$ is still random. The proof establishes these two claims and then combines them.

*Claim 1: conditional on $\mathcal{G}_t$, the p-value $p_t$ is uniform on $\{1/t, 2/t, \dots, 1\}$.* Exchangeability implies that permuting $Z_1,\dots,Z_t$ while holding $Z_{t+1},\dots,Z_T$ fixed does not change the joint law, and every such permutation produces the same value of the conditioning information $\mathcal{G}_t$. Hence, conditional on $\mathcal{G}_t$, the vector $(Z_1,\dots,Z_t)$ is a uniformly random ordering of the multiset $\widehat{P}_t$; this is the conditioning device of Lemma 10.1. Because the score algorithm is symmetric, the fitted score function $z \mapsto S(z;\mathcal{D}_t)$ depends on $\mathcal{D}_t$ only through $\widehat{P}_t$. Under the conditioning it is therefore a fixed function, and the multiset of score values $\{R_{t,1},\dots,R_{t,t}\}$ is fixed as well. The only randomness left in $p_t$ is which element of this fixed multiset sits in position $t$. The scores are distinct almost surely, and the ordering is conditionally uniform, so the rank of $R_{t,t}$ among the $t$ scores is conditionally uniform on $\{1,\dots,t\}$. The event

$\{p_t = k/t\}$ is exactly the event that $R_{t,t}$ is the $k$th largest score, so

$$p_t \mid \mathcal{G}_t \ \sim \ \text{uniform on } \{1/t, 2/t, \ldots, 1\}.$$

*Claim 2: for every $s > t$, the p-value $p_s$ is a function of $\mathcal{G}_t$.* Let $\sigma$ be any permutation of $\{1, \ldots, T\}$ that permutes $\{1, \ldots, t\}$ and fixes every index larger than $t$, and relabel the stream by $\sigma$. At time $s > t$, the relabeled data set contains the same points as $\mathcal{D}_s$, only in a different order, and the test point is still $Z_s$ because $\sigma(s) = s$. Symmetry of the score algorithm gives $S(z; (\mathcal{D}_s)_\sigma) = S(z; \mathcal{D}_s)$ for every $z$, and the p-value computed from the relabeled stream is

$$\frac{1}{s}\sum_{i=1}^{s} \mathbf{1}\{S(Z_{\sigma(i)}; \mathcal{D}_s) \geq S(Z_s; \mathcal{D}_s)\} = \frac{1}{s}\sum_{i=1}^{s} \mathbf{1}\{S(Z_i; \mathcal{D}_s) \geq S(Z_s; \mathcal{D}_s)\} = p_s,$$

because the sum runs over the same collection of indices. Thus $p_s$ is invariant to permutations of the first $t$ observations: it depends on the data only through the unordered multiset $\widehat{P}_t$ and $Z_{t+1}, \ldots, Z_T$, which is the claim.

*Combining the claims.* By Claim 2, the vector $(p_{t+1}, \ldots, p_T)$ is a function of $\mathcal{G}_t$. Claim 1 and the tower property of conditional expectation therefore give

$$p_t \mid (p_{t+1}, \ldots, p_T) \ \sim \ \text{uniform on } \{1/t, 2/t, \ldots, 1\},$$

a conditional law that does not depend on the values of the later p-values. Now induct backward from $t = T$. For any grid values $a_t \in \{1/t, \ldots, 1\}$, the chain rule of conditional probability gives

$$\mathbb{P}(p_1 = a_1, \ldots, p_T = a_T) = \prod_{t=1}^{T} \mathbb{P}(p_t = a_t \mid p_{t+1} = a_{t+1}, \ldots, p_T = a_T) = \prod_{t=1}^{T} \frac{1}{t}.$$

The joint probability mass function factorizes into the product of the marginal mass functions, which is mutual independence, and each marginal is uniform on its grid. □

The assumptions matter in three ways. First, the independence is exact, not asymptotic, and it holds even though every pair of p-values is computed from overlapping data. The proof explains why: once the multiset of the first $t$ observations and the later observations $Z_{t+1}, \ldots, Z_T$ are fixed, all later p-values are determined, while the position of the $t$th observation within that multiset is still completely random. Second, the result is about online full conformal retraining with a symmetric score. If the training or calibration data are frozen and all future test points are compared with the same reference sample, as in Section 7 of Chapter 10, the p-values are dependent, and the right tool is the PRDS property of Theorem 10.6. The theorem applies to the symmetric full conformal construction at every step; it does not apply to repeated use of a fixed calibration set. Third, distinct scores matter: with ties the p-values need not be independent, although ties only push conformal prediction toward conservatism, and the long-run coverage below survives as an inequality. Exact continuous uniformity and independence can instead be recovered by smoothing the tied rank at every time $t$ with a fresh $U_t \sim \text{Unif}(0, 1)$, where the $U_t$'s are mutually independent and independent of the data; the backward-rank proof then applies to the smoothed ranks [5, 190].

**Corollary B.9: Independent errors and long-run coverage**

Suppose $Z_1, Z_2, \ldots$ is an infinite sequence whose every finite initial segment satisfies the assumptions of Theorem B.8. Fix $\alpha \in (0,1)$ and let

$$\mathrm{err}_t = \mathbf{1}\{Y_t \notin C_t(X_t)\}$$

be the miscoverage indicator of the online full conformal set at level $\alpha$. Then $\mathrm{err}_1, \mathrm{err}_2, \ldots$ are mutually independent Bernoulli variables with

$$\mathbb{P}(\mathrm{err}_t = 1) = \frac{\lfloor t\alpha \rfloor}{t} \in \Big(\alpha - \frac{1}{t},\ \alpha\Big],$$

and the running coverage frequency converges:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \in C_t(X_t)\} \longrightarrow 1-\alpha \qquad \text{almost surely as } T \to \infty.$$

**Proof of Corollary B.9**

The miscoverage event is exactly $\{p_t \le \alpha\}$, so $\mathrm{err}_t = \mathbf{1}\{p_t \le \alpha\}$ is a function of $p_t$ alone, and functions of mutually independent random variables are mutually independent. Since $p_t$ is uniform on $\{1/t, \ldots, 1\}$, the number of grid points at most $\alpha$ is $\lfloor t\alpha \rfloor$, which gives the stated success probability, and $\alpha - 1/t < \lfloor t\alpha \rfloor / t \le \alpha$. For the running frequency, the summands are independent with variances at most $1/4$, so $\sum_t \mathrm{Var}(\mathrm{err}_t)/t^2 < \infty$, and Kolmogorov's version of the strong law of large numbers for independent, not identically distributed summands gives $T^{-1}\sum_{t \le T}\{\mathrm{err}_t - \mathbb{E}[\mathrm{err}_t]\} \to 0$ almost surely. Since $|\mathbb{E}[\mathrm{err}_t] - \alpha| < 1/t$, the averaged means converge to $\alpha$. Hence the miscoverage frequency converges almost surely to $\alpha$, which is the displayed statement. □

Three dependence regimes. The dependence among conformal p-values depends on how the data are reused. (i) With online full conformal retraining, a symmetric score, and almost surely distinct scores, Theorem B.8 makes the p-values exactly independent under the exchangeable-stream model. BH then has its independent-null guarantee for a fixed finite family of these p-values [20]. This theorem does not establish independence under a stream that mixes observations from different distributions; a multiple-testing interpretation with both true and false nulls requires its own joint-model argument. (ii) With a single shared calibration set, mutually independent test points that are independent of the reference sample, and a continuous true-null score distribution, the p-values are PRDS on the true nulls by Theorem 10.6; BH remains valid by the results of Chapter 6. (iii) Under arbitrary or unknown dependence, one falls back on the BY correction [21], or on conformal e-values combined with the e-BH procedure of Chapter 8 [15, 114, 133]. Figure B.2 summarizes the three regimes.

**An anytime-valid test of exchangeability.** The independence theorem has a second use: it turns the stream of conformal p-values into a sequential test of exchangeability itself. Recall from Chapter 8 that a nonnegative process $M_0, M_1, \ldots$ with $M_0 = 1$ whose conditional expectation given the past cannot increase under the null is a nonnegative test supermartingale, obtained for instance as a running product of conditional e-values (Proposition A.6); such processes are examples of the e-processes of that chapter. Ville's inequality (Theorem A.8) guarantees that, under the null, the entire path exceeds $1/\alpha$ with probability at most $\alpha$, so the process can be monitored at every time without a multiplicity penalty.

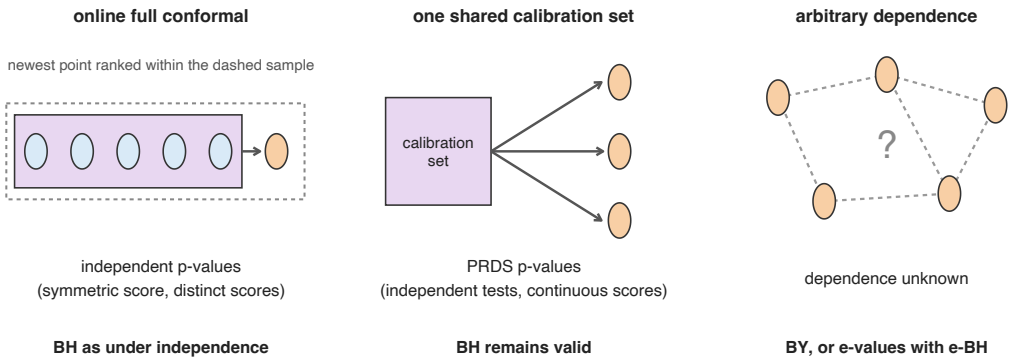



FIGURE B.2. Three dependence regimes for conformal p-values, and the matching multiple-testing tool. Online full conformal retraining with a symmetric score and distinct scores gives exactly independent p-values under an exchangeable stream, so BH applies to a fixed finite family under that model. Fresh independent smoothing gives the analogous conclusion when ties are possible. A single shared calibration set gives PRDS p-values when the test points are mutually independent and independent of the reference sample and the true-null score distribution is continuous, so BH remains valid. Under arbitrary or unknown dependence, the guarantees fall back on the BY correction or on conformal e-values with e-BH.

Here the null hypothesis is exchangeability of the stream: for every $T$, the vector $(Z_1, \dots, Z_T)$ is exchangeable. This is the assumption behind the ordinary conformal rank guarantee. It can fail through drift, changepoints, or feedback from the deployed model itself. The next theorem shows that betting against the online conformal p-values produces a valid e-process for this null.

The bets in the theorem use only earlier conformal p-values. These p-values generate the filtration for the guarantee. The theorem does not justify choosing bets from all past raw observations: $p_t$ need not be independent of that larger information set.

**Theorem B.10: Conformal test supermartingale**

Let $p_1, p_2, \dots$ be the online conformal p-values of this section, computed with a symmetric score algorithm whose scores are almost surely distinct at every step. For each $t$, let $f_t : [0,1] \to [0,\infty)$ be a nonincreasing function with $\int_0^1 f_t(u)\,du \le 1$, where $f_t$ may be chosen as a function of $p_1, \dots, p_{t-1}$. Define

$$M_0 = 1, \qquad M_T = \prod_{t=1}^{T} f_t(p_t).$$

If the stream is exchangeable, then $(M_T)_{T\ge 0}$ is a nonnegative supermartingale with respect to the filtration $\mathcal{F}_T = \sigma(p_1, \dots, p_T)$, and for every $\alpha \in (0,1)$,

$$\mathbb{P}\Big(\sup_{T\ge 0} M_T \ge \frac{1}{\alpha}\Big) \le \alpha.$$

In particular, the sequential test that rejects exchangeability at the first time $M_T \ge 1/\alpha$ has level at most $\alpha$ simultaneously over all stopping rules.

