Multiplicity, Selection, and Modern Statistical Inference

Xianyang Zhang

Draft version, August 2026.

Comments, corrections, and suggestions are welcome.

Contents

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.
- Basic real analysis sufficient to read ϵ-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 familywise and false-discovery error rates, including closure, graphical procedures, BH, 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 e-BH; advanced sequential derivations are collected in Appendix A. Chapter 9 develops conditional randomization tests, 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 the debiased Lasso, 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. Chapter 13 uses applications in genomics, watermarking, contamination auditing, and reference-relative factuality to show how the same statistical principles reappear in modern scientific and AI workflows.

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 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)
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}	ith 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 $R = \lvert\mathcal{R}\rvert$
U	true nulls not rejected, $U = m_0 - V$
V	number of false rejections, $V = \lvert\mathcal{R} \cap \mathcal{I}_0\rvert$
M	nonnulls not rejected (missed discoveries), $M = m_1 - D$
D	number of true discoveries, $D = R - V$
$\widehat{\mathcal{S}}$	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}
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_{safe}	strict-safe score for reference-relative factuality
T_{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
Φ, ϕ	standard normal CDF and density
$\mathrm{Unif}(0,1)$	uniform distribution on $[0,1]$
χ^2_k, χ_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

Regression and high-dimensional models

$X \in \mathbb{R}^{n\times p}$	design matrix (rows are observations, columns are features)
$X_j \in \mathbb{R}^n$	jth column of X (the jth feature across all observations)
X_{-j}	the design matrix with column j removed
$x_i^\top \in \mathbb{R}^{1\times p}$	ith 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 β
$\mathcal{S}_\beta$	coefficient support $\{j : \beta_j \neq 0\}$
θ, Θ_0	generic parameter and null parameter set
Σ, Γ	covariance matrix and precision matrix Σ^{-1}
$\widetilde{X}$	knockoff copy of X (Chapter 9)

Conventions

- Boldface is not used for vectors; whether a symbol is scalar or vector is clear from context.
- Uppercase letters denote random variables and lowercase letters denote realizations whenever the distinction matters.
- $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$ are the real, integer, and natural numbers.
- All limits are as $m \to \infty$ or $n \to \infty$ unless otherwise stated; in particular, $o(\cdot)$ and $O(\cdot)$ refer to that limit.
- The symbol q is reserved for an FDR target. The symbol α denotes a single-test Type I error, familywise error budget, or prediction miscoverage level, as stated locally.
- A hypothesis-specific q-value is written with an uppercase Q, for example Q_i, $Q_{(i)}$, or $Q(z)$. This distinguishes it from the prespecified scalar FDR target q.
- Threshold rules use $R(t) \vee 1$ to avoid division by zero before any rejection has been made. Supremum and infimum rules follow the convention stated where the rule is introduced; when a proof uses the inequality at the selected threshold, the chapter either searches over a finite grid or assumes enough one-sided continuity or closedness for the selected threshold to be attained.
- The blackboard symbols $\mathbb{P}, \mathbb{E}$ denote the abstract probability and expectation operators. Italic $P_0, P_1, P_\theta, E_\theta$ and similar denote *specific* probability measures and the corresponding expectations; the subscript identifies which measure.

- Common distributions are written in roman: $\mathrm{Bin}(n, p)$ for binomial, $\mathrm{Bernoulli}(p)$, and $\mathrm{Unif}(0, 1)$ for uniform.
- Group counts use G, while benchmark tasks and model–task cells use B_{task} and B_{cell} when a count is needed. Cluster separation is δ_{sep}, distinct from the normal density ϕ. Time-indexed processes may still use T as a terminal time, but decision-table counts use D and M.

CHAPTER 1

Introduction: Multiplicity, Selection, and Modern Inference

The trouble often starts after the first successful plot. A genomic screen has a striking gene near the top of a volcano plot. A high-dimensional regression has a small set of variables selected by the Lasso. A model evaluation table has several cells where a new method appears to beat the baseline. Each finding may be real. Each may also be the best-looking accident among many chances to look.

For example, suppose 20,000 null genes are tested and the smallest p-value is 10^{-5}. On its own, 10^{-5} looks impressive. But if all 20,000 p-values were independent uniforms, then

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

The surprise fades once the family of questions is visible. The calculation is elementary, but the lesson is deep: evidence is interpreted relative to the search that produced it.

The same issue appears in less obvious forms. If a regression analyst fits a Lasso, chooses one selected coefficient, and then reports the ordinary least-squares confidence interval as if that coefficient had been fixed in advance, the interval is answering the wrong question. If a benchmark paper compares many models across many tasks and highlights only the biggest gains, unadjusted p-values exaggerate the evidence. If a prediction method is tuned on many subgroups and then advertised as calibrated, we need to know which coverage statement survived the tuning.

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, and selective inference. 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. 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 Benjamini–Hochberg discovery list controls an average false discovery proportion, but it does not promise that every selected gene is real. 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 Grammar

Route: Core

Most chapters follow the same grammar:

$$\text{define the claim} \quad \longrightarrow \quad \text{calibrate evidence} \quad \longrightarrow \quad \text{report a guaranteed object.}$$

The evidence may be a p-value, an e-value, a residual score, a feature statistic, or a randomized comparison. The guarantee may concern FWER, FDR, finite-sample coverage, optional stopping, or a post-selection law. What changes from chapter to chapter is the symmetry or null structure that makes calibration possible.

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.

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 familywise error control with false-discovery 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 conditional randomization tests, knockoff filters, and multilayer knockoffs. Chapter 10 develops distribution-free prediction through conformal methods, including conformal risk control and beyond-exchangeability extensions. Chapters 11 and 12 treat inference after high-dimensional modeling and data-dependent selection, including data thinning and fission for valid inference after unsupervised discovery steps such as clustering. Chapter 13 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 read the chapters nearly in order. Applied researchers may instead start from the 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.

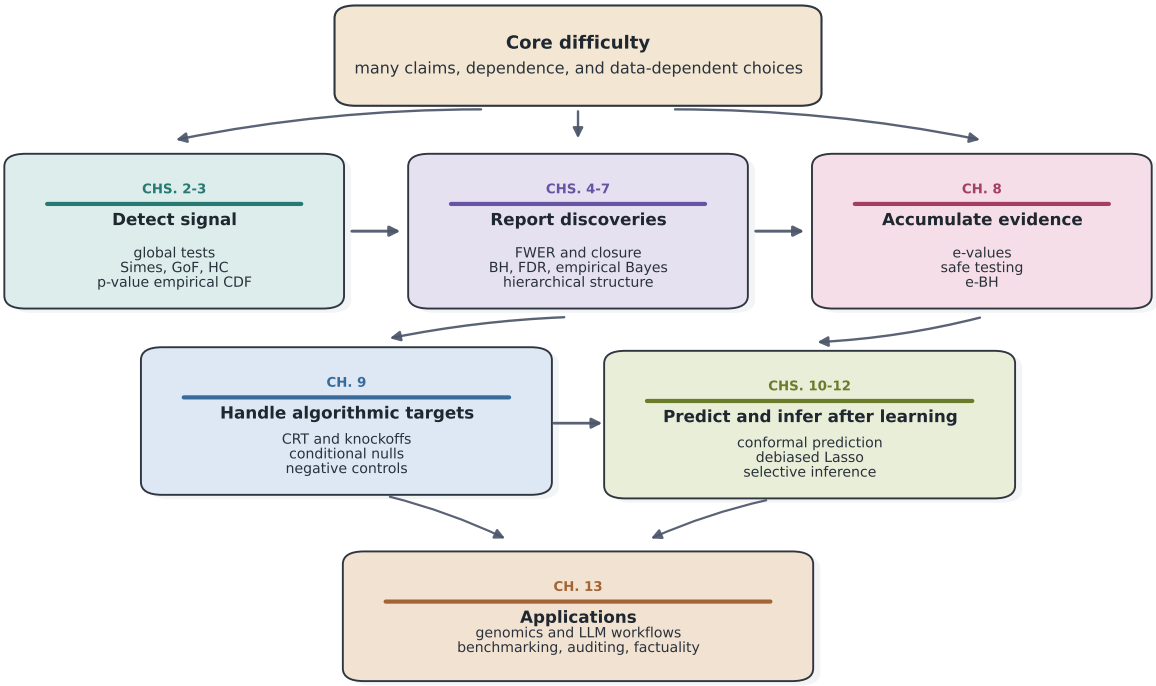

FIGURE 1.1. Conceptual map of the book. Global testing asks whether signal is present. FWER, FDR, hierarchical, and e-value methods turn calibrated evidence into discovery rules. Knockoffs, conformal prediction, debiased Lasso, and selective inference handle settings in which algorithms or data-dependent choices determine the object being reported. The applications chapter revisits the same questions in genomics and language-model evaluation.

Statisticians who already know classical multiple testing may use the later chapters as a bridge to knockoffs, conformal prediction, and selective inference.