**Proof of Theorem B.10**

Fix $t$ and condition on $\mathcal{F}_{t-1}$. The function $f_t$ is then fixed. By Theorem B.8, applied at every horizon, $p_t$ is independent of $p_1, \dots, p_{t-1}$ and uniform on $\{1/t, 2/t, \dots, 1\}$, so

$$\mathbb{E}[f_t(p_t) \mid \mathcal{F}_{t-1}] = \frac{1}{t} \sum_{k=1}^{t} f_t(k/t).$$

This is where the nonincreasing hypothesis does its work. The p-value is discrete uniform, not continuous uniform, so the conditional expectation is a grid average rather than an integral; monotonicity converts one into the other. Since $f_t$ is nonincreasing, $f_t(k/t) \le f_t(u)$ for every $u \in ((k-1)/t,\, k/t]$, hence

$$\frac{1}{t} f_t(k/t) \le \int_{(k-1)/t}^{k/t} f_t(u)\, du,$$

and summing over $k = 1, \dots, t$ gives

$$\mathbb{E}[f_t(p_t) \mid \mathcal{F}_{t-1}] \le \int_0^1 f_t(u)\, du \le 1.$$

Without monotonicity this step fails: a nonnegative $f_t$ with unit integral can concentrate its mass near the grid points $k/t$ and make the grid average arbitrarily large. Therefore

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}\, \mathbb{E}[f_t(p_t) \mid \mathcal{F}_{t-1}] \le M_{t-1},$$

and $M_T \ge 0$ with $M_0 = 1$, so $(M_T)_{T \ge 0}$ is a nonnegative supermartingale starting at one. Ville's inequality (Theorem A.8 in Appendix A) gives $\mathbb{P}(\sup_T M_T \ge 1/\alpha) \le \alpha$. For any stopping rule $\tau$, the rejection event $\{M_\tau \ge 1/\alpha\}$ is contained in $\{\sup_T M_T \ge 1/\alpha\}$, so the level bound holds uniformly over stopping rules. □

A simple concrete choice is the one-parameter betting family

$$f_t(u) = \frac{1 - \lambda_t u}{1 - \lambda_t/2}, \qquad \lambda_t \in [0, 1],$$

which is nonnegative and nonincreasing on $[0, 1]$ and integrates to exactly one; the parameter $\lambda_t$, which may be tuned from the past p-values, controls how aggressively the test bets on small p-values [190].

Under the theorem's assumptions, each factor $f_t(p_t)$ has conditional expectation at most one given the earlier p-values. Their product can therefore be monitored using Ville's inequality, even though the observations are reused when computing later scores. A value $M_T \ge 1/\alpha$ rejects exchangeability at level $\alpha$. Failure to cross the threshold does not establish exchangeability. Power depends on whether the chosen scores and betting functions respond to the departure from the null.

When exchangeability cannot be assumed, one alternative is to update the conformal threshold online so that the realized coverage frequency tracks the target $1 - \alpha$ no matter how the data are generated, an approach known as adaptive conformal inference [71]. Its guarantee is deterministic and long-run: the average miscoverage over time approaches $\alpha$ for every sequence, random or not. This is a different kind of statement from the distributional coverage studied in this appendix, and we do not develop it here [5].

## 6. When Exchangeability Fails

**Route: Advanced**

Conformal prediction is often described as distribution-free. The phrase means distribution-free under exchangeability. If the calibration data and the future test point are drawn from different distributions, the rank of the test score among the calibration scores need not be uniform.

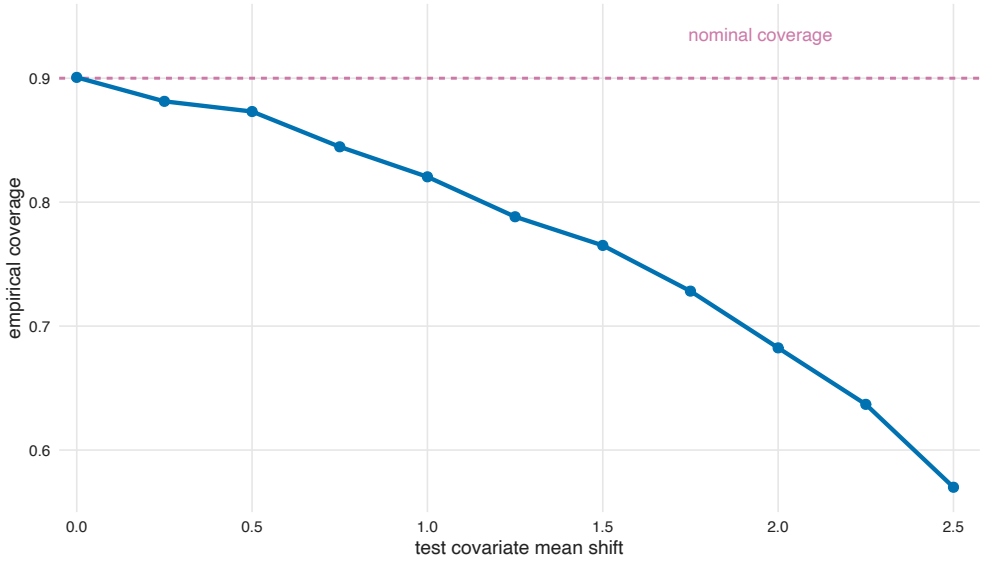



FIGURE B.3. A split conformal interval calibrated for nominal 90% coverage under one covariate distribution can undercover after test covariate shift. The issue is not model misspecification alone; the exchangeable-rank argument no longer applies.

Weighted and distribution-shift conformal methods modify the calibration weights or the coverage target using an assumed relationship between the calibration and test distributions. Their guarantees depend on that relationship being known or estimated accurately.

**Weighted conformal prediction.** When the test distribution differs from the calibration distribution by a known likelihood-ratio function $w(x, y) = \mathrm{d}P_{\text{test}}/\,\mathrm{d}P_{\text{cal}}$, weighted conformal prediction [178] replaces the empirical CDF of calibration scores by a weighted distribution that includes an atom at $+\infty$. The likelihood ratio is the density of the test law with respect to the calibration law; when both laws have densities, it is their ratio, and we use a nonnegative version that is finite at every point. The ratio is allowed to vanish. To match the displays below, the split-conformal calibration set of Section 4 of Chapter 10 is indexed by $1, \dots, n$ in this subsection, the score function $S$ is trained on a separate split, and $S_i = S(X_i, Y_i)$ are the calibration scores. For a candidate response $y$, write

$$d_y = \sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y).$$

When $d_y > 0$, the weighted distribution is

$$\widehat{\mathcal{F}}_y^w = \sum_{i=1}^{n} \frac{w(X_i, Y_i)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)}\, \delta_{S_i} + \frac{w(X_{n+1}, y)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)}\, \delta_{+\infty}.$$

If $d_y = 0$, define $\widehat{\mathcal{F}}_y^w = \delta_{+\infty}$. This makes the cutoff infinite and includes that candidate response. Thus the prediction set is defined even when a candidate and every calibration observation have zero weight. For $d_y > 0$, the finite-score part of this distribution has CDF

$$\widehat{F}_n^w(s) = \frac{\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbf{1}\{S_i \le s\}}{\sum_{i=1}^{n} w(X_i, Y_i) + w(X_{n+1}, y)}.$$

The atom at $+\infty$ accounts for the test point's share of the total weight. We cannot use its unknown true response to compute a calibration score. Placing this mass above every finite score gives a conservative cutoff. In particular, if the test point's normalized weight exceeds $\alpha$, the finite scores carry less than $1 - \alpha$ of the total mass, so the $(1 - \alpha)$-quantile is infinite. The proof below makes this comparison precise by replacing $+\infty$ with the true test score.

For a finite collection of atoms $z_1, \ldots, z_m \in \mathbb{R} \cup \{+\infty\}$ carrying nonnegative weights $w_1, \ldots, w_m$ that sum to one, define the weighted quantile

$$\text{Quantile}\Big(\sum_{j=1}^{m} w_j \delta_{z_j};\, \tau\Big) = \inf\Big\{t \in \mathbb{R} : \sum_{j=1}^{m} w_j\, \mathbf{1}\{z_j \le t\} \ge \tau\Big\},$$

the smallest point at which the cumulative weight reaches $\tau$, with the convention that the infimum of the empty set is $+\infty$. The weighted conformal prediction set is

$$\widehat{C}_n^w(x) = \{y : S(x, y) \le \text{Quantile}(\widehat{\mathcal{F}}_y^w;\, 1 - \alpha)\}.$$

The zero-normalizer convention never changes coverage at the true response. Indeed, the likelihood-ratio identity gives

$$P_{\text{test}}\{w(Z) = 0\} = \int_{\{w=0\}} w(z)\, dP_{\text{cal}}(z) = 0.$$

Hence $w(X_{n+1}, Y_{n+1}) > 0$ almost surely under the test law, and the normalizer at $y = Y_{n+1}$ is positive. With a constant weight function $w \equiv 1$, every atom receives weight $1/(n+1)$ and the set reduces to the split conformal set of Section 4 of Chapter 10: the atom at $+\infty$ reproduces the finite-sample correction $\lceil (1 - \alpha)(n + 1) \rceil$ of Proposition 10.2.

The analysis rests on two ingredients: a deterministic property of weighted quantiles, and a description of where the test point sits inside the pooled sample once its distribution has shifted. The deterministic ingredient, due to Harrison [80], states that a weighted quantile always captures at least its nominal fraction of the weighted atoms.

**Lemma B.11: Deterministic weighted-quantile bound**

Fix real numbers $z_1, \ldots, z_m$ and nonnegative weights $w_1, \ldots, w_m$ with $\sum_{i=1}^{m} w_i = 1$. Then for every $\tau \in [0, 1]$,

$$\sum_{i=1}^{m} w_i\, \mathbf{1}\Big\{z_i \le \text{Quantile}\Big(\sum_{j=1}^{m} w_j \delta_{z_j};\, \tau\Big)\Big\} \ge \tau.$$

**Proof of Lemma B.11**

Write $G(t) = \sum_{j=1}^{m} w_j \mathbf{1}\{z_j \le t\}$; this is a nondecreasing, right-continuous step function. If $\tau = 0$ the claim is trivial, so assume $\tau > 0$. Then $q = \inf\{t : G(t) \ge \tau\}$ is finite, since $G(t) = 1 \ge \tau$ for $t \ge \max_j z_j$ while $G(t) = 0 < \tau$ for $t$ below the smallest atom carrying positive weight. Taking $t_k \downarrow q$ with $G(t_k) \ge \tau$ and using right-continuity gives $G(q) \ge \tau$, which is the claim. □

With equal weights $1/(n_2+1)$ on the $n_2$ calibration scores and the test score used in Section 4 of Chapter 10, Lemma B.11 says that at least $\lceil (1-\alpha)(n_2+1) \rceil$ of the $n_2+1$ scores lie at or below their empirical quantile; this is exactly the deterministic step inside the rank argument of Proposition 10.2.

The probabilistic ingredient replaces the uniform-rank statement of Proposition 10.2. Under a distribution shift the test point is no longer equally likely to occupy each position in the pooled sample, but its conditional position remains explicit: given the observed values, it lands on each of them with probability proportional to the likelihood ratio.

**Proposition B.12: Weighted exchangeability**

Let $Z_1, \dots, Z_n$ be independent draws from a distribution $P$, and let $Z_{n+1} \sim Q$ be drawn independently of them, where $Q$ is absolutely continuous with respect to $P$ with likelihood ratio $w = \mathrm{d}Q/\mathrm{d}P$ satisfying $w(z) < \infty$ for every $z$. Let

$$\widehat{P}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{Z_i}$$

denote the empirical distribution of the pooled sample; conditioning on $\widehat{P}_{n+1}$ means observing the multiset of values $\{Z_1, \dots, Z_{n+1}\}$ without their order. Then

$$Z_{n+1} \mid \widehat{P}_{n+1} \;\sim\; \sum_{i=1}^{n+1} p_i \, \delta_{Z_i}, \qquad p_i = \frac{w(Z_i)}{\sum_{j=1}^{n+1} w(Z_j)}.$$

When $Q = P$, the likelihood ratio is constant, every $p_i$ equals $1/(n+1)$, and the proposition reduces to the exchangeability statement behind the rank argument: the test point is uniform over the pooled values. Under a shift, the test point preferentially occupies the values at which $w$ is large.