The chapters use a common guarantee-oriented grammar: identify the inferential target and reported object, state the assumptions that support the method, and then state the resulting error or coverage guarantee. These elements recur throughout each chapter, but they are not always collected in a single opening "contract." Section labels mark *Core*, *Advanced*, *Optional proof*, or *Research frontier*. A first course can follow the Core route and treat advanced guarantees and proofs as reading assignments; Appendices A and B collect the most technical safe-inference and conformal derivations. Assumption diagnostic cards separate what can be checked empirically from what must be defended scientifically or by design.

Before applying any method, keep five questions in view. What is the reported object? What family or selection rule does the guarantee cover? What null, conditional null, or exchangeability condition justifies the calibration? How does dependence affect the argument? And is the stated error rate the one the scientific claim actually needs?

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 means that at least one individual null is false. It does not identify which one. The localization problem, where we decide which hypotheses to reject while controlling familywise error or false discovery rate, 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 [48, 114]. 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 task j does not provide evidence that system A improves on system B. For the benchmark Gaussian approximation, assume that within each task the paired differences $D_{j1}, \dots, D_{jn_j}$ 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 $N(\mu_j, 1)$, by the central limit theorem and Slutsky's theorem, when n_j is large enough for that approximation to be accurate. The boundary case $\delta_j = 0$, equivalently $\mu_j = 0$, corresponds to no mean improvement on task j, while $\delta_j > 0$ 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 Θ_{0j} denote all parameter values satisfying H_{0j}. A p-value for this possibly composite null is valid if it is uniformly super-uniform over that set:

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

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 is

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

whereas the one-sided upper-tail p-value 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 α 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 α, 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 α. 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 α, 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, within half a percentage point of the nominal level. So under independence Bonferroni gives up almost nothing; it is a near-exact level-α 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 α/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 under the global null,

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

Fisher's test rejects for large T_F. The exact chi-square calibration is an independence statement; under dependence, the same statistic may still be useful, but its null distribution must be justified separately.

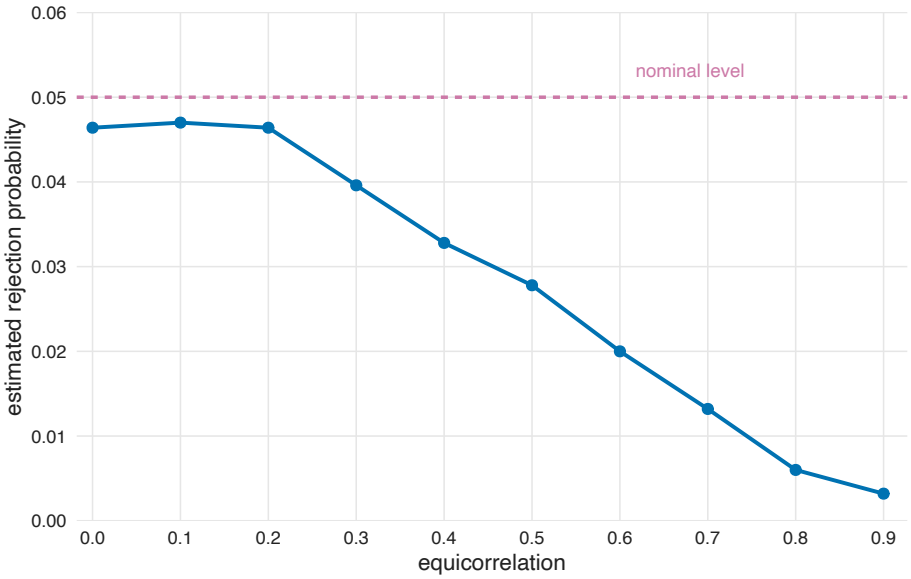

FIGURE 2.1. Estimated size of Bonferroni's two-sided global test under an equicorrelated Gaussian null with $m = 200$, $\alpha = 0.05$, and 5000 Monte Carlo repetitions. The test remains valid, but positive correlation makes it increasingly conservative.

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 χ^2_{2m} under independent uniform p-values and rejects for large values; it is sensitive when unusually *large* p-values are evidence against the null (the opposite tail from Fisher's combination of unusually small p-values). Stouffer's method maps p-values to z-scores. For one-sided p-values, define $X_j = \Phi^{-1}(1 - p_j)$. Under independent nulls,

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

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 evidence. These choices encode assumptions about the shape of the alternative before any asymptotic theory is invoked. This opposite-tail case is not just a mathematical curiosity: if the p_j's are one-sided upper-tail p-values for beneficial treatment effects, then many unusually large p-values are evidence that the treatment may be harmful rather than beneficial.

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. For p-values obtained from jointly normal test statistics, the upper tail of W matches a standard Cauchy variable C under broad dependence conditions in the explicit tail limit

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

[116]. The asymptotic variable here is the tail threshold t, not the number of hypotheses m. Thus the Cauchy test is often useful when one wants an analytic tail p-value calculation without assuming independence.

The same idea extends beyond the Cauchy distribution: one can combine p-values through other heavy-tailed calibrators. The point is not that Cauchy is magic; the point is that heavy tails preserve sensitivity to extremes while allowing a tractable tail approximation.

Figure 2.2 previews how the same ordered p-values support two different questions. The horizontal Bonferroni cutoff tests whether any p-value is exceptionally small. The Simes global test uses the rising line $\alpha i/m$, whereas the Benjamini–Hochberg (BH) step-up procedure uses qi/m at FDR target q. The two lines coincide in this preview because $q = \alpha$. BH reports a discovery set while controlling the false discovery rate under its stated conditions. The inset shows why the first few ranks drive the comparison.

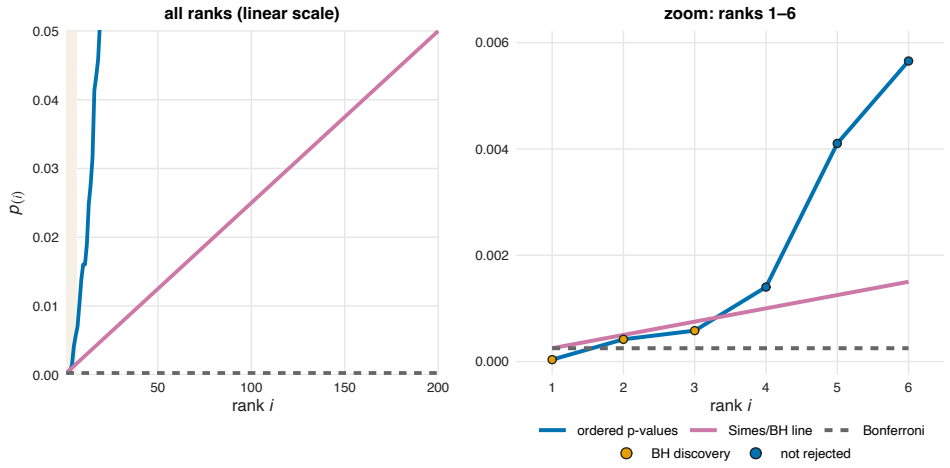

FIGURE 2.2. Sorted p-values from one simulated screen with $m = 200$ two-sided z-score tests at level $\alpha = 0.05$; six coordinates have mean 3.2, the rest are null. Bonferroni rejects p-values below the horizontal threshold $\alpha/m = 2.5 \times 10^{-4}$. Rank-based tests, specifically the Simes global test and the Benjamini–Hochberg (BH) step-up procedure, compare ordered p-values with rising lines. Simes uses $\alpha i/m$, while BH uses qi/m; here $q = \alpha = 0.05$, so the two lines coincide. These procedures appear in later chapters.

4. The Gaussian Sequence Model

Route: Advanced

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

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

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

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 ϕ 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 $(1-u_m)^m \le \exp(-m u_m) \to 0$, 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 has a positive mean $\mu^* > 0$, while all others have mean zero. If

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

then the maximum test has power tending to one. The signal coordinate alone crosses the null extreme-value threshold with probability tending to one.

If instead

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

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,

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

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 α, $1-e^{-\alpha} = \alpha + O(\alpha^2)$, 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-α test has power tending to one uniformly over the unknown-location class

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

where $e_j \in \mathbb{R}^m$ is the jth standard basis vector, with a one in coordinate j and zeros elsewhere. Thus $\mu^* e_j$ represents a signal of size μ^* 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.

Choose J uniformly from $\{1,\dots,m\}$, set $\mu_J = \mu^*$, and set all other means to zero. Under the null, the density of $Z = (Z_1,\dots,Z_m)$ is f_0. Under this mixture alternative, the likelihood ratio 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 $0 < \mu^* \le (1-\epsilon)\sqrt{2\log m}$ for some fixed $\epsilon > 0$, then $L_m(Z) \xrightarrow{P_0} 1$. Consequently, for every sequence of level-α rejection regions A_m, the mixture-average power satisfies

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

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 ϵ is sufficiently large: $\epsilon > 1 - 1/\sqrt{2}$. For smaller ϵ, μ^* 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 $e^{(\mu^*)^2} \le m^{1/2+o(1)}$, so $m^{-1}E_0\widetilde{Y}_{m1}^2 \to 0$. If $r_m > 1/4$, Mills' ratio gives

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

up to a polynomial factor in $\log m$. Therefore

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

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 rewards aggregate energy: any μ of large enough ℓ_2 norm will inflate T_m above its null mean, regardless of how that norm is distributed across coordinates.

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

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

Thus an asymptotic level-α L_2 test rejects when

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

Under a fixed mean vector μ,

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

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

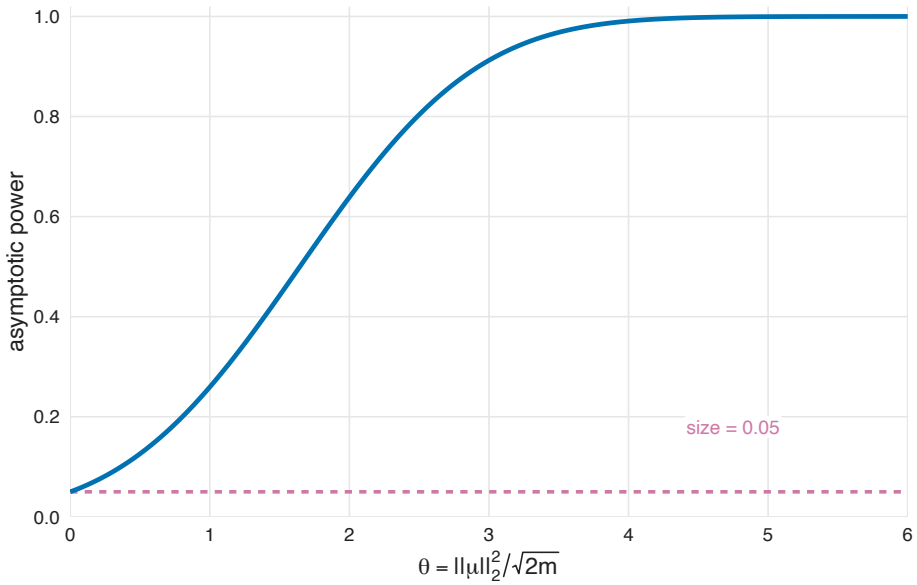

FIGURE 2.3. Asymptotic power curve of the L_2 test as a function of $\theta = \|\mu\|_2^2/\sqrt{2m}$, for $\alpha = 0.05$. Dense alternatives are visible once the total squared signal is on the $\sqrt{m}$ scale.

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 spherical-mixture and contiguity argument 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. The calculation is the same second-moment argument used in Le Cam's first lemma. 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. By symmetry, odd moments of U_1 vanish, and

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

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 $\rho^2/\sqrt{2m} \to 0$ up to constants, 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 next proposition records exactly which Le Cam facts are being used. 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 tools used in the lower bounds

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, $L_m \to 1$ under P_m. The tilt in Le Cam's third lemma is therefore trivial: any statistic with a null limiting distribution has the same limiting distribution under the spherical alternative. Such a statistic cannot yield asymptotic power above its size. This is the formal version of the phrase "the two experiments are asymptotically indistinguishable."

7. Comparing the Maximum and L_2 Tests

Route: Advanced

The maximum test and the L_2 test measure different features of μ:

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

Neither dominates the other.

Example 2.6: Many strong but still sparse signals

Suppose $m^{3/8}$ coordinates of μ 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 energy 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 μ equal a fixed constant a, and the rest are zero. Then $\|\mu\|_\infty = a$, which is far below $\sqrt{2\log m}$, so the maximum 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 2.4. 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. Monte Carlo rejection probabilities for four global tests. The Gaussian sequence simulation uses $m = 500$, $\alpha = 0.05$, and 1500 Monte Carlo repetitions.

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

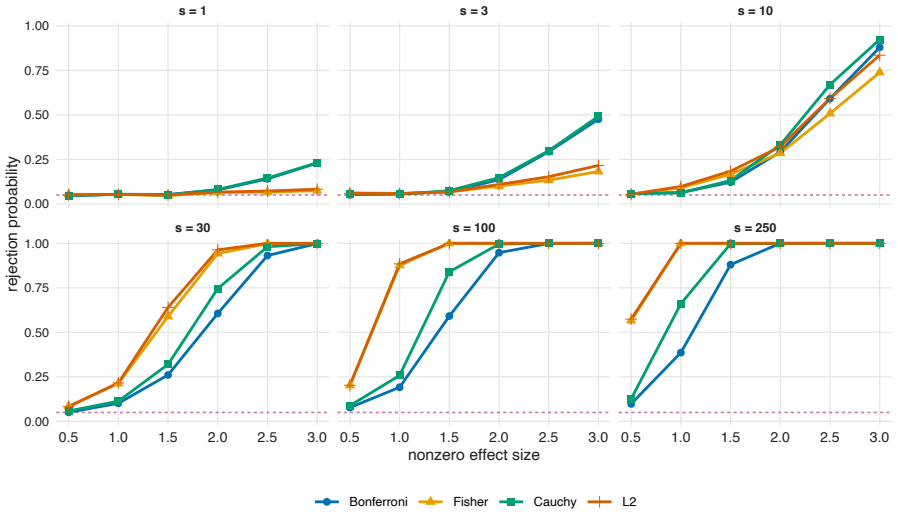

FIGURE 2.4. Estimated rejection probabilities at level $\alpha = 0.05$ in a Gaussian sequence model with $m = 500$ coordinates. Each panel fixes the number s of nonzero means and varies their common effect size; each curve is based on 1500 Monte Carlo replicates. Bonferroni reacts fastest when s is small 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.

8. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Inspect p-value histograms, dependence plots, calibration simulations, and sensitivity across sparse and dense tests.
Not identified from these data	The exact alternative sparsity and the full joint null law are not identified from one realized testing vector.
Sensitivity analysis	Repeat conclusions across maximum-, aggregate-, and dependence-robust calibrations.
Conservative fallback	Use Bonferroni or a valid permutation/global randomization test when dependence is uncertain.

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 much less sensitive to Gaussian dependence, but it is still a calibration statement that should be checked when the p-values come from a different model. 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 energy tests are natural for many weak signals, not to claim that one test is best in every finite-sample study.

9. Bibliographic Notes

Route: Core

Fisher's combination test is classical [56], and Stouffer's normal-score combination was developed in the social-science meta-analysis literature [152]. Sidak's exact independent-threshold calibration is due to Šidák [145]. The Cauchy combination test and its dependence-robust tail approximation are from Liu and Xie [116]. The contiguity arguments used in the lower bounds are Le Cam arguments; see van der Vaart [163, Chapter 6], Le Cam [101], and Le Cam and Yang [102]. Ingster's work on Gaussian sequence testing made precise how sparse alternatives can lead to non-Gaussian and infinitely divisible likelihood-ratio limits [86]; modern high-dimensional detection-boundary treatments include Donoho and Jin [47] and Arias-Castro et al. [5].

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 α/m and the Sidak threshold $1-(1-\alpha)^{1/m}$. Verify numerically that they agree to leading order and that the Bonferroni size under independence is close to $1 - e^{-\alpha}$.

Exercise 2.9: Fisher's chi-square distribution

Let $p_1, \dots, p_5$ be independent $\mathrm{Unif}(0,1)$. Compute the distribution of $T_F = -2\sum_{j=1}^{5} \log p_j$ and find the 0.95 quantile of T_F. 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 p-value 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 $p \sim \mathrm{Unif}(0,1)$ 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: Maximum scale

Use Mills' ratio to prove

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

for independent standard normal Z_j's. Then show that $\Phi^{-1}(1-\alpha/m) = \sqrt{2\log m}\{1 + o(1)\}$.

Exercise 2.13: 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 more general approximation with denominator $\sqrt{2m + 4\|\mu\|_2^2}$. Use the fact that $E\xi_j^3 = 0$.

Exercise 2.14: Sparse mixture likelihood ratio

For the one-sparse mixture alternative with $\mu^* = (1-\epsilon)\sqrt{2\log m}$, derive the likelihood ratio

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

Show that $E_0 L_m = 1$. For which values of ϵ does the elementary second-moment argument prove $L_m \to 1$ in $L^2(P_0)$? Why is a truncation argument needed for the full range $\epsilon > 0$?

Exercise 2.15: Cauchy combination under independence and under dependence

Suppose $p_1, \dots, p_m$ are independent $\mathrm{Unif}(0,1)$, so that the transformed quantities $C_j = \tan\{\pi(1/2 - p_j)\}$ are iid standard Cauchy. For any nonnegative weights w_j with $\sum_j w_j = 1$, use the stability of the Cauchy distribution under positive-weight linear combinations to prove that $W = \sum_j w_j C_j$ is *exactly* standard Cauchy.

Now suppose $p_1, \dots, p_m$ come from jointly normal test statistics with arbitrary covariance. Following Liu and Xie [116], illustrate numerically that W is no longer exactly Cauchy by comparing its empirical distribution to the standard Cauchy in the body of the distribution, and verify that

the upper-tail approximation $\mathbb{P}(W \geq t) \approx 1/(\pi t)$ becomes accurate as the threshold t grows large. (A simulation cannot prove an asymptotic tail equivalence; for the proof of $\mathbb{P}(W \geq t) = \{1/(\pi t)\}\{1+o(1)\}$ under broad dependence, see Liu and Xie [116].) Explain why this tail equivalence is the genuinely useful result: it justifies the Cauchy calibration of W under broad dependence, where the exact distribution is intractable.