**Proof of Proposition B.12**

The claimed conditional law is characterized by the following statement: for all measurable sets $A$ and $B$,

$$\mathbb{P}(Z_{n+1} \in A,\ \widehat{P}_{n+1} \in B) = \mathbb{E}\Big[\sum_{i=1}^{n+1} p_i \, \mathbf{1}\{Z_i \in A\} \, \mathbf{1}\{\widehat{P}_{n+1} \in B\}\Big].$$

Indeed, $\sum_{i=1}^{n+1} p_i \mathbf{1}\{Z_i \in A\}$ is invariant under permutations of $(Z_1, \dots, Z_{n+1})$, hence a function of $\widehat{P}_{n+1}$ alone, and the display is the defining property of a conditional probability given $\widehat{P}_{n+1}$. The second tool is a change-of-measure identity: for any bounded measurable

function $h$,

$$\mathbb{E}_{(Z_1,\dots,Z_{n+1})\sim P^n\times Q}\big[h(Z_1,\dots,Z_{n+1})\big] = \mathbb{E}_{(Z_1,\dots,Z_{n+1})\sim P^{n+1}}\big[w(Z_{n+1})\,h(Z_1,\dots,Z_{n+1})\big],$$

which is the defining property of the likelihood ratio, applied to the last coordinate. Apply the identity with $h = \mathbf{1}\{Z_{n+1}\in A\}\mathbf{1}\{\widehat{P}_{n+1}\in B\}$:

$$\mathbb{P}(Z_{n+1}\in A,\ \widehat{P}_{n+1}\in B) = \mathbb{E}_{P^{n+1}}\big[w(Z_{n+1})\,\mathbf{1}\{Z_{n+1}\in A\}\,\mathbf{1}\{\widehat{P}_{n+1}\in B\}\big].$$

Under $P^{n+1}$, extend the definition of the $p_i$ by setting $p_i = 0$ on the event $\sum_{j=1}^{n+1} w(Z_j) = 0$; this event can have positive probability under $P^{n+1}$ even though it is null under the original law. With this convention the identity

$$w(Z_{n+1}) = p_{n+1}\sum_{j=1}^{n+1} w(Z_j)$$

holds almost surely under $P^{n+1}$, both sides vanishing when the sum does. Substituting and expanding the sum,

$$\mathbb{P}(Z_{n+1}\in A,\ \widehat{P}_{n+1}\in B) = \sum_{j=1}^{n+1}\mathbb{E}_{P^{n+1}}\big[p_{n+1}\,w(Z_j)\,\mathbf{1}\{Z_{n+1}\in A\}\,\mathbf{1}\{\widehat{P}_{n+1}\in B\}\big].$$

Under $P^{n+1}$ the coordinates are independent and identically distributed, so the $j$th expectation is unchanged when the coordinates $j$ and $n+1$ are swapped. The swap sends $p_{n+1}$ to $p_j$, sends $w(Z_j)$ to $w(Z_{n+1})$, sends $\mathbf{1}\{Z_{n+1}\in A\}$ to $\mathbf{1}\{Z_j\in A\}$, and leaves $\widehat{P}_{n+1}$ unchanged. Hence the $j$th term equals

$$\mathbb{E}_{P^{n+1}}\big[w(Z_{n+1})\,p_j\,\mathbf{1}\{Z_j\in A\}\,\mathbf{1}\{\widehat{P}_{n+1}\in B\}\big] = \mathbb{E}_{P^n\times Q}\big[p_j\,\mathbf{1}\{Z_j\in A\}\,\mathbf{1}\{\widehat{P}_{n+1}\in B\}\big],$$

where the equality applies the change-of-measure identity in the reverse direction. Summing over $j$ gives the characterizing display and completes the proof. □

The two ingredients combine into the coverage guarantee. Because the weights in $\widehat{\mathcal{F}}_y^w$ are normalized by their own sum, the construction requires the likelihood ratio only up to a positive multiplicative constant: any unknown normalizing constant of $w$ cancels.

**Theorem B.13: Coverage of weighted conformal prediction**

Let the score function $S$ be trained on data independent of the calibration sample and the test point, as in Section 4 of Chapter 10. Let $Z_1,\dots,Z_n$ be independent draws from $P_{\text{cal}}$ and let $Z_{n+1}\sim P_{\text{test}}$ be independent of them, with likelihood ratio $w = \mathrm{d}P_{\text{test}}/\mathrm{d}P_{\text{cal}}$ finite everywhere and known up to a positive multiplicative constant. Then

$$\mathbb{P}\{Y_{n+1}\in\widehat{C}_n^w(X_{n+1})\}\geq 1-\alpha.$$

**Proof of Theorem B.13**

Condition on the training data, so $S$ is fixed. The normalizer at the true response is positive almost surely by the preceding likelihood-ratio calculation. Write $S_i = S(X_i,Y_i)$ for $i = 1,\dots,n+1$ and $p_i = w(Z_i)/\sum_{j=1}^{n+1} w(Z_j)$. At the candidate value $y = Y_{n+1}$, the distribution $\widehat{\mathcal{F}}_{Y_{n+1}}^w$ places mass $p_i$ at $S_i$ for $i\leq n$ and mass $p_{n+1}$ at $+\infty$. Moving the last atom from $+\infty$ down to $S_{n+1}$ can only decrease the quantile, so the coverage event

$S_{n+1} \le \text{Quantile}(\widehat{\mathcal{F}}^{w}_{Y_{n+1}}; 1-\alpha)$ contains the event

$$E = \Big\{ S_{n+1} \le \text{Quantile}\Big( \sum_{i=1}^{n+1} p_i \delta_{S_i};\, 1-\alpha \Big) \Big\}.$$

Condition further on the multiset $\{Z_1, \dots, Z_{n+1}\}$. The weights $p_i$, the scores $S_i$, and hence the quantile in $E$ are functions of the multiset alone. By Proposition B.12, the test point takes the value $Z_j$ with conditional probability $p_j$, so

$$\mathbb{P}(E \mid \{Z_1, \dots, Z_{n+1}\}) = \sum_{j=1}^{n+1} p_j \, \mathbf{1}\Big\{ S_j \le \text{Quantile}\Big( \sum_{i=1}^{n+1} p_i \delta_{S_i};\, 1-\alpha \Big) \Big\} \ge 1-\alpha,$$

where the inequality is Lemma B.11 with $\tau = 1-\alpha$. The tower rule, averaging first over the multiset and then over the training data, gives the marginal statement. □

Two useful special cases are covariate shift and label shift. In covariate shift, the test and calibration laws share the conditional distribution of $Y$ given $X$, so the likelihood ratio $w(x,y) = w(x)$ depends only on the covariate [178]. In label shift, the two laws share the conditional distribution of $X$ given $Y$, and $w(x,y) = w(y)$ depends only on the response. The same theorem applies, with the unknown test label supplied by the candidate value $y$ inside $\widehat{\mathcal{F}}^{w}_{y}$. In practice the likelihood ratio must usually be estimated, and the estimation error enters the coverage gap; the theorem itself assumes the ratio is known up to its normalizing constant.

Weighted conformal prediction requires the shift to be known through a likelihood ratio. A complementary strategy makes no attempt to model the shift. Fix the weights in advance to encode how much each calibration point is trusted (for example, downweighting older observations in a time series), and accept a coverage penalty that grows with the actual, unknown drift [14]. The penalty is measured in total variation distance, the largest difference between the probabilities that two laws assign to a common event, as recalled in Section 1.

**Theorem B.14: Fixed-weight conformal prediction under drift [14]**

Let $w_1, \dots, w_{n+1}$ be fixed nonnegative constants with $\sum_{i=1}^{n+1} w_i = 1$ and $w_{n+1} \ge w_i$ for every $i \le n$. Let $Z = (Z_1, \dots, Z_{n+1})$ have an arbitrary joint distribution, and let $Z^{(i)}$ denote the sequence with the test point and calibration point $i$ swapped:

$$Z^{(i)} = (Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \dots, Z_n, Z_i).$$

Let $S$ be a fixed measurable score function, or a random score function trained using data independent of $Z$, and set $S_i = S(X_i, Y_i)$. Define the fixed-weight conformal set

$$C^{w}(x) = \Big\{ y : S(x,y) \le \text{Quantile}\Big( \sum_{i=1}^{n} w_i \delta_{S_i} + w_{n+1} \delta_{+\infty};\, 1-\alpha \Big) \Big\}.$$

Then

$$\mathbb{P}\{ Y_{n+1} \in C^{w}(X_{n+1}) \} \ge 1 - \alpha - \sum_{i=1}^{n} w_i \, d_{\mathrm{TV}}(Z,\, Z^{(i)}),$$

where $d_{\mathrm{TV}}(Z, Z^{(i)})$ denotes the total variation distance between the laws of the two sequences.

No exchangeability is assumed. If the data are exchangeable, then $Z$ and $Z^{(i)}$ have the same law, every penalty term vanishes, and the bound is the usual $1-\alpha$. Away from exchangeability,

each calibration point pays a penalty proportional to its own weight: points suspected of being far from the test distribution should receive small weight, and points believed to be close to it may keep large weight.

Proof idea. If $S$ was trained independently, condition on its training data first; the distribution of $Z$ and every total variation term in the theorem are unchanged. Thus the rest of the argument treats $S$ as fixed. This independence condition is essential for the displayed data-level penalty: a merely disjoint training split need not be independent when $Z$ has an arbitrary joint distribution. The set $C^w$ contains the variant in which the atom at $+\infty$ is replaced by an atom at the candidate score $S(x, y)$, so it suffices to bound the coverage of that smaller set; this is the form analyzed by Barber et al. [14], whose result covers any score computed symmetrically from the pooled sample, the split score used here being the simplest case. The argument has three steps. First, a data-processing step. The coverage event depends on the data only through the score vector $(S_1, \ldots, S_{n+1})$, and swapping $Z_i$ with $Z_{n+1}$ induces exactly the corresponding swap of the score vector. The proof therefore establishes the sharper bound in which each penalty term is the total variation distance between the laws of the original and swapped *score vectors*; since total variation distance cannot increase under a measurable transformation, that sharper bound implies the data-level bound displayed in the theorem. In high dimensions the difference can be large: two raw data sequences can be nearly mutually singular while their score vectors have nearly identical laws, so the score-level form of the penalty is the one to use when the displayed bound looks vacuous. Second, a mixture comparison. The coverage probability is expanded as a weighted sum over which index occupies the test position, and each summand is compared with its swapped counterpart at a cost of one total variation term. The mixture comparison used in Barber et al. [14] is where the condition $w_{n+1} \geq w_i$ enters, since the mixture weight $w_{n+1} - w_i$ must be nonnegative. Third, the deterministic weighted-quantile bound of Lemma B.11 guarantees that the weighted fraction of scores at or below the weighted quantile is at least $1 - \alpha$, exactly as in the proof of Theorem B.13.

The bound subtracts an explicit drift penalty from $1 - \alpha$, much as the FDR bounds in Chapter 6 add a penalty for departures from their dependence assumptions. Giving less weight to older observations can reduce the drift penalty, but also reduces their contribution to calibration. The penalty involves the unknown joint law, so the bound quantifies robustness rather than determining an achieved coverage level from the observed data. Localized variants that reweight calibration points by proximity to the test covariate deliver approximate test-conditional coverage by the same weighting mechanism [76, 86].

## 7. Conformal Risk Control and Learn-Then-Test

**Route: Research frontier**

The conformal guarantee controls miscoverage, the expectation of a 0/1 loss. Other applications require control of a different loss, such as the false-negative rate of a medical-image segmentation, the intersection-over-union of a detected object, the token-level F1 score of a generated summary, or the maximum sub-group loss of a fairness-constrained predictor. Angelopoulos et al. [4] introduce CRC, which generalizes coverage control to the expectation of any bounded, monotone loss function.

For both CRC and LTT below, choose the prediction algorithm and the family of loss functions before using the calibration observations. If the prediction algorithm is fitted on a separate sample, condition on that sample so that only calibration remains to be done. The parameter chosen on the calibration sample controls how the fitted predictor is used; it does not authorize unrestricted refitting on the same observations.

The two methods give different guarantees. CRC bounds expected loss, averaging over both the calibration sample and a future observation. The FWER version of LTT uses an additional probability level $\delta$: with probability at least $1-\delta$ over the calibration sample, every certified parameter has population risk at most $\alpha$. The choice between them therefore depends on the desired guarantee as well as on whether the loss is monotone.