Computational.

Exercise 2.16: Bonferroni under equicorrelation

Let $Z \sim N(0, \Sigma_\rho)$, 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 size of the two-sided Bonferroni global test at level 0.05 by Monte Carlo with at least 5000 replicates. Compare your output with Figure 2.1. Explain why positive correlation changes the observed size but not the validity guarantee.

Exercise 2.17: Sparse versus dense simulation

Reproduce Figure 2.4. 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 2.18: Fisher under dependence

Simulate equicorrelated Gaussian z-scores under the global null and compute two-sided p-values. Apply Fisher's test using the independent chi-square calibration. How does the Type I error change with correlation? Compare this behavior with Bonferroni.

Advanced.

Exercise 2.19: Dense lower bound

For the spherical prior alternative $\mu = \rho U$, derive the likelihood ratio and verify that

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

where U_1 is the first coordinate of a uniform point on $\mathcal{S}^{m-1}$. Use this to show that the likelihood ratio converges to one in P_0-probability when $\rho^2/\sqrt{2m} \to 0$, so no test can have asymptotic power above its size against the spherical mixture. Then show the converse: when $\rho^2/\sqrt{2m} \to \infty$, the L_2 test itself has power tending to one. (Note that the likelihood ratio cannot diverge under P_0 — its P_0-mean equals 1; divergence is a statement under the alternative P_1.)

Exercise 2.20: Truncation for the sparse lower bound

Complete the truncation step in Proposition 2.4. Define the truncated likelihood ratio $\widetilde{L}_m = m^{-1} \sum_j \exp(\mu^* Z_j - (\mu^*)^2/2)\, \mathbf{1}\{Z_j \leq c_m\}$ for a sequence $c_m \to \infty$, choose c_m so the second moment is $1 + o(1)$, and bound the tail contribution $E_0|L_m - \widetilde{L}_m|$.

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 goodness-of-fit test for the uniform distribution, with different procedures looking at different parts of the empirical CDF.

This viewpoint is useful because it makes the geometry of rare signals visible. If a few alternatives produce very small p-values, the empirical CDF 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), higher criticism (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 empirical CDF

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

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.

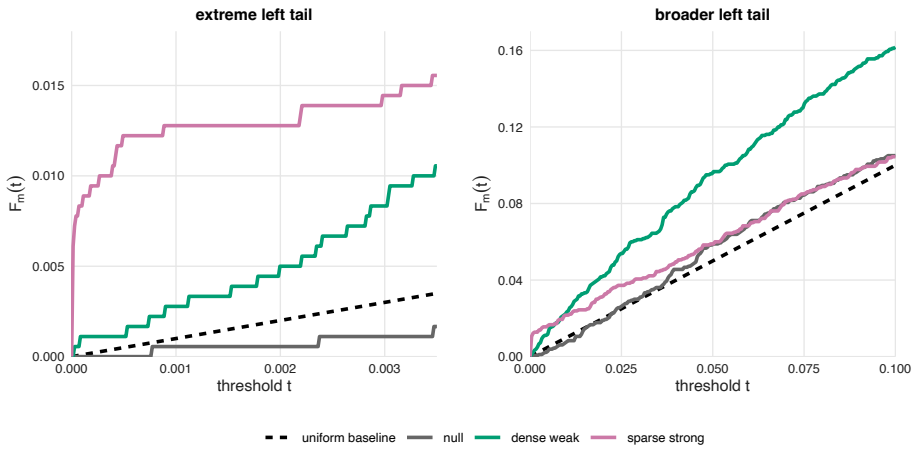

FIGURE 3.1. Left-tail empirical CDFs of one-sided p-values under a null model, a dense weak alternative, and a sparse stronger alternative. In this stylized draw, $m = 1800$, the dense weak case shifts 360 coordinates by 1.0, and the sparse strong case shifts 24 coordinates by 4.2. The diagonal is the 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 CDF 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-α Simes test rejects when $p_{\text{Simes}} \le \alpha$.

Figure 3.2 shows the same rule rank by rank. The first ordered p-value below the line certifies a global departure, while the rank of that crossing 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

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

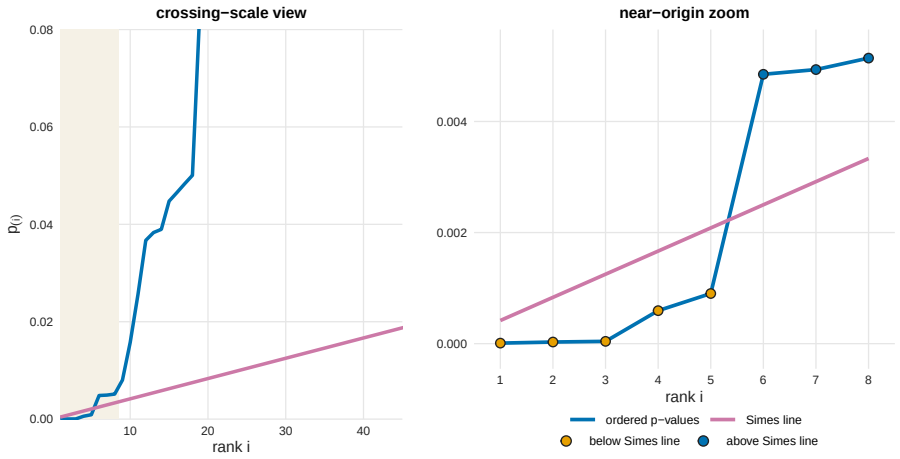

FIGURE 3.2. The Simes crossing rule. Ordered p-values are compared with the line $\alpha i/m$. A crossing gives a global rejection. Later, the same linear boundary becomes the basis of the BH multiple-testing procedure. The illustration uses $m = 120$, $\alpha = 0.05$, and one-sided p-values from eight 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 α:

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

The second term is α^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.

The one-sided Kolmogorov-Smirnov 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 Cramer-von Mises 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 Anderson-Darling 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 higher criticism: if the alternatives are rare, the left tail should be standardized more aggressively than the center of the distribution.

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 Anderson-Darling statistic is symmetric in the two tails.

Proposition 3.2: Brownian bridge covariance

If $p_1, \dots, p_m$ 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 α_m. The case $t < s$ is symmetric. □

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

Route: Advanced

The empirical CDF at a fixed threshold has variance $t(1-t)/m$. This suggests the standardized statistic

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

Higher criticism maximizes this quantity over left-tail thresholds, taking only the positive part so that empty crossings do not contribute negative values. 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 empirical CDF is at or below the diagonal are skipped.

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 calculation in the spirit of Darling-Erdős 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, by permutation, 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.

Berk-Jones replaces the normal approximation in HC by an exact binomial likelihood ratio. Let

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

This is the Kullback–Leibler divergence $D(\text{Bernoulli}(x) \,\|\, \text{Bernoulli}(t))$, 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. The one-sided Berk-Jones statistic is

$$\text{BJ}_m = \max_{1 \le i \le \lfloor m/2 \rfloor} m\,h\left(\frac{i}{m}, p_{(i)}\right) \mathbf{1}\left\{p_{(i)} < \frac{i}{m}\right\}.$$

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 $N(t)/m = i/m$) at each rank i and takes the maximum, scanning over which rank gives the most extreme binomial likelihood-ratio departure.

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.

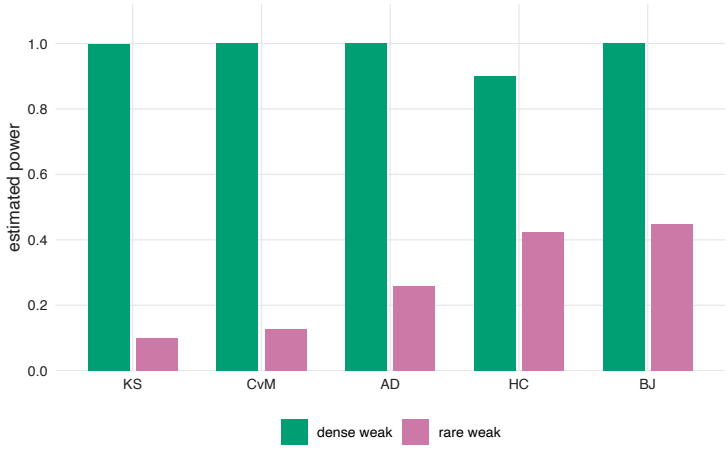

FIGURE 3.3. Monte Carlo power for KS, CvM, AD, HC, and BJ with $m = 700$ one-sided Gaussian p-values and nominal level 0.05. Each method's cutoff is its empirical 95th percentile from 1,200 null replicates of 700 iid Unif$(0,1)$ p-values; power uses 900 replicates per alternative. In each alternative 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 $N(0,1)$. The generator uses seed 6892026.