**Theorem B.15: Conformal risk control**

Let $\lambda$ range over a fixed compact interval $[\lambda_{\min}, \lambda_{\max}]$, and let $L_\lambda(X, Y)$ be a loss function such that, for every realization of the data, $\lambda \mapsto L_\lambda(X, Y)$ is non-increasing, right-continuous, and $0 \le L_\lambda(X, Y) \le B$. Assume also that $L_{\lambda_{\max}}(X, Y) = 0$ for every realization. Fix a target loss level $\alpha \in [B/(n+1), B]$, and let $\widehat{\lambda}$ be the smallest $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ such that

$$\frac{1}{n+1}\left(\sum_{i=1}^{n} L_\lambda(X_i, Y_i) + B\right) \le \alpha.$$

Then the test-point loss satisfies

$$\mathbb{E}[L_{\widehat{\lambda}}(X_{n+1}, Y_{n+1})] \le \alpha$$

under exchangeability of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$.

**Proof of Theorem B.15**

Write $Z_i = (X_i, Y_i)$, and define the empirical surrogate

$$R_n(\lambda) = \frac{1}{n+1}\left(\sum_{i=1}^{n} L_\lambda(X_i, Y_i) + B\right).$$

The map $\lambda \mapsto R_n(\lambda)$ is non-increasing because $\lambda \mapsto L_\lambda(X_i, Y_i)$ is non-increasing for every $i$. The endpoint condition and $\alpha \ge B/(n+1)$ make the feasible set nonempty. Define $\widehat{\lambda} = \inf\{\lambda \in [\lambda_{\min}, \lambda_{\max}] : R_n(\lambda) \le \alpha\}$; compactness and right-continuity ensure that the infimum is achieved, so $R_n(\widehat{\lambda}) \le \alpha$.

To justify the test-point expectation, use a leave-one-out symmetrization. For each $i \in \{1, \ldots, n+1\}$, let $\widehat{\lambda}^{(-i)}$ be the threshold obtained by applying the same rule to all observations except $Z_i$:

$$\frac{1}{n+1}\left(\sum_{\ell \ne i} L_\lambda(Z_\ell) + B\right) \le \alpha.$$

The endpoint condition makes every one of these leave-one-out feasible sets nonempty. By exchangeability,

$$\mathbb{E}[L_{\widehat{\lambda}}(Z_{n+1})] = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{E}[L_{\widehat{\lambda}^{(-i)}}(Z_i)].$$

It remains to show the deterministic inequality

$$\sum_{i=1}^{n+1} L_{\widehat{\lambda}^{(-i)}}(Z_i) \le (n+1)\alpha.$$

The smallest leave-one-out threshold gives the largest losses, because each loss decreases with the threshold. We use it to bound all terms in the sum at once. Let $i_*$ be an index for

which $\widehat{\lambda}^{(-i_*)}$ is smallest. Since losses are non-increasing in $\lambda$,

$$L_{\widehat{\lambda}^{(-i)}}(Z_i) \le L_{\widehat{\lambda}^{(-i_*)}}(Z_i) \qquad \text{for every } i.$$

The defining inequality for $\widehat{\lambda}^{(-i_*)}$ gives

$$\sum_{i \neq i_*} L_{\widehat{\lambda}^{(-i_*)}}(Z_i) + B \le (n+1)\alpha.$$

Because $L_{\widehat{\lambda}^{(-i_*)}}(Z_{i_*}) \le B$, the full sum at $\widehat{\lambda}^{(-i_*)}$ is at most $(n+1)\alpha$. Combining the last two displays proves the deterministic inequality, and hence $\mathbb{E}[L_{\widehat{\lambda}}(X_{n+1}, Y_{n+1})] \le \alpha$. □

The proof uses exactly the same exchangeability device as the split-conformal rank argument: a symmetric function of the data inherits the exchangeable behavior of the data, which lets us prove a bound that transfers to the test point. When $L_\lambda(X, Y) = \mathbf{1}\{Y \notin C_\lambda(X)\}$ and $C_\lambda$ is the prediction set indexed by quantile level $\lambda$, CRC reduces to ordinary split conformal coverage. For richer losses such as $1 - \mathrm{IoU}(C_\lambda(X), Y)$, $1 - \mathrm{F1}(C_\lambda(X), Y)$, or per-token false-negative rates, the same theorem yields an expected-loss guarantee when the loss is uniformly bounded and decreases with $\lambda$ for every observation. That monotonicity must be checked for the particular prediction family: enlarging a predicted region need not improve intersection-over-union or F1. For nonmonotone losses, the LTT construction below uses a separate test for each candidate value.

**Beyond monotone losses: Learn Then Test.** CRC requires monotonicity of the loss in the calibration parameter. Many operational losses are not monotone: a fairness-adjusted error metric may rise and fall with calibration parameter, and token-F1 in a multi-step generation need not respond monotonically to a single threshold. The LTT framework of Angelopoulos et al. [2] reduces calibration of an arbitrary risk-controlling parameter to a multiple-testing problem.

Fix a finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$, and define the population risk of a fixed candidate by

$$R(\lambda) = \mathbb{E}_{(X,Y)\sim P}[L_\lambda(X, Y)].$$

For each candidate $\lambda$, test the null hypothesis

$$H_\lambda : R(\lambda) > \alpha$$

on an i.i.d. calibration sample of size $n$ using a concentration-based p-value $p_\lambda$. This independence assumption is stronger than the exchangeability used by CRC above; it is what justifies the usual Hoeffding calibration test. For losses bounded in $[0, B]$, Hoeffding's inequality gives the closed-form valid p-value

$$p_\lambda^{\mathrm{Hoeff}} = \exp\left(-\frac{2n}{B^2}(\alpha - \bar{L}_\lambda)_+^2\right), \qquad \bar{L}_\lambda = \frac{1}{n}\sum_{i=1}^{n} L_\lambda(X_i, Y_i),$$

which is super-uniform under $H_\lambda$ (a direct application of Hoeffding to $\bar{L}_\lambda$ bounded around its mean). Sharper Bernstein-type p-values can exploit small variance. If the experimenter needs anytime validity, the calibration evidence must instead be built as an e-process or by another valid sequential construction.

There are two ways to use the grid. If we want a high-probability guarantee that every accepted tuning value is safe, apply Bonferroni, Holm, or another FWER-controlling rule from Chapter 4. If we want to keep many candidate tuning values and control the average fraction of unsafe ones, BH can be used when the grid p-values satisfy the dependence conditions needed

for BH, such as independence or PRDS. Without such a condition, one should use a dependence-robust rule such as BY, or use e-values with e-BH. The theorem below states the simpler FWER version.

**Theorem B.16: LTT validity**

Suppose each $p_\lambda$ is a valid super-uniform p-value for $H_\lambda$ under the calibration sample. Apply Bonferroni at level $\delta$ to the family $\{p_\lambda : \lambda \in \Lambda\}$, so $\widehat{\Lambda}_{\text{Bonf}} = \{\lambda : p_\lambda \le \delta/K\}$. Then

$$\mathbb{P}\Big(\widehat{\Lambda}_{\text{Bonf}} \subseteq \{\lambda \in \Lambda : R(\lambda) \le \alpha\}\Big) \ge 1 - \delta.$$

If $\widehat{\Lambda}_{\text{Bonf}}$ is nonempty, the analyst may return any $\widehat{\lambda}$ in it; otherwise the procedure abstains. The probability that it returns an unsafe candidate is at most $\delta$.

**Proof of Theorem B.16**

The event "$\widehat{\Lambda}_{\text{Bonf}}$ contains a null $\lambda$" is exactly the event "some $p_\lambda \le \delta/K$ for $\lambda \in \Lambda$ with $R(\lambda) > \alpha$". By the union bound,

$$\mathbb{P}\Big(\widehat{\Lambda}_{\text{Bonf}} \text{ contains a null}\Big) \le \sum_{\lambda : H_\lambda} \mathbb{P}(p_\lambda \le \delta/K) \le K \cdot \delta/K = \delta,$$

using super-uniformity of each $p_\lambda$ under $H_\lambda$. Hence with probability at least $1 - \delta$, every $\lambda$ in the rejected set is *not* a null, i.e., $R(\lambda) \le \alpha$. The bound transfers to any selection rule $\widehat{\lambda}$ that returns an element of $\widehat{\Lambda}_{\text{Bonf}}$ whenever that set is nonempty. □

This probability is over the calibration sample and includes the possibility of abstention. It is not a conditional guarantee given that some candidate was accepted. The risk bound $\alpha$ concerns future observations; $\delta$ bounds the chance that calibration certifies an unsafe candidate.

Using BH instead of Bonferroni preserves the rejection structure, but the usual BH guarantee requires the same kind of dependence care discussed in Chapter 6. Under independence or PRDS, BH controls FDR on the calibration grid: the expected fraction of rejected $\lambda$'s that fail to satisfy $R(\lambda) \le \alpha$ is at most $\delta$. For bounded losses calibrated with Hoeffding p-values, the concentration scale at familywise level $\delta$ is $B\sqrt{\log(K/\delta)/n}$, matching the usual uniform-concentration rate.

LTT applies a multiple-testing procedure from earlier chapters (BH, Bonferroni, or e-BH on e-value calibration tests) to a calibration grid. Risk-controlling prediction sets in the sense of Bates et al. [16] are a special case in which the test is constructed from a known concentration inequality.

APPENDIX C

# Details for Decorrelated-Score Inference

**Route: Advanced**

Section 8 of Chapter 11 constructs a score that is less sensitive to estimation error in nuisance parameters. This appendix makes the required error bounds explicit and proves Proposition 11.5. The argument separates the cancellation supplied by decorrelation from the additional conditions needed for a normal approximation. It follows the general approach of Ning and Liu [141, Section 3], using the fitted Hessian throughout the expansion so that its connection to the computational recipe is visible. The final section explains the changes needed for a general loss or a misspecified likelihood.

## 1. From Estimation Errors to Score Errors

**Route: Advanced**

Write the parameter as $\beta = (\theta, \gamma^\top)^\top$, with scalar target $\theta$ and nuisance vector $\gamma$. The true parameter is $\beta^\star = (\theta^\star, \gamma^{\star\top})^\top$. For independent, identically distributed observations $U_1, \dots, U_n$, let $\ell_n(\beta)$ be the average negative log-likelihood and set

$$g_n(\beta) = \nabla \ell_n(\beta), \qquad H_n(\beta) = \nabla^2 \ell_n(\beta).$$

Thus $g_n$ is the vector of first derivatives and $H_n$ is the matrix of second derivatives. Subscripts $\theta$ and $\gamma$ select their corresponding coordinates or blocks. We assume that the loss is twice continuously differentiable in a neighborhood containing the parameters used below, so the Hessian is symmetric.

Let $I = \mathbb{E} H_n(\beta^\star)$, and suppose that $I_{\gamma\gamma}$ is invertible. Recall the population projection and partial information

$$w^\star = I_{\gamma\gamma}^{-1} I_{\gamma\theta}, \qquad J = I_{\theta\theta} - w^{\star\top} I_{\gamma\theta}.$$

The dimension, these population quantities, and the true parameter may depend on $n$. We assume that $J$ remains bounded away from zero and infinity. This prevents the target direction from becoming either uninformative or arbitrarily informative per observation.

Use the initial fit $\widehat{\beta} = (\widehat{\theta}, \widehat{\gamma}^\top)^\top$ and define $\widehat{H} = H_n(\widehat{\beta})$. The estimate $\widehat{w}$ in (21) satisfies

$$\|\widehat{H}_{\gamma\theta} - \widehat{H}_{\gamma\gamma}\widehat{w}\|_\infty \le \lambda_w.$$

Here $\|u\|_1 = \sum_j |u_j|$ and $\|u\|_\infty = \max_j |u_j|$. This inequality is the part of the optimization problem used directly in the proof. Establishing that the problem has a solution and that its solution estimates $w^\star$ well requires conditions on the particular model and design.

To shorten the displays, put

$$\widehat{v} = (1, -\widehat{w}^\top)^\top, \qquad S_n^\star = g_{n,\theta}(\beta^\star) - w^{\star\top} g_{n,\gamma}(\beta^\star).$$

The quantity $S_n^\star$ is the score that would be available if the true nuisance parameter and projection were known. The fitted score and curvature from the main chapter are

$$\widehat{S}_n(\theta) = \widehat{v}^\top g_n(\theta, \widehat{\gamma}), \qquad \widehat{J} = \widehat{H}_{\theta\theta} - \widehat{w}^\top \widehat{H}_{\gamma\theta}.$$

The notation $g_n(\theta, \widehat{\gamma})$ means that the gradient is evaluated at the full vector $(\theta, \widehat{\gamma}^\top)^\top$. Both $\widehat{\gamma}$ and $\widehat{w}$ are held fixed as $\theta$ varies.

We need expansions at two points: the fitted vector $\widehat{\beta}$ and the vector

$$\widehat{\beta}^0 = (\theta^\star, \widehat{\gamma}^\top)^\top,$$

which replaces only the fitted target coordinate by its true value. For testing, we need to evaluate the score at $\theta^\star$, so we use $\widehat{\beta}^0$. For the one-step estimator, we evaluate it at $\widehat{\theta}$, so we use $\widehat{\beta}$. Both evaluations use the same fitted nuisance parameter. For either point $b$, define the projected nonlinear remainder

$$\mathcal{R}_n(b) = \widehat{v}^\top \{g_n(b) - g_n(\beta^\star) - \widehat{H}(b - \beta^\star)\}. \tag{49}$$

It measures the error from replacing the gradient difference by a linear approximation with the fitted Hessian. No independence between the fitted quantities and the observations is assumed.

The lemma below gives sufficient conditions for these fitting errors to be negligible compared with the $n^{-1/2}$ fluctuations of the score at the true parameter. The sequences $r_\gamma$ and $r_w$ describe estimation error, while $a_n$ describes the largest nuisance-score fluctuation.

**Lemma C.1: Sufficient bounds for the score expansions**

Suppose that $\lambda_w$ is deterministic, that the constraint above holds with probability tending to one, and that, for deterministic nonnegative sequences $r_\gamma, r_w, a_n$,

$$\|\widehat{\gamma} - \gamma^\star\|_1 = O_{\mathbb{P}}(r_\gamma), \qquad \|\widehat{w} - w^\star\|_1 = O_{\mathbb{P}}(r_w), \qquad \|g_{n,\gamma}(\beta^\star)\|_\infty = O_{\mathbb{P}}(a_n).$$

Assume further that

$$\sqrt{n}\{\lambda_w r_\gamma + a_n r_w\} \longrightarrow 0, \qquad \max_{b \in \{\widehat{\beta}^0, \widehat{\beta}\}} |\mathcal{R}_n(b)| = o_{\mathbb{P}}(n^{-1/2}). \tag{50}$$

Then

$$\widehat{S}_n(\theta^\star) - S_n^\star = o_{\mathbb{P}}(n^{-1/2}), \tag{51}$$

$$\widehat{S}_n(\widehat{\theta}) - S_n^\star - \widehat{J}(\widehat{\theta} - \theta^\star) = o_{\mathbb{P}}(n^{-1/2}). \tag{52}$$

**Proof of Lemma C.1**

Write $\delta_\gamma = \widehat{\gamma} - \gamma^\star$ and $\delta_w = \widehat{w} - w^\star$. Define the row vector

$$\widehat{d}^\top = \widehat{H}_{\theta\gamma} - \widehat{w}^\top \widehat{H}_{\gamma\gamma}.$$

Symmetry of $\widehat{H}$ and the constraint give $\|\widehat{d}\|_\infty \le \lambda_w$. If $b$ is either of the two points in the lemma, its nuisance coordinate is $\widehat{\gamma}$. Expanding the definition of the remainder gives the exact identity

$$\widehat{S}_n(b_\theta) = S_n^\star + \widehat{J}(b_\theta - \theta^\star) + \widehat{d}^\top \delta_\gamma - \delta_w^\top g_{n,\gamma}(\beta^\star) + \mathcal{R}_n(b). \tag{53}$$

Here $b_\theta$ denotes the target coordinate of $b$. The elementary inequality $|u^\top v| \le \|u\|_\infty \|v\|_1$ bounds the two error products by

$$|\widehat{d}^\top \delta_\gamma| \le \lambda_w \|\delta_\gamma\|_1, \qquad |\delta_w^\top g_{n,\gamma}(\beta^\star)| \le \|\delta_w\|_1 \|g_{n,\gamma}(\beta^\star)\|_\infty.$$

Both are $o_{\mathbb{P}}(n^{-1/2})$ by (50), as is the nonlinear remainder. Taking $b = \widehat{\beta}^0$ gives the first expansion; taking $b = \widehat{\beta}$ gives the second. □

The first product is small because the estimated projection nearly cancels the effect of nuisance-estimation error. The second is small because the projection-estimation error multiplies a score fluctuation centered at zero. The nonlinear remainder is a separate requirement. Population orthogonality alone does not control any of these sample errors.

An example of the rate condition. Before considering the remainder, it helps to substitute familiar rates into the first condition of (50). Suppose a model-specific analysis gives

$$\lambda_w, a_n = O\left(\sqrt{\frac{\log p}{n}}\right), \qquad r_\gamma = O\left(s_\gamma\sqrt{\frac{\log p}{n}}\right), \qquad r_w = O\left(s_w\sqrt{\frac{\log p}{n}}\right),$$

where $s_\gamma$ and $s_w$ are the numbers of nonzero coordinates in $\gamma^\star$ and $w^\star$. Then

$$\sqrt{n}(\lambda_w r_\gamma + a_n r_w) = O\Big(\frac{(s_\gamma + s_w)\log p}{\sqrt{n}}\Big).$$

Thus $(s_\gamma + s_w)\log p = o(\sqrt{n})$ makes these two error products negligible. This is the same rate calculation used for the debiased Lasso in Chapter 11, now accounting for errors in both the nuisance parameter and the projection. It does not yet control $\mathcal{R}_n(b)$.

When is the nonlinear remainder small? For a quadratic loss the Hessian does not depend on $\beta$, so $\mathcal{R}_n(b) = 0$ exactly. For a nonlinear loss, the fundamental theorem of calculus gives

$$\mathcal{R}_n(b) = \int_0^1 \widehat{v}^\top\{H_n(\beta^\star + t(b-\beta^\star)) - \widehat{H}\}(b-\beta^\star)\,dt.$$

One sufficient condition is a Lipschitz bound on the Hessian: throughout a convex neighborhood containing $\beta^\star, \widehat{\beta}^0, \widehat{\beta}$, suppose

$$\|H_n(u) - H_n(v)\|_{\max} \le L_n\|u - v\|_1, \qquad \|A\|_{\max} = \max_{j,k}|A_{jk}|.$$

Using $\|b - \beta^\star\|_1 \le \|\widehat{\beta} - \beta^\star\|_1$ for the two points of interest yields

$$|\mathcal{R}_n(b)| \le \frac{3}{2}L_n(1 + \|\widehat{w}\|_1)\|\widehat{\beta} - \beta^\star\|_1^2.$$

Thus it suffices for the right-hand side, multiplied by $\sqrt{n}$, to converge to zero in probability. This convenient bound may be stronger than necessary: model-specific analysis can use the integral directly. It also explains why the sparsity requirements for a nonlinear model need not match those for least squares.

## 2. Normal Approximation and the One-Step Correction

**Route: Advanced**

The preceding lemma handles the errors introduced by fitting the model. Two further ingredients are needed: a CLT for the score at the true parameter, and consistent estimation of its scale. Under a correctly specified regular likelihood, the information identity equates the covariance matrix of the per-observation negative-log-likelihood gradient with its expected Hessian. Consequently the per-observation projected score has variance $J$. This identity identifies the variance; it does not by itself prove a CLT when the dimension and projection change with $n$.

Assume, as in Proposition 11.5, that

$$\frac{\sqrt{n}\,S_n^\star}{\sqrt{J}} \;\Rightarrow\; N(0,1), \qquad \frac{\widehat{J}}{J} \xrightarrow{\mathbb{P}} 1. \tag{54}$$

The first statement is the oracle CLT; the second is relative consistency of the estimated information, meaning that the ratio of the estimate to the population quantity tends to one.

Since $J$ is bounded away from zero, $\widehat{J} > 0$ with probability tending to one. All the following statistics are evaluated on that event.

**Proof of Proposition 11.5**

Under the null $\theta^\star = \theta_0$, the first score expansion gives

$$\frac{\sqrt{n}\,\widehat{S}_n(\theta_0)}{\sqrt{\widehat{J}}} = \frac{\sqrt{n}\,S_n^\star}{\sqrt{J}}\sqrt{\frac{J}{\widehat{J}}} + o_{\mathbb{P}}(1) \;\Rightarrow\; N(0,1).$$

Slutsky's theorem therefore gives an asymptotic level-$\alpha$ two-sided test by comparing the absolute value of this statistic with $z_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the standard normal distribution.

For the one-step estimator in (23), the second score expansion gives

$$\begin{aligned}\widetilde{\theta} - \theta^\star &= \widehat{\theta} - \theta^\star - \frac{\widehat{S}_n(\widehat{\theta})}{\widehat{J}} \\ &= -\frac{S_n^\star}{\widehat{J}} + o_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

The term involving the initial target-estimation error cancels. Hence

$$\sqrt{n\widehat{J}}\,(\widetilde{\theta} - \theta^\star) = -\frac{\sqrt{n}\,S_n^\star}{\sqrt{J}}\sqrt{\frac{J}{\widehat{J}}} + o_{\mathbb{P}}(1) \;\Rightarrow\; N(0,1).$$

Symmetry of the standard normal law removes the minus sign from the limiting distribution. The interval

$$\widetilde{\theta} \;\pm\; \frac{z_{1-\alpha/2}}{\sqrt{n\widehat{J}}}$$

therefore has coverage tending to $1-\alpha$. □

These are guarantees for a prespecified target coordinate under sequences of models satisfying the stated conditions. Uniform validity over a parameter class requires uniform versions of the probability bounds and normal approximation. Selecting the coordinate after inspecting the data, or reporting many intervals, requires the additional arguments discussed in the main chapter.

## 3. General Losses and Misspecified Models

**Route: Advanced**

A loss function can still define a useful target when it is not the negative log-likelihood of the true data distribution. Suppose

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} L(U_i;\beta), \qquad \beta^\dagger = \arg\min_{\beta} \mathbb{E}L(U_i;\beta).$$

Assume that this population minimizer is uniquely defined, lies in the interior of the parameter space, and satisfies $\mathbb{E}\nabla L(U_i;\beta^\dagger) = 0$. Expectations now refer to the actual data distribution. If $L$ comes from a misspecified likelihood, $\beta^\dagger$ describes the best-fitting member of that model in expected negative log-likelihood. It need not equal a scientifically meaningful coefficient unless that interpretation is justified separately.

Define the population curvature and its target direction by

$$H = \mathbb{E}\nabla^2 L(U_i;\beta^\dagger), \qquad w^\dagger = H_{\gamma\gamma}^{-1}H_{\gamma\theta}, \qquad v^\dagger = (1, -w^{\dagger\top})^\top,$$

assuming that $H_{\gamma\gamma}$ is invertible. Put

$$J_\dagger = H_{\theta\theta} - w^{\dagger\top} H_{\gamma\theta}, \qquad V_\dagger = \mathrm{Var}\{v^{\dagger\top} \nabla L(U_i; \beta^\dagger)\}.$$

Here $J_\dagger$ is the change in the population projected score per unit change in the target; $V_\dagger$ is the variance of one projected score contribution. Curvature describes how a change in the parameter changes the mean score; variance describes how the observed score fluctuates around its mean. The two quantities serve different purposes in the normal approximation. Both are assumed bounded away from zero and infinity. The projection still cancels the nuisance derivative because $H_{\theta\gamma} - w^{\dagger\top} H_{\gamma\gamma} = 0$. Without the information identity, however, it need not make the projected gradient uncorrelated with the nuisance gradient, and $V_\dagger$ need not equal $J_\dagger$. This is the variance distinction emphasized in Ning and Liu [141, Section 5].