The average likelihood ratio statistic smooths this idea across ranks:

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

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:

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

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 choice such as $w_i \propto 1/(i \log m)$ spreads prior weight over tail locations on a scale-invariant grid. 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 β 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-α tests whose power tends to one. For $r < \rho^*(\beta)$, no sequence of level-α 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; we cite the Donoho-Jin formulation because it is the one most directly connected with higher criticism [47, 86].

Figure 3.4 is a map of this theorem rather than a second test definition. For a fixed sparsity exponent β, move vertically in signal exponent r: crossing the solid boundary changes the asymptotic problem from undetectable to detectable, and the dashed curve shows where an extreme-coordinate rule alone reaches that transition.

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

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

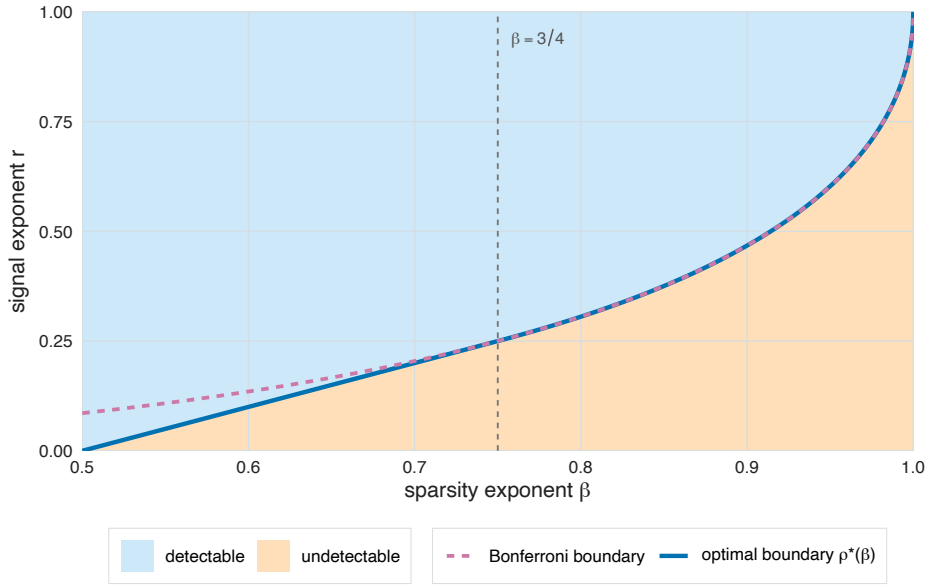

FIGURE 3.4. Rare/weak detection boundary for Gaussian mixtures. Above $\rho^*(\beta)$, detection is possible; below it, detection is impossible. Bonferroni is rate-optimal in the very sparse region $3/4 \le \beta < 1$, but not throughout the whole rare/weak regime. The vertical dotted line marks $\beta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

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. Higher criticism and Berk-Jones 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 $\Phi^{-1}(1-m^{-u}) = \sqrt{2u\log m}\{1 + o(1)\}$. 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}, a standardized tail count such as higher criticism 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 [47] 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, higher criticism has asymptotic power tending to one whenever $r > \rho^*(\beta)$, without knowing β or r in advance. If $r < \rho^*(\beta)$, no sequence of level-α 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 β, 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 separates a calibration failure from a power gain. Its first panel asks whether an iid cutoff preserves size under first-order autoregressive (AR(1)) noise; only after answering that question should the second panel's rejection rates be interpreted as power.

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

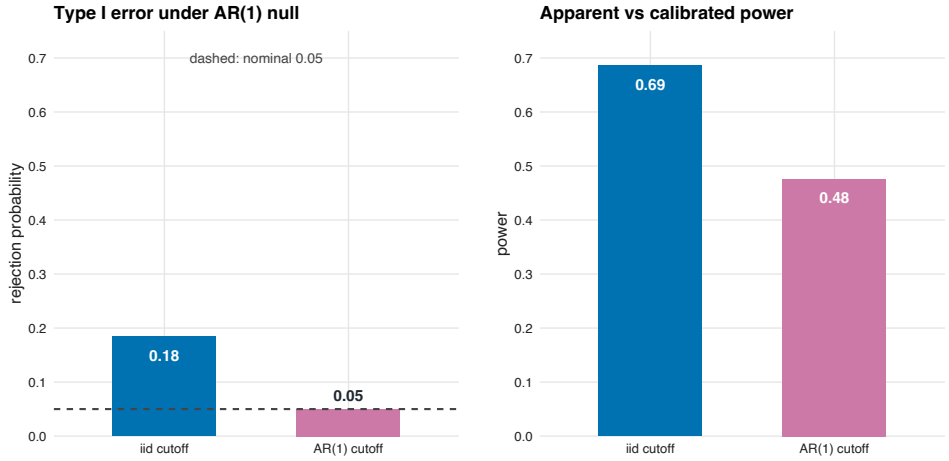

FIGURE 3.5. Dependence changes both calibration and apparent power. The simulation uses $m = 400$ one-sided Gaussian p-values with AR(1) null correlation $\rho = 0.85$. Cutoffs are Monte Carlo 95th-percentile null cutoffs, either under iid uniforms or under the AR(1) null. Under the AR(1) null, the iid cutoff is too low and rejects too often; under a sparse alternative with 8 randomly located signals of size 2.4, the same too-low cutoff also gives inflated apparent power. The AR(1)-calibrated cutoff restores size control and reports the honest lower power. The dashed line in the left panel marks the nominal level 0.05; it has no analogue in the power panel. Each null and alternative estimate uses 5,000 Monte Carlo replications.

higher criticism. If $X = \mu + Z$ with $Z \sim N(0, \Sigma)$, then a whitening operator A_{wh} satisfying $A_{\mathrm{wh}}\Sigma A_{\mathrm{wh}}^{\top} = I$ turns the noise white:

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

For many correlated models, such as Toeplitz or decaying correlations, the transformation can increase the effective strength of sparse signals. The iid case can therefore be the hardest case at a fixed sparsity and signal strength. The catch is that ordinary HC applied to the untransformed coordinates ignores this information, so it can suffer from both wrong calibration and lost power.

A related likelihood-ratio calculation explains why precision adjustment is a principled statistic for sparse high-dimensional mean testing. Consider n independent observations $X_1, \ldots, X_n$ in $\mathbb{R}^m$ with mean μ and known covariance matrix Σ, and let $\Gamma = \Sigma^{-1}$ be the precision matrix. The sample mean $\bar{X} = n^{-1}\sum X_\ell$ has covariance Σ/n, and we define the precision-adjusted vector

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

Up to the positive factor n, S_μ is the Gaussian likelihood score for the mean vector at $\mu = 0$. Thus $S_{\mu,j}$ is the score for a local shift in coordinate j with the other mean coordinates held fixed; this is a statement about that specified score direction, not a blanket locally-most-powerful claim when the other means are unrestricted nuisance parameters. Because its covariance is Γ/n, not a multiple of the identity, this operation is a precision adjustment rather than a whitening.

For a fixed candidate support S, maximizing the Gaussian likelihood over mean vectors supported on S gives the likelihood-ratio statistic

$$T_S = n\, S_{\mu,S}^{\top}\Gamma_{S,S}^{-1}S_{\mu,S},$$

and T_S has a $\chi^2_{|S|}$ distribution under $\mu = 0$. 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,

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

where γ_{jj} is the jth diagonal of Γ. This is the one-sparse case of the likelihood-ratio derivation in Zhang [180] and the basis of the two-sample precision-transformed maximum test of Cai et al. [35].

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 μ_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.

The practical lesson is simple, but conditional on what can be learned from the data. If dependence is weak or well approximated by a known null model, simulate or permute from that model. 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 this is a real modeling requirement, not a free preprocessing step: estimating Γ usually requires assumptions such as sparse precision rows or a sparse graphical model, with methods such as nodewise Lasso, constrained ℓ_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

Assumption diagnostic	
Checkable	Check null calibration on negative controls and compare empirical dependence with the calibration model.
Not identified from these data	Exact tail dependence and the rare/weak asymptotic regime cannot be certified from the observed extremes alone.
Sensitivity analysis	Vary truncation ranges and compare Simes, higher criticism, Berk–Jones, and permutation calibration.
Conservative fallback	Use an exact randomization test or a conservative combination bound.

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. Higher criticism and Berk-Jones rely on tail standardization; their finite-sample thresholds should be calibrated under the actual null dependence when possible. 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.

8. Bibliographic Notes

Route: Core

The Simes global test is due to Simes [146]. Higher criticism and the rare/weak detection boundary are presented in the form of Donoho and Jin [47]. The earlier minimax testing work of Ingster [86] 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 higher criticism follows from law-of-the-iterated-logarithm arguments going back to Darling and Erdős [42]. Berk-Jones statistics go back to Berk and Jones [24]; modern comparisons of higher criticism, Berk-Jones, and related phi-divergence statistics include Jager and Wellner [89]. Innovated higher criticism and the possibility that dependence can improve sparse-signal detection are developed by Hall and Jin [73]; related heterogeneous-mixture boundaries appear in Cai et al. [33]. Precision-transformed maximum and sparse likelihood-ratio tests for high-dimensional means are developed by Cai et al. [35] and Zhang [180]. Precision-matrix estimation methods used in such settings include nodewise Lasso [121], CLIME [34], and graphical Lasso [60].