The decomposition in (53) remains algebraically valid with the truth replaced by $\beta^\dagger$. Suppose its two expansions hold, the oracle projected score satisfies a CLT with variance $V_\dagger$, and $\widehat{J}/J_\dagger \to 1$ and $\widehat{V}/V_\dagger \to 1$ in probability. The same proof then gives

$$\frac{\sqrt{n}\, \widehat{S}_n(\theta^\dagger)}{\sqrt{\widehat{V}}} \;\Rightarrow\; N(0,1), \qquad \frac{\sqrt{n}\, \widehat{J}(\widetilde{\theta} - \theta^\dagger)}{\sqrt{\widehat{V}}} \;\Rightarrow\; N(0,1).$$

The interval SE is therefore $\sqrt{\widehat{V}}/(\sqrt{n}\,|\widehat{J}|)$. Its squared population counterpart, $V_\dagger/(nJ_\dagger^2)$, is the scalar sandwich variance: the score variance lies between two factors of inverse curvature. For a correctly specified likelihood, the information identity gives $V_\dagger = J_\dagger$, reducing the population variance to $1/(nJ_\dagger)$. Without that identity, estimating curvature alone is not enough to obtain a standard error.

For example, a possible variance estimator uses the fitted contributions

$$\widehat{\psi}_i = \widehat{v}^\top \nabla L(U_i; \widehat{\beta}), \qquad \widehat{V} = \frac{1}{n}\sum_{i=1}^{n} (\widehat{\psi}_i - \overline{\widehat{\psi}})^2, \qquad \overline{\widehat{\psi}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi}_i.$$

Consistency of this estimator must be proved using moment and estimation bounds for the chosen loss; it is not implied by score orthogonality. Thus extending the method requires checking three distinct points: what parameter the loss targets, whether the fitted score has the required expansion, and whether its variance is estimated consistently.

APPENDIX D

# Method-Selection Matrix

Start with the target and output columns, then check whether the required model or dependence condition describes the data. The last column lists alternative methods and diagnostic checks. An alternative may require different assumptions or address a different target or error criterion.

| Method | Target | Output | Guarantee | Status | Dependence/model requirement | Main failure mode | Fallback or diagnostic |
|---|---|---|---|---|---|---|---|
| Bonferroni or Holm | Individual nulls in one family | Rejection set | Strong FWER $\leq \alpha$ | Finite | Marginally valid p-values | Family or weights chosen after seeing results | Prespecified Bonferroni |
| Closed testing | Elementary nulls through intersection nulls | Rejections and simultaneous claims | Strong FWER $\leq \alpha$ | Finite | Valid local tests for every intersection | Invalid local intersection tests or computational shortcuts | Closed Bonferroni |
| BH | Flat discovery family | Rejection set, adjusted p-values | FDR $\leq m_0 q/m$ | Finite | Independent or PRDS super-uniform null p-values | Reused filters/weights or hostile dependence | BY |
| BY | Flat discovery family | Rejection set | FDR $\leq q$ | Finite | Arbitrary dependence, valid marginal p-values | Severe conservatism | Holm if FWER target is acceptable |
| Storey procedure/q-values | Flat discovery family | Adaptive rejection set, estimated q-values | Procedure-specific FDR | Procedure-specific | Finite version: fixed tail cutoff and independence; see Chapter 6 | Unsupported tail tuning or null-fraction estimate | BH or BY with valid p-values |
| Local FDR | Posterior null status | Posterior probabilities/ranking | Oracle posterior interpretation; fitted approximation | Model-based | Correct two-groups model; density estimation adds uncertainty | Misspecified mixture/null density | BH with valid p-values and suitable dependence |
| Group BH | Flat groups | Rejected groups | Group-level FDR | Finite | Valid group p-values and BH dependence condition | Interpreted as leaf-level control | Group-level BY/Bonferroni |
| p-filter | One jointly filtered set at several layers | Layer-consistent rejection set | Simultaneous per-layer FDR | Finite | Simes within each fixed partition; base p-vector PRDS with super-uniform nulls | Interpreted as recursive selected-family control | Closed testing with valid intersection tests (FWER) |
| TreeBH | Tree-structured selected families | Rejected branches/leaves | Recursive selected-family FDR | Finite | Prespecified tree, independent valid leaves, recursive Simes, prescribed path levels | Read as unconditional levelwise FDR | Valid hierarchical FWER method if that target is acceptable |
| Partial conjunction | Truth in at least $r$ studies | Replicability p-value/claim | Error control for $r$-of-$K$ claims | Finite | Valid study p-values and aggregation condition | Read as truth in every study | Bonferroni across studies |
| E-test | One null at a specified analysis | E-value/rejection | Type I error $\leq \alpha$ for $E \geq 1/\alpha$ | Finite | Expectation $\leq 1$ under every null law | Optional continuation of a merely fixed-time e-value | Valid fixed-horizon p-test |

*Continued on next page*

**Method-Selection Matrix (continued)**

| Method | Target | Output | Guarantee | Status | Dependence/model requirement | Main failure mode | Fallback or diagnostic |
|---|---|---|---|---|---|---|---|
| E-process | Sequential null | Anytime evidence; confidence sequence by inversion | Time-uniform Type I error | Finite | Stopped-value validity; e.g., products of conditional e-values | Hindsight tuning or filtration errors | Valid tests at prespecified looks with alpha spending |
| e-BH | Many e-value nulls | Rejection set | FDR $\leq m_0 q/m$ | Finite | Coordinatewise null e-value validity; arbitrary dependence | Invalid input e-values | BY on valid p-values |
| CRT | Conditional feature null $Y \perp X_j \mid X_{-j}$ | Feature p-value | Type I error $\leq \alpha$ | Finite | Correct conditional feature law | Feature-model misspecification | Diagnostic: sensitivity to feature-model choices |
| Fixed-X knockoffs | Coefficient nulls in fixed Gaussian linear model | Selected features | Knockoff+ FDR at target $q$ | Finite | Fixed $X$, Gaussian linear errors, Gram construction, swap-equivariant statistic | Invalid construction or target confusion | Holm on valid prespecified coefficient tests |
| Model-X knockoffs | Conditional feature nulls | Selected features | Knockoff+ FDR at target $q$ | Finite | Joint swap exchangeability, response-independent construction given $X$, flip-sign statistic | Approximate or nonexchangeable knockoffs | CRT still needs a valid feature law; sensitivity is diagnostic |
| Split conformal | Next response | Prediction set | Marginal coverage $\geq 1-\alpha$ | Finite | Exchangeability and fixed score pipeline | Read as conditional coverage; tuning after calibration | Freeze pipeline and recalibrate on fresh data |
| Conformal p-values + BH | Outlier/test-point nulls | Rejection set | FDR $\leq m_0 q/m$ | Finite | Super-uniform conformal p-values and conformal PRDS | Adversarially dependent test points | BY |
| Debiased Lasso | Prespecified high-dimensional coordinates | Estimates and intervals | Asymptotic coordinate/simultaneous coverage | Asymptotic | Sparsity, inverse control, CLT, consistent variance; stronger uniform bounds for many coordinates | Remainder, leverage, variance mismatch | Identifiable lower-dimensional target with valid inference |
| FCR adjustment | Fixed parameters, selected reporting set | Selected intervals | FCR $\leq \alpha$ | Finite | Independent coordinate intervals, or stated concordance/nesting extension | Read as per-selected-interval conditional coverage | Valid simultaneous intervals |
| POSI | All selected-model projection targets | Intervals after arbitrary selection | Simultaneous coverage | Finite, model-based | Full-rank candidate models, Gaussian errors, calibrated simultaneous constant | Wide intervals or structural-target confusion | Independent inference data with a valid target/model |
| Polyhedral selective inference | Contrast after a recorded selection event | Conditional interval/p-value | Selection-conditional coverage/error | Finite Gaussian | Complete linear selection event, fixed tuning | Omitted tuning/analyst choices | Independent inference data, or full carving reference law |
| Prediction-powered inference | Prespecified population parameter | Estimate and interval | Asymptotic coverage | Asymptotic | Matched samples, valid fitting/reuse, CLT, consistent variance | Sample mismatch, overfitting, omitted correction variance | Labeled-only inference; independent fitting or justified cross-fitting |
| Active statistical inference | Prespecified population mean | Weighted estimate and interval | Asymptotic coverage | Asymptotic | Known positive label probabilities, fixed predictor/rule, random labeling, CLT | Ignored weights, tiny probabilities, optional stopping | Uniform labeling; probability floor and weight checks |
| Key resampling | Text–key independence | Audit p-value | Super-uniform randomization p-value | Finite/ideal model | Exchangeable keys, text independence, identical scan | Fixed-window reference for a maximum scan | Full-scan resampling/Bonferroni |
| Reference-relative factuality | Entailment by a reference object | Backed-off answer | Marginal reference-entailment $\geq 1-\alpha$ | Finite | Exchangeable pairs, frozen pipeline, finite grid with terminal abstention | Interpreted as truth or pointwise coverage | Terminal abstention; recalibration after pipeline changes |

APPENDIX E

# Glossary

**Familywise error rate (FWER):** The probability of at least one false rejection in a specified family: $\mathrm{FWER} = \mathbb{P}(V \geq 1)$.

**False discovery rate (FDR):** The expected false discovery proportion: $\mathrm{FDR} = E\{V/(R \vee 1)\}$.

**Positive FDR (pFDR):** The expected false discovery proportion conditional on making at least one rejection: $E(V/R \mid R > 0)$, when $\mathbb{P}(R > 0) > 0$.

**Marginal FDR (mFDR):** The ratio of expected false to expected total discoveries, $E(V)/E(R)$, when $E(R) > 0$. State explicitly if a method uses a modified denominator.

**BH-adjusted p-value:** The smallest nominal level at which an index is rejected by the BH rule. It is a property of that rejection rule, not a posterior null probability.

**Q-value:** A population q-value is the smallest pFDR attainable by a region in a specified nested rejection family that includes the observation. An estimated Storey q-value summarizes entry into a fitted adaptive rule; its FDR validity depends on the procedure's assumptions. See Section 10 and Chapter 6.

**Local FDR:** The posterior probability of a null state given an observation under a specified two-groups model. Fitted local FDR values estimate this model-dependent probability; they do not automatically give frequentist FDR control.

**False coverage rate (FCR):** The expected proportion of noncovering intervals among the intervals selected for reporting.

**Selected-family FDR:** The expectation of an average of within-family FDPs over selected families, with the selection-path weights specified by the method. In TreeBH this is a recursive average along selected branches, distinct from the FDP pooled over all rejections at a depth; see Chapter 7.

**Super-uniformity:** A null p-value property: $\mathbb{P}(P \leq t) \leq t$ for every $t \in [0, 1]$. Exact continuous uniformity is sufficient but not required.

**Negative control:** A prespecified outcome, exposure, sample, or hypothesis expected, for design or scientific reasons, to satisfy the relevant null while sharing the primary analysis pipeline. An unexpected signal can reveal confounding or miscalibration; a null result does not prove validity. A null simulation instead probes an assumed data-generating model, an empirical null is a fitted reference distribution, and knockoffs are engineered competitors justified by swap symmetry.

**PRDS:** Positive regression dependence on a subset: conditioning a true-null p-value to be larger makes every coordinatewise increasing event no less likely, in the precise formulation used by the theorem. Ordinary BH controls FDR under PRDS and super-uniform null p-values.

**Exchangeability:** Invariance of a joint distribution under permutation of the indexed units. It is weaker than independence and supplies the rank symmetry behind conformal and randomization procedures.

**Marginal coverage:** Coverage averaged over the training/calibration data and the next test unit:

$$\mathbb{P}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha.$$

**Conditional coverage:** Coverage after conditioning on specified information, such as a covariate value, training sample, or selection event. The conditioning sigma-field must be stated.

**Simultaneous coverage:** The probability that all intervals or all times cover together is at least $1 - \alpha$. This is stronger than separate marginal coverage.

**E-value:** A nonnegative random variable with expectation at most one under every null law. The rule $E \geq 1/\alpha$ is a valid level-$\alpha$ e-test. The reciprocal conversion $P = \min\{1, 1/E\}$, with $P = 1$ when $E = 0$, gives a valid p-value. Validity at one analysis time does not by itself permit optional continuation.

**E-process:** An adapted nonnegative process whose stopped value has expectation at most one for every stopping time and every null law. It therefore supports time-uniform testing. A test supermartingale is one construction; the general definition and infinite-time convention appear in Appendix A.

**Selective coverage:** Coverage defined after a data-dependent selection step. It may refer to FCR, coverage conditional on a complete selection event, or simultaneous protection; these are distinct guarantees.