9. Exercises

Route: Core

Basic.

Exercise 3.4: Simes crossing

For $m = 5$ p-values

$$0.003,\ 0.021,\ 0.040,\ 0.31,\ 0.78,$$

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

Exercise 3.5: Variance of the empirical CDF

Assume the p-values are independent uniforms. Derive

$$\operatorname{Var}\{F_m(t)\} = \frac{t(1-t)}{m}.$$

Use this calculation to motivate the denominator in higher criticism.

Exercise 3.6: Brownian bridge covariance

Derive the covariance function

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

for the limiting empirical process.

Intermediate.

Exercise 3.7: 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 3.8: Berk-Jones likelihood ratio

Fix a threshold t and let $N(t) = mF_m(t)$. Under the null, $N(t) \sim \mathrm{Bin}(m, p)$ with $p = t$. Writing $\widehat{p} = N(t)/m$ for the empirical fraction, show that the binomial log likelihood ratio for testing the null $p = t$ against the data-driven $p = \widehat{p} > t$ is

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

Conclude that the Berk-Jones statistic is, at each rank, a binomial log-likelihood ratio against the null success probability $t = p_{(i)}$.

Exercise 3.9: Simulation comparison

Simulate one-sided Gaussian p-values under a sparse mixture. Compare Kolmogorov-Smirnov, Anderson-Darling, higher criticism, and Berk-Jones using Monte Carlo null calibration. Report both size and power.

Computational.

Exercise 3.10: Average likelihood ratio

Implement the average likelihood ratio statistic with weights $w_i \propto \{i \log m\}^{-1}$ for $1 \le i \le m/2$. Compare its power with HC and BJ in rare/weak simulations.

Exercise 3.11: Rare/weak phase diagram

For a grid of (β, r), simulate the Gaussian mixture model with $\varepsilon_m = m^{-\beta}$ and $\mu_m = \sqrt{2r\log m}$. Estimate the power of Bonferroni, HC, and BJ. Overlay the theoretical boundary $\rho^*(\beta)$ and summarize where each method is strongest.

Exercise 3.12: Dependence calibration of HC

Reproduce Figure 3.5 for an AR(1) Gaussian null with $\rho = 0.5$ and $m \in \{200, 1000\}$. Compute the iid-calibrated and AR(1)-calibrated cutoffs by Monte Carlo and report the realized Type I error of each under the AR(1) null.

Advanced.

Exercise 3.13: Bonferroni in the very sparse regime

Show that the maximum test detects the rare/weak mixture when

$$\sqrt{r} + \sqrt{1-\beta} > 1.$$

Use this to explain why Bonferroni is rate-optimal for $3/4 \le \beta < 1$.

Exercise 3.14: Dependence-aware sparse testing

In a Gaussian mean model with known covariance Σ, compare ordinary maximum testing with the precision-adjusted statistic

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

Construct an example where the precision-adjusted statistic has larger power, and one where it has smaller power. What property of Γ relative to the true μ determines which regime applies?

Exercise 3.15: Simes uniformity by induction

Refine the $m = 2$ calculation in the proof of Theorem 3.1 to a general inductive argument. Show that under iid uniform p-values, the Simes statistic $T_m = \min_i mp_{(i)}/i$ is exactly $\mathrm{Unif}(0, 1)$, not merely super-uniform.

CHAPTER 4

Familywise Error Rate, Closure, and Graphical Procedures

The first two chapters asked whether a collection of tests contains any signal. That is a global question. In practice, the analyst usually wants individual decisions: which endpoints, genes, models, subgroups, or validation checks can be declared significant? Once individual decisions are reported, the relevant error is no longer only the Type I error of a global test. The analyst must control the chance of making at least one false rejection among the reported discoveries.

This chapter develops familywise error rate control as an error-budgeting problem. Bonferroni spends the budget evenly. Holm spends it sequentially. Closure explains why these procedures are strongly valid. Graphical procedures then give a practical language for more structured testing plans, where alpha can flow 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. Familywise error control asks for a guarantee on that family as a whole: with probability at least $1 - \alpha$, every reported rejection of a true null should be absent. This is a stricter goal than controlling the expected proportion of mistakes, and it is often the right goal when a single false claim can trigger a costly follow-up experiment, regulatory conclusion, or scientific headline.

A second theme is that FWER procedures are easiest to understand when alpha is treated as a scarce resource. Bonferroni divides the resource before seeing the data. Holm recovers some efficiency by allowing earlier, smaller p-values to clear the way for larger thresholds later. Closure gives the theorem that makes such shortcuts legitimate. Graphical procedures make the same resource accounting explicit, which is why they are useful in confirmatory studies whose testing order is part of the design rather than an afterthought.

1. Why Strong Control Is Needed

Route: Core

Consider a platform trial with one primary endpoint and several secondary endpoints. If the primary endpoint succeeds, the study team may want to test secondary endpoints. If one secondary endpoint succeeds, the team may want to test a related subgroup. This is not an unstructured screen: the order and priorities are part of the scientific design. Still, every reported rejection can be wrong, and the sponsor or reader wants a guarantee on the whole family of claims.

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 m_0 for the number of true null hypotheses 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:

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

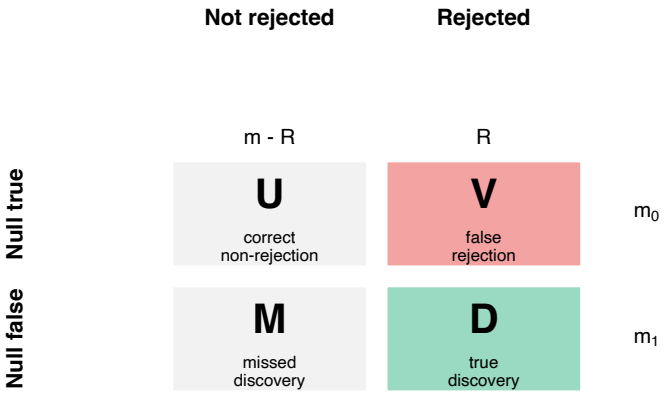

FIGURE 4.1. Multiple-testing decision table. Rows indicate whether the null hypothesis is true or false; columns indicate the procedure's decision. Among true nulls, U are correctly not rejected and V are falsely rejected; among false nulls, M are missed discoveries and D are true discoveries. Row totals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

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

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 familywise error rate is

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

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

Definition 4.1: Weak and strong FWER control

A procedure has weak FWER control at level α 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 α 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 α.

The distinction is not pedantic. Suppose a procedure first tests the global null at level α, 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 α.

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 $p_1, \ldots, p_m$ be valid p-values for hypotheses $H_1, \ldots, H_m$. Let $\mathcal{I}_0$ denote the unknown set of true null indices.

Definition 4.2: Super-uniform p-value

A p-value p_i is valid, or super-uniform, for H_i if, whenever H_i is true,

$$\mathbb{P}(p_i \leq t) \leq t \qquad \text{for every } 0 \leq t \leq 1.$$

If equality holds for every t, the p-value is exactly uniform. Most FWER arguments in this chapter require only marginal super-uniformity for each true-null p-value; dependence assumptions enter only for procedures, such as Sidak, Hochberg, and Hommel, that use sharper information than a union bound.

The Bonferroni procedure rejects

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

Proposition 4.3: Bonferroni strong FWER control

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

Proof of Proposition 4.3

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

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

□

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 α. 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

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

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

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

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

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

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.

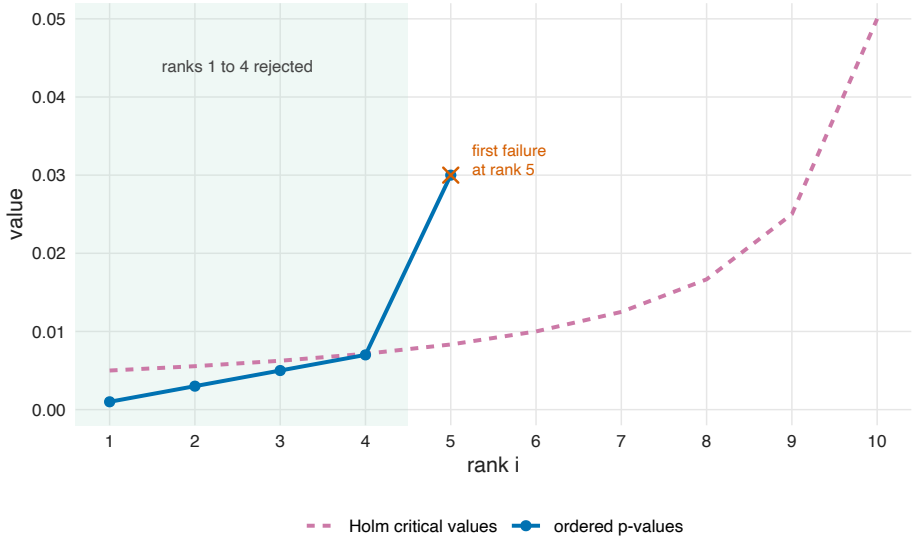

FIGURE 4.2. Holm's step-down rule for $m = 10$ and $\alpha = 0.05$. The solid blue curve shows the ordered p-values through the first failure, at rank 5 (red $\times$); the shaded ranks 1–4 are rejected, and later p-values are not inspected. The dashed purple curve shows all Holm critical values $\alpha/(m-i+1)$,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

Theorem 4.4: Holm strong FWER control

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

Proof of Theorem 4.4

Let $m_0 = |\mathcal{I}_0|$. 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 $H_{(1)}, \ldots, H_{(i)}$.