**Least-squares projection target:** For fixed mean vector $\mu$ and a selected design $X_M$ with full column rank, the coefficient $(X_M^\top X_M)^{-1} X_M^\top \mu$. It is not automatically a population or structural regression coefficient.

**Prediction-powered inference:** Inference for a population quantity that combines many predictions with observed prediction errors on a smaller labeled sample. The labeled data correct prediction bias; model quality affects precision under the stated sampling and fitting assumptions. See Chapter 13.

**Active statistical inference:** Choosing which outcomes to label in order to estimate a specified population quantity efficiently. Known randomized acquisition probabilities allow correction for unequal labeling, provided relevant cases can receive labels. Acquisition adaptivity and optional stopping require their own guarantees.

**Cross-fitting:** Dividing observations into folds so that each observation is evaluated using a predictor fitted without that fold. This separates fitting from evaluation for each observation; overlapping training sets still require analysis of the combined estimator's dependence and uncertainty.

**Canonicalization:** A deterministic map from equivalent representations to one representation. In watermark auditing it removes representation selection; key exchangeability and identical observed/reference scans still supply resampling validity.

# Bibliography


[1] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

[2] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052, 2021.

[3] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

[4] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the 12th International Conference on Learning Representations*. ICLR, 2024.

[5] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2026. Pre-publication version.

[6] Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *The Annals of Applied Statistics*, 20(3), 2026. doi: 10.1214/26-AOAS2215.

[7] Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011. doi: 10.1214/11-AOS910.

[8] Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Selective conformal inference with false coverage-statement rate control. *Biometrika*, page asae010, 2024.

[9] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

[10] Rina Foygel Barber and Aaditya Ramdas. The p-filter: Multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B*, 79(4): 1247–1268, 2017. doi: 10.1111/rssb.12218.

[11] Rina Foygel Barber, Emmanuel J. Candès, and Richard J. Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3):1409–1431, 2020. doi: 10.1214/19-AOS1852.

[12] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021. doi: 10.1214/20-AOS1965.

[13] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

[14] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

[15] Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with FDR control via conformal e-values. *Advances in Neural Information Processing Systems*, 36:65585–65596, 2023.

[16] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

[17] Stephen Bates, Emmanuel Candès, Lucas Janson, and Wenshuo Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, 116(535):1413–1427, 2021. doi: 10.1080/01621459.2020.1729163.

[18] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244.

[19] Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008. doi: 10.1111/j.1541-0420.2007.00984.x.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[21] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[22] Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469): 71–81, 2005. doi: 10.1198/016214504000001907.

[23] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006. doi: 10.1093/biomet/93.3.491.

[24] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

[25] Robert H. Berk and Douglas H. Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:47–59, 1979. doi: 10.1007/BF00533250.

[26] Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

[27] Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for fdr control. *Electronic Journal of Statistics*, 2:963–992, 2008. doi: 10.1214/08-EJS180.

[28] Gilles Blanchard and Etienne Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

[29] Marina Bogomolov and Ruth Heller. Assessing replicability of findings across two studies of multiple features. *Biometrika*, 105(3):505–516, 2018. doi: 10.1093/biomet/asy029.

[30] Marina Bogomolov, Christine B. Peterson, Yoav Benjamini, and Chiara Sabatti. Hypotheses on a tree: New error rates and testing strategies. *Biometrika*, 108(3):575–590, 2021. doi: 10.1093/biomet/asaa086.

[31] F. Jay Breidt and Jean D. Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205, 2017. doi: 10.1214/16-STS589.

[32] Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009. doi: 10.1002/sim.3495.

[33] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin, 2011. doi: 10.1007/978-3-642-20192-9.

[34] T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017. doi: 10.1214/16-AOS1461.

[35] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B*, 73(5):629–662, 2011. doi: 10.1111/j.1467-9868.2011.00778.x.

[36] T. Tony Cai, Weidong Liu, and Xi Luo. A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

[37] T. Tony Cai, Weidong Liu, and Yin Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B*, 76(2):349–372, 2014. doi: 10.1111/rssb.12034.

[38] Emmanuel J. Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

[39] Hongyuan Cao, Jun Chen, and Xianyang Zhang. Optimal false discovery rate control for large-scale multiple testing with auxiliary information. *The Annals of Statistics*, 50(2): 807–857, 2022. doi: 10.1214/21-AOS2128.

[40] Ivi Chatzi, Nina Corvelo Benz, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Tokenization multiplicity leads to arbitrary price variation in LLM-as-a-service. arXiv preprint arXiv:2506.06446v2, 2025. URL https://arxiv.org/abs/2506.06446v2. First submitted 2025; revised January 2026.

[41] Yiqun T. Chen and Daniela M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

[42] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013. doi: 10.1214/13-AOS1161.

[43] Giovanni Cherubin. Majority vote ensembles of conformal predictors. *Machine Learning*, 108(3):475–488, 2019.

[44] Daniel Csillag, Claudio José Struchiner, and Guilherme Tegoni Goedert. Prediction-powered E-values. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11493–11514. PMLR, 2025. URL https://proceedings.mlr.press/v267/csillag25a.html.

[45] Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplement Issue*, 1:205–237, 1984.

[46] D. A. Darling and P. Erdős. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal*, 23(1):143–155, 1956. doi: 10.1215/S0012-7094-56-02313-4.

[47] Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical Science*, 30(4): 533–558, 2015. doi: 10.1214/15-STS527.

[48] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, 2017. doi: 10.1007/s11749-017-0554-2.

[49] Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L. Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, 2024. doi: 10.1080/01621459.2024.2353948. Forthcoming.

[50] Alex Dmitrienko, Ajit C. Tamhane, and Brian L. Wiens. General multistage gatekeeping procedures. *Biometrical Journal*, 50(5):667–677, 2008. doi: 10.1002/bimj.200710464.

[51] Edgar Dobriban. The Benjamini–Hochberg procedure can fail to control the FDR for correlated two-sided Gaussian tests, 2026. URL https://arxiv.org/abs/2607.12208.

[52] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004. doi: 10.1214/009053604000000265.

[53] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128.

[54] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, 2008.

[55] Robin Dunn, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337, 2023. doi: 10.1093/biomet/asac064.

[56] Cynthia Dwork, Weijie J. Su, and Li Zhang. Differentially private false discovery rate control. *Journal of Privacy and Confidentiality*, 11(2), 2021. doi: 10.29012/jpc.755.

[57] Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004. doi: 10.1198/016214504000000089.

[58] Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Number 1 in Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge, 2012.

[59] European Medicines Agency. Guideline on multiplicity issues in clinical trials. EMA/CHMP/44762/2017, 2017. URL https://www.ema.europa.eu/en/multiplicity-issues-clinical-trials-scientific-guideline.

[60] Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3541. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c9fcd02e6445c7dfbad6986abee53d0d-Abstract.html.

[61] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 4 edition, 1932.

[62] William Fithian and Lihua Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022. doi: 10.1214/21-AOS2137.

[63] William Fithian, Dennis L. Sun, and Jonathan Taylor. Optimal inference after model selection, 2017.

[64] Dean P. Foster and Robert A. Stine. Alpha-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B*, 70(2):429–444, 2008. doi: 10.1111/j.1467-9868.2007.00643.x.

[65] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. doi: 10.1093/biostatistics/kxm045.

[66] Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *Proceedings of the International Conference on Machine Learning*, pages 18495–18527, 2025.

[67] Lucy L. Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024. doi: 10.1080/01621459.2022.2116331.

[68] Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. arXiv preprint arXiv:2401.09379, 2024.

[69] Etienne Gauthier, Francis Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction. arXiv preprint arXiv:2503.13050, 2025. Preprint.

[70] Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1504–1512, 2024.

[71] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[72] Jelle J. Goeman, Rosa J. Meijer, Thijmen J. P. Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106 (4):841–856, 2019. doi: 10.1093/biomet/asz041.

[73] Peter Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024. doi: 10.1073/pnas.2302098121.

[74] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

[75] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

[76] Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

[77] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.

[78] Peter Hall and Jiashun Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732, 2010. doi: 10.1214/09-AOS764.

[79] Ruijian Han, Lan Luo, Yuanhang Luo, Yuanyuan Lin, and Jian Huang. Adaptive debiased SGD in high-dimensional GLMs with steaming data. arXiv preprint arXiv:2405.18284v1, 2024. URL https://arxiv.org/abs/2405.18284v1. Title as posted in the May 2024 version.

[80] Matthew T Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

[81] Ruth Heller and Marina Bogomolov. Replicability across multiple studies. *Statistical Science*, 38(4):602–620, 2023. doi: 10.1214/23-STS890.

[82] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. doi: 10.1093/biomet/75.4.800.

[83] Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2): 901–928, 2023.

[84] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[85] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 1988. doi: 10.1093/biomet/75.2.383.

[86] Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2024.

[87] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

[88] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2): 1055–1080, 2021. doi: 10.1214/20-AOS1991.

[89] Nikolaos Ignatiadis and Wolfgang Huber. Covariate-powered cross-weighted multiple testing. *Journal of the Royal Statistical Society: Series B*, 83(4):720–751, 2021. doi: 10.1111/rssb.12411.

[90] Nikolaos Ignatiadis, Bernd Klaus, Judith B. Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, 2016. doi: 10.1038/nmeth.3885.

[91] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. E-values as unnormalized weights in multiple testing. *Biometrika*, 111(2):417–439, 2024. doi: 10.1093/biomet/asad057.

[92] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. Asymptotic and compound e-values: Multiple testing and empirical bayes. arXiv preprint arXiv:2409.19812, 2024.

[93] Yu. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[94] International Council for Harmonisation. ICH harmonised tripartite guideline E9: Statistical principles for clinical trials. ICH guideline, 1998. URL https://www.ich.org/page/efficacy-guidelines.

[95] International Council for Harmonisation. ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials. ICH guideline addendum, 2020. URL https://www.ich.org/page/efficacy-guidelines.

[96] Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053, 2007. doi: 10.1214/0009053607000000244.

[97] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[98] Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46(2):526–554, 2018. doi: 10.1214/17-AOS1559.

[99] Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *Biometrika*, page asag053, 2026. doi: 10.1093/biomet/asag053. Advance article.

[100] Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

[101] Ying Jin and Zhimei Ren. Confidence on the focal: Conformal prediction with selection-conditional coverage. *Journal of the Royal Statistical Society: Series B*, 87(4):1239–1259, 2025.

[102] Samuel Karlin and Yosef Rinott. Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980. doi: 10.1016/0047-259X(80)90065-2.

[103] Eugene Katsevich and Chiara Sabatti. Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1–33, 2019. doi: 10.1214/18-AOAS1185.

[104] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 2023.

[105] Dan M. Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. arXiv preprint arXiv:2501.18577, 2025.

[106] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000. doi: 10.1214/aos/1015957397.

[107] Nick W. Koning. Post-hoc $\alpha$ hypothesis testing and the post-hoc $p$-value. arXiv preprint arXiv:2312.08040, 2024.

[108] Chenchen Kuai, Jiwan Jiang, Zihao Zhu, Hao Wang, Keshu Wu, Zihao Li, Yunlong Zhang, Chenxi Liu, Zhengzhong Tu, Zhiwen Fan, and Yang Zhou. How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles. arXiv preprint arXiv:2604.07650v1, 2026. URL https://arxiv.org/abs/2604.07650v1. April 2026 version.

[109] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=FpaCL1MO2C.

[110] Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. doi: 10.1214/24-AOS2487.

[111] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

[112] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2 edition, 2000.

[113] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

[114] Junu Lee and Zhimei Ren. Boosting e-BH via conditional calibration. *arXiv preprint arXiv:2404.17562*, 2024.

[115] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014. doi: 10.1111/rssb.12021.

[116] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[117] Lihua Lei and William Fithian. AdaPT: An interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B*, 80(4):649–679, 2018. doi: 10.1111/rssb.12274.

[118] Lihua Lei, Aaditya Ramdas, and William Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267, 2021. doi: 10.1093/biomet/asaa064.

[119] James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, 120(549):135–146, 2025. doi: 10.1080/01621459.2023.2270148.

[120] Ang Li and Rina Foygel Barber. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B*, 81(1):45–74, 2019. doi: 10.1111/rssb.12298.

[121] Guanxun Li and Xianyang Zhang. A general framework for multiple testing via e-value aggregation and data-dependent weighting. arXiv preprint arXiv:2312.02905, 2023.