Figure 4.3 isolates the consequence of scan direction. Holm must certify every comparison from the left, 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.

Hochberg can reject more hypotheses than Holm, but it is not an arbitrary-dependence procedure. Its validity relies on a Simes-type condition. This is the first place where Chapter 3 enters the multiple-testing story. Recall that the Simes global test rejects $\bigcap_{i=1}^{m} H_i$ at level α when

$$p_{(i)} \le \frac{i\alpha}{m} \quad \text{for at least one } i.$$

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.

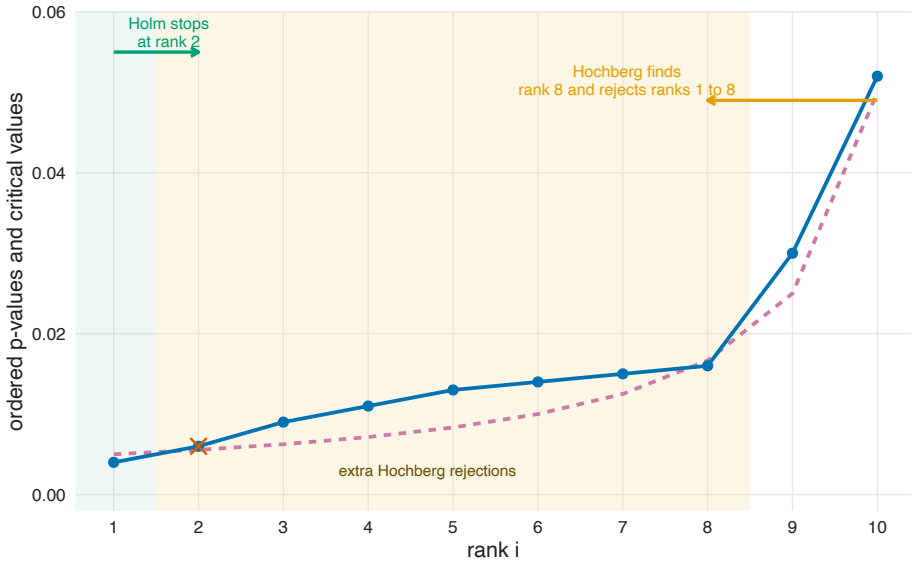

FIGURE 4.3. Step-down and step-up scans using the same critical curve. Holm starts from the smallest p-value and stops at the first failure, here at rank 2. Hochberg starts from the largest p-value and looks backward for the first success, here at rank 8, so it rejects ranks 1 through 8. The illustration 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.

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

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

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-α Simes global test rejects. Under independence, or under positive regression dependence on a subset (PRDS; see 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 α. Weak control is not strong control: when some nulls are false, BH may reject true nulls with probability larger than α. 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 $I \subseteq \{1, \dots, m\}$, define the intersection hypothesis

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

A closed testing procedure specifies a level-α local test for every intersection H_I. It rejects an elementary hypothesis H_i if and only if every intersection hypothesis 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.

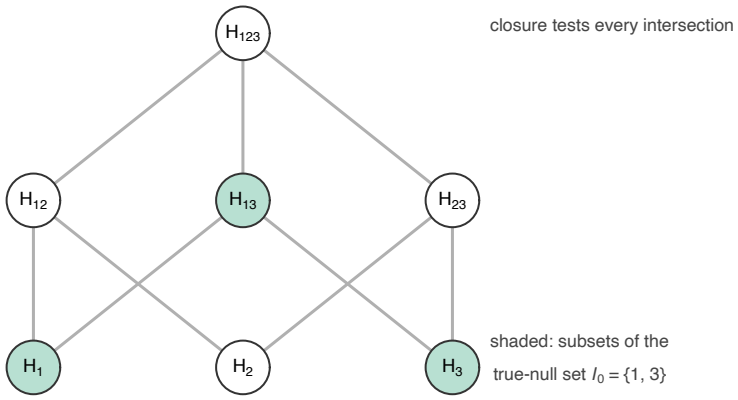

FIGURE 4.4. Closure lattice for three hypotheses. To reject H_1, closed testing requires rejection of every intersection containing 1: $H_1, H_{12}, H_{13}, H_{123}$. In the example shading, the true-null set is $\mathcal{I}_0 = \{1, 3\}$, so the shaded nodes are exactly the intersection hypotheses H_I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

Theorem 4.5: Closure principle

If every intersection hypothesis H_I is tested by a level-α local test, then the closed testing procedure controls FWER at level α 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 α. Therefore

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

□

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.

Closure is not limited to classical FWER. A modern line of work uses the same idea to produce simultaneous lower bounds on the number of true discoveries, and hence upper bounds on the false discovery proportion, for post hoc selected sets of rejections [185, 186]. Recent extensions replace p-value local tests by e-value local tests, including online closed testing and a general closure principle for FDR control [184, 187]. This chapter keeps the focus on FWER, but these extensions explain why closure remains a central organizing principle beyond familywise error.

5. Closed Bonferroni, Closed Simes, and Hommel

Route: Core

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

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

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 α 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.

Each hypothesis H_i receives an initial local alpha budget α_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.

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.

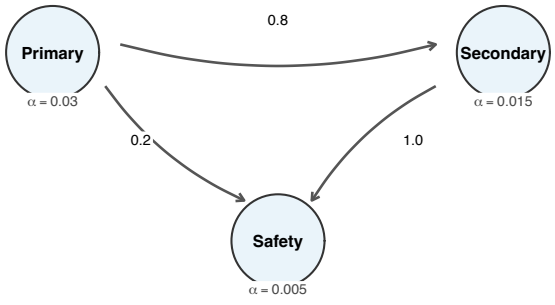

FIGURE 4.5. A graphical testing procedure. Initial alpha is assigned to nodes. After a rejection, the rejected node's local alpha is redistributed along outgoing edges.

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

$$g_{jk} \leftarrow \frac{g_{jk} + g_{ji}g_{ik}}{1 - g_{ji}g_{ij}}, \qquad j \neq k, \quad j, k \neq i,$$

and $g_{jj} = 0$.

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 the omitted reverse edges have 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 failed, 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 α; only if H_1 is rejected does H_2 get tested, again at level α; 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 α; 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 α_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: an early failure lowers the later testing levels, but it does not necessarily end the entire testing plan.

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 for the global intersection:

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

Since $\sum_i \alpha_i \le \alpha$, this local intersection test has level at most α. 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.

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

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

so its current local level is

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

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.

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 formula requires $g_{ji}g_{ij} < 1$. Since both weights lie in $[0, 1]$, this fails only in the boundary case $g_{ji} = g_{ij} = 1$, a two-cycle that sends all alpha back and forth without escape and should be excluded from the design. The update and its validity are derived by Bretz et al. [30].

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 α. 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

Graphical procedures are not valid because the graph is intuitive. They are valid because they correspond to closed tests based on weighted Bonferroni local 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

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

By the union bound, this is a level-α 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 $i \in J$,

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

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. In the resulting closed procedure, rejection of an intersection therefore points to at least one elementary hypothesis that can be rejected. Consonance is what lets a closed procedure behave like a sequential rejection rule rather than stopping with only a rejected global intersection.

The closed weighted Bonferroni algorithm can be read as a sequential shortcut. At any stage, let I be the set of hypotheses not yet removed. Test each remaining H_i, $i \in I$, at local level $\alpha w_i(I)$. If at least one hypothesis crosses its local weighted Bonferroni threshold, reject one such hypothesis, remove it from I, update the weights for the smaller intersection, and continue. If no remaining hypothesis crosses its local weighted Bonferroni threshold, stop. Monotonicity and consonance are the properties that make this shortcut agree with the full closed-testing rule rather than merely resemble it.

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

The practical tradeoff among FWER procedures is conservative validity versus power. Figure 4.6 compares Bonferroni, Holm, and Hochberg in a one-sided Gaussian simulation with many near-boundary nonnull effects and strong positive equicorrelation. Holm dominates Bonferroni because it recycles unused alpha through the ordering. Hochberg can gain additional power when its dependence assumptions are appropriate; the third panel isolates this increment over Holm directly. The simulation is deliberately chosen to make the step-up advantage visible. Read the error panel first: any method above the dashed line is invalid for that setting, regardless of its power in the companion panel.