[122] Guanxun Li and Xianyang Zhang. A note on e-values and multiple testing. *Biometrika*, 112(1):asae050, 2025. doi: 10.1093/biomet/asae050.

[123] Jonathan Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

[124] Puheng Li, Tijana Zrnic, and Emmanuel J. Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2312. URL https://papers.nips.cc/paper_files/paper/2025/hash/638947056421498360 4d1ac81631c2c5-Abstract-Conference.html.

[125] Xingchi Li, Guanxun Li, and Xianyang Zhang. Segmenting watermarked texts from language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

doi: 10.52202/079017-0467. URL https://papers.nips.cc/paper_files/paper/2024/hash/1a8d295871250443f9747d239925b89d-Abstract-Conference.html.

[126] Xingchi Li, Guanxun Li, and Xianyang Zhang. A likelihood based approach for watermark detection. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1675–1683. PMLR, 2025. URL https://proceedings.mlr.press/v258/li25d.html.

[127] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

[128] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(1):145–156, 2022. doi: 10.1093/biomet/asab017.

[129] Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[130] Joshua R. Loftus and Jonathan E. Taylor. Selective inference in regression models with groups of variables. arXiv preprint arXiv:1511.01478, 2015. Preprint; foundational for CV-selected models.

[131] Xiaokun Luan, Yihao Zhang, Pengcheng Su, Feiran Lei, and Meng Sun. VOW: Verifiable and oblivious watermark detection for large language models. arXiv preprint arXiv:2604.27666, 2026. URL https://arxiv.org/abs/2604.27666. Preprint.

[132] Pranav Mani, Peng Xu, Zachary Chase Lipton, and Michael Oberst. No free lunch: Non-asymptotic analysis of prediction-powered inference. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 86296–86335. PMLR, 2026. URL https://proceedings.mlr.press/v306/mani26b.html.

[133] Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

[134] Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. doi: 10.1093/biomet/63.3.655.

[135] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, Chichester, 2000. doi: 10.1002/9780470316979.

[136] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281.

[137] Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *Journal of Machine Learning Research*, 26 (179):1–31, 2025. URL https://www.jmlr.org/papers/v26/24-0056.html.

[138] Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*,

volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR, 2024.

[139] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, 2023.

[140] Anna Neufeld, Ameer Dharamshi, Lucy L. Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

[141] Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017. doi: 10.1214/16-AOS1448.

[142] Yuki Nishino, Tomohiro Shiraishi, Teruyuki Katsuoka, and Ichiro Takeuchi. Statistical test for saliency maps of graph neural networks via selective inference. arXiv preprint arXiv:2501.18452, 2025. Preprint.

[143] Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models, 2023.

[144] Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

[145] Mehrdad Pournaderi and Yu Xiang. Differentially private model-X knockoffs via johnson–lindenstrauss transform. arXiv preprint arXiv:2508.04800, 2025. Preprint.

[146] Aaditya Ramdas and Tudor Manole. Randomized and exchangeable improvements of markov's, chebyshev's and chernoff's inequalities. *Statistical Science*, 41(1):121–142, 2026. doi: 10.1214/24-STS952.

[147] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1-2):1–390, 2025. doi: 10.1561/3600000002.

[148] Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4286–4294. PMLR, 2018.

[149] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv preprint arXiv:2009.03167, 2020.

[150] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022. doi: 10.1016/j.ijar.2021.06.017.

[151] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

[152] Zhimei Ren and Rina Foygel Barber. Derandomised knockoffs: Leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society: Series B*, 86(1):122–154, 2024. doi: 10.1093/jrsssb/qkad085.

[153] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

[154] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[155] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020. doi: 10.1080/01621459.2019.1660174.

[156] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

[157] Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking, 2025.

[158] Sanat K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1):239–257, 2002. doi: 10.1214/aos/1015362192.

[159] Matteo Sesia, Chiara Sabatti, and Emmanuel J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2019. doi: 10.1093/biomet/asy033.

[160] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647.

[161] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[162] Tomohiro Shiraishi, Daiki Miwa, Vo Nguyen Le Duy, and Ichiro Takeuchi. Selective inference for change point detection by recurrent neural network. *Neural Computation*, 37 (6):1213–1252, 2025. doi: 10.1162/neco_a_01756.

[163] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967. doi: 10.1080/01621459.1967.10482935.

[164] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986. doi: 10.1093/biomet/73.3.751.

[165] Erik Skalnes and Michael Oberst. Fixed-size active statistical inference. arXiv preprint arXiv:2610.27389, 2026. Accepted at NeurIPS 2026.

[166] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI–EBI GWAS catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1): D977–D985, 2023. doi: 10.1093/nar/gkac1010.

[167] Branko Sorić. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610, 1989. doi: 10.1080/01621459.1989.10478811.

[168] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64(3):479–498, 2002.

[169] John D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. doi: 10.1214/aos/1074290335.

[170] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B*, 66(1):187–205, 2004. doi: 10.1111/j.1467-9868.2004.00439.x.

[171] Samuel A. Stouffer, Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams. *The American Soldier: Adjustment During Army Life*, volume 1 of *Studies in Social Psychology in World War II*. Princeton University Press, Princeton, NJ, 1949.

[172] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. doi: 10.1093/biomet/ass043.

[173] Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

[174] Wenguang Sun, Brian J. Reich, T. Tony Cai, Michele Guindani, and Armin Schwartzman. False discovery control in large-scale multiple testing. *Journal of the Royal Statistical Society: Series B*, 77(1):59–83, 2015.

[175] Jonathan Taylor and Robert Tibshirani. Post-selection inference for l1-penalized likelihood models. *The Canadian Journal of Statistics*, 46(1):41–61, 2018. doi: 10.1002/cjs.11313.

[176] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[177] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. doi: 10.1080/01621459.2015.1108848.

[178] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2530–2540, 2019.

[179] U.S. Food and Drug Administration. Multiple endpoints in clinical trials: Guidance for industry. FDA guidance document, 2022. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry.

[180] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014. doi: 10.1214/14-AOS1221.

[181] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[182] Vladimir Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society: Series B*, 55(2):317–351, 1993.

[183] Vladimir Vovk. On-line confidence machines are well-calibrated. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 187–196. IEEE, 2002.

[184] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, 2012.

[185] Vladimir Vovk. Transductive conformal predictors. In *Proceedings of the International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360, 2013.

[186] Vladimir Vovk and Ivan Petej. Venn–Abers predictors. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.

[187] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4): 791–808, 2020.

[188] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

[189] Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Technical report, Royal Holloway, University of London, 2003. Technical report, https://alrw.cs.rhul.ac.uk/old/04.pdf.

[190] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[191] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489.

[192] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117.

[193] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

[194] Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3997–4005. PMLR, 2024.

[195] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *Electronic Journal of Statistics*, 18(1):2292–2338, 2024. doi: 10.1214/24-EJS2253.

[196] Wei Yang, Shixuan Li, Heng Ping, Peiyu Zhang, Paul Bogdan, and Jesse Thomason. Auditing multi-agent LLM reasoning trees outperforms majority vote and LLM-as-judge. arXiv preprint arXiv:2602.09341v1, 2026. URL https://arxiv.org/abs/2602.09341v1. February 2026 version.

[197] Daniel Yekutieli. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008. doi: 10.1198/016214507000001373.

[198] Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019. doi: 10.1214/18-AOS1756.

[199] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.

[200] Xianyang Zhang. Testing High Dimensional Mean Under Sparsity, 2015. URL https://arxiv.org/abs/1509.08444. arXiv:1509.08444v2.

[201] Xianyang Zhang. Mirror and Knockoff+ thresholds under dependence, 2026. URL https://arxiv.org/abs/2607.17084. arXiv:2607.17084v2.

[202] Xianyang Zhang and Jun Chen. Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association*, 117(537):411–427, 2022. doi: 10.1080/01621459.2020.1783273.

[203] Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017. doi: 10.1080/01621459.2016.1166114.

[204] Tijana Zrnic. A note on the prediction-powered bootstrap. arXiv preprint arXiv:2405.18379, 2024. Version 3 revised January 2025.

[205] Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010. PMLR, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

[206] Tijana Zrnic and Emmanuel J. Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024. doi: 10.1073/pnas.2322083121.

[207] Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values, 2024. URL https://arxiv.org/abs/2407.15733.

[208] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011. doi: 10.1214/11-STS356.

[209] Jelle J. Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218–1238, 2021. doi: 10.1214/20-AOS1999.

[210] Ziyu Xu, Aldo Solari, Lasse Fischer, Rianne de Heide, Aaditya Ramdas, and Jelle Goeman. Bringing closure to false discovery rate control: A general principle for multiple testing, 2026. URL https://arxiv.org/abs/2509.02517.

# Index


active statistical inference, 255, 352
AdaPT, 97
adjusted p-value, 351
alpha recycling, 44
approximate inverse, 213
arbitrary dependence, 66
average likelihood ratio, 25

Bayes factor, 130
benchmark multiplicity, 279
Benjamini–Hochberg procedure (BH), 54
Benjamini–Yekutieli procedure (BY), 66
Berk–Jones statistic, 25
Bonferroni procedure, 5, 38

calibrator, 125
canonicalization, 352
Cauchy combination test, 6
closed testing, 41
  Bonferroni, 42
  Simes, 42
closure principle, 41
conditional feature null, 169
conditional randomization test (CRT), 159
conditional sign-flip property, 79
confidence sequence, 312
conformal entailment, 291
conformal p-value, 194
conformal PRDS, 194
conformal prediction
  full, 188
  split, 189
conformalized quantile regression (CQR), 197
contamination audit, 289
covariate shift, 335
coverage
  conditional, 314, 351
  marginal, 185, 314, 351
  selective, 352
  simultaneous, 351
cross-fitting, 264, 352
cross-prediction-powered inference, 264
cross-validation
  selective inference, 249

data carving, 249
data fission, 245
data thinning, 245
debiased Lasso, 213
  diagnostics, 227
decorrelated score, 221
dense alternatives, 8
desparsified Lasso, 213
detection boundary, 27

e-BH, 141
e-process, 300, 352
e-value, 125, 352
  assembling, 150
  combining, 150
  multiple testing, 141
exchangeability, 185, 351
exchangeability audit, 289

false coverage rate (FCR), 235, 351
false discovery rate (FDR), 54, 351
familywise error rate (FWER), 351
Fisher combination test, 6
fixed-design inference, 162
fixed-design target, 207

Gaussian sequence model, 8
graphical multiple testing, 44
group BH, 107
group-level FDR, 107

heteroskedastic robust interval, 227
higher criticism, 25
Holm procedure, 38
Hommel procedure, 42
hypothesis weights, 62

independent hypothesis weighting (IHW), 62
inferential unit, 279

jackknife prediction interval, 200
jackknife+, 200

key resampling, 285
knockoff+ threshold
  under dependence, 79
knockoffs

fixed-X, 162
model-X, 169

language-model benchmark, 279
least-squares projection target, 207
likelihood ratio, 130
LLM-as-a-judge, 281
local FDR, 89, 351
logistic regression
decorrelated score, 223

marginal FDR (mFDR), 351
measurement bias, 281
mirror estimation, 68
mirror procedure
under dependence, 79
mirror threshold, 68
multilayer FDR, 110

negative control, 17, 351
Neyman orthogonality, 221
nodewise Lasso, 218
null proportion estimation, 85

one-step estimation, 221
optional continuation, 300
optional stopping, 300

p-filter, 110
partial conjunction, 116
polyhedral selective inference, 239
POSI, 238
positive dependence, 74
positive FDR (pFDR), 100, 351
post-selection inference, 238
PRDS, 74, 351
precision matrix estimation, 218
prediction-powered inference, 255, 352
power tuning, 259
recalibration, 265
projection target, 352

q-value, 88, 351

random-design target, 207
randomized selective inference, 249
rare/weak model, 27
reference-relative factuality, 291
replicability, 116

scan statistic, 285
selected-family FDR, 111, 351
sequential conditional independent pairs (SCIP), 173
Sidak procedure, 5, 38
Simes procedure, 22
sparse alternatives, 8
Storey estimator, 85
Stouffer combination test, 6
super-uniform p-value, 38, 351

TreeBH, 111
truncated-normal pivot, 239
two-groups model, 89

universal inference, 130

watermark detection, 285
weighted BH, 62
weighted conformal prediction, 335