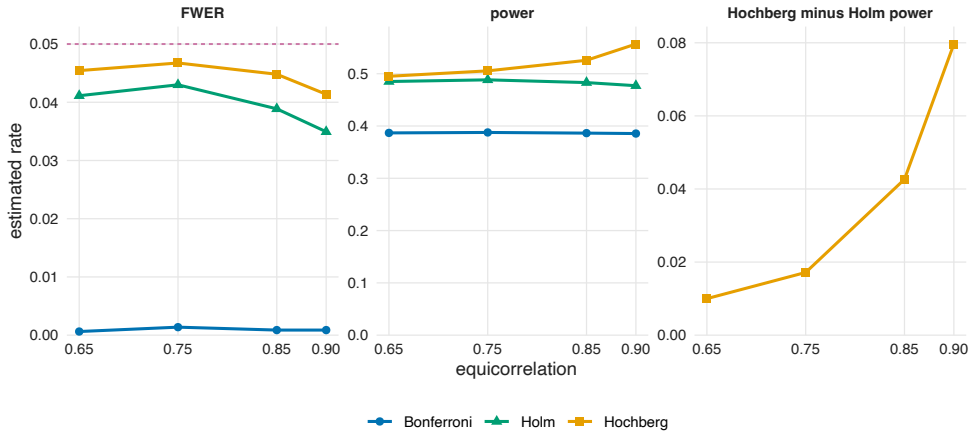

FIGURE 4.6. Monte Carlo FWER, average power, and Hochberg's power gain over Holm for Bonferroni, Holm, and Hochberg in one-sided Gaussian tests with $m = 100$, ninety-eight $N(3, 1)$ nonnull coordinates, two $N(0, 1)$ null coordinates, and equicorrelations from 0.65 to 0.9. Each correlation uses 16,000 Monte Carlo replications at $\alpha = 0.05$. The dashed line in the FWER panel marks 0.05, and the three panels use separate vertical scales. The right panel isolates the Holm–Hochberg difference, which is most visible in this high-dimensional, highly correlated near-boundary configuration.

9. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Audit marginal p-value validity, family membership, weights, and every graphical transition before unblinding.
Not identified from these data	No observed-data diagnostic can prove that the family was not informally changed after seeing results.
Sensitivity analysis	Report results under Holm/Bonferroni and under any sharper dependence-based procedure.
Conservative fallback	Use Holm or closed Bonferroni with prespecified weights.

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 are valid when their alpha recycling corresponds to a closed family of weighted Bonferroni local tests. A graph is a design interface, not a proof by itself.

10. Bibliographic Notes

Route: Core

Bonferroni and Sidak are classical. Holm's step-down procedure is due to Holm [79]. The closure principle is due to Marcus et al. [119]. Hochberg's step-up procedure was introduced by Hochberg [77], and the closed Simes shortcut leading to Hommel's procedure is due to Hommel [80]. Simes' test, used inside closed Simes procedures, is from Simes [146]. Modern graphical testing procedures and gatekeeping formulations are described by Bretz et al. [30] and Dmitrienko et al. [45].

Closure has also become a general tool for post hoc error statements beyond FWER. The exploratory-research formulation of Goeman and Solari [185] uses closed testing to give simultaneous true-discovery guarantees for sets chosen after seeing the data. Goeman et al. [186] show that closed testing is essentially unavoidable for admissible control of false-discovery proportion tail probabilities. More recent e-value work extends the same logic to online testing [184] and to broad classes of expected-loss error rates, including FDR [187].

The regulatory side of multiple testing in clinical trials is shaped by a small set of official guidance documents. The ICH E9 statistical-principles guideline [87] and its E9(R1) addendum on estimands [88] set the language for primary and secondary endpoints and the inferential targets they certify. The EMA guideline on multiplicity issues [54] and the FDA guidance on multiple endpoints in clinical trials [161] 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 $\mathrm{FDP} = V/(R \vee 1)$, and determine whether a familywise error occurred.

Exercise 4.14: Bonferroni strong control

Prove Bonferroni strong FWER control under arbitrary dependence using only super-uniformity of the true-null p-values.

Exercise 4.15: Weak is not strong

Construct a procedure that controls FWER under the global null but fails to control FWER when one null is false. Compute the FWER in your example.

Exercise 4.16: Nonconsonance numerically

Verify the Fisher-combination nonconsonance example in the text numerically: compute $T_{12} = -2\{\log(p_1) + \log(p_2)\}$ for $p_1 = p_2 = 0.07$, compare it with $\chi^2_{4,0.95}$, and compute the Fisher combined p-value. Confirm that the intersection test rejects at level 0.05, while neither singleton test $p_i \leq 0.05$ rejects.

Intermediate.

Exercise 4.17: Holm by direct proof

Prove Holm's strong FWER control by bounding the rank of the smallest true-null p-value among the rejected hypotheses, mirroring the rank argument sketched in Theorem 4.4. Fill in any step that the chapter compressed.

Exercise 4.18: Closed Bonferroni

Show that closing Bonferroni local tests gives exactly Holm's procedure. Verify both directions: closed Bonferroni rejects every hypothesis Holm rejects, and rejects nothing Holm fails to reject.

Exercise 4.19: 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 closed Bonferroni at level 0.05, and identify the rejection set of the shortcut Holm procedure. Check that the two rejection sets agree at the elementary-hypothesis level.

Exercise 4.20: Graph update

Consider three hypotheses with initial alpha budgets

$$(0.03,\ 0.015,\ 0.005)$$

and transition matrix

$$G = \begin{pmatrix} 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Trace the graphical procedure by hand for p-values $(0.02, 0.03, 0.018)$: determine which hypotheses are rejected and in what order, applying the graph-update formula after each rejection. In particular, show that after rejecting the first hypothesis, the updated transition weight from the second hypothesis to the third remains one.

Computational.

Exercise 4.21: 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 both FWER and average power, since a method whose FWER exceeds α is not a valid comparison even if its power is larger.

Exercise 4.22: Platform-trial gatekeeping

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. Simulate FWER under the global null and under at least two partial-null configurations, and simulate power under one realistic alternative.

Advanced.

Exercise 4.23: Hochberg and Simes

Show that Hochberg's step-up rejections are contained in the closed Simes rejections, and state clearly the dependence condition needed for validity. Identify a small example in which Hochberg rejects a strict superset of what Holm rejects.

Exercise 4.24: 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. State the dependence conditions under which the local Simes tests are valid, and contrast Hommel's rejection set with Holm's on an example where Hommel rejects more hypotheses.

CHAPTER 5

False Discovery Rate and the BH Principle

Familywise error 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. False discovery rate is the standard error rate for that regime.

The central procedure in this chapter is the Benjamini-Hochberg procedure. The right way to understand BH is not as a list of sorted p-value instructions, but as a self-consistent threshold: choose a cutoff only when the estimated fraction of false discoveries below that cutoff is small enough.

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 false discovery proportion is

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

The false discovery rate is its expectation:

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

The convention $R \vee 1$ 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 [67]. BH should therefore be read as an expected-error-rate guarantee, not as a certificate that every selected list has small FDP.

2. The BH Step-Up Rule

Route: Core

Let $p_1, \ldots, p_m$ be p-values and write

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

for the order statistics. The BH procedure at level q finds

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

with $\widehat{k} = 0$ if the set is empty. It 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.

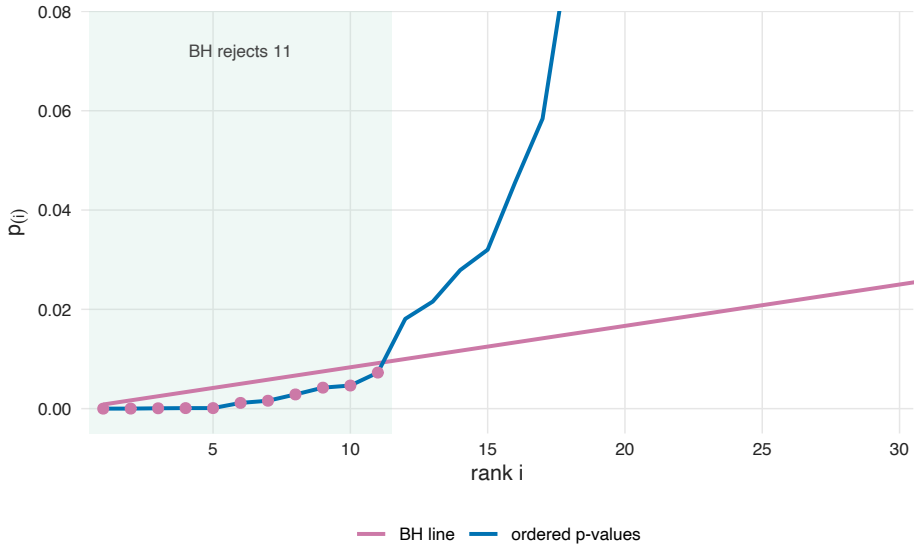

FIGURE 5.1. BH in ordered-p-value form. The procedure finds the largest rank at which $p_{(i)} \le qi/m$ and rejects all p-values up to that rank. The simulated example has $m = 120$, $q = 0.10$, and 16 one-sided Gaussian signals of mean 2.6; the display is a small-p-value zoom of the first 30 ranks.

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 false discovery proportion.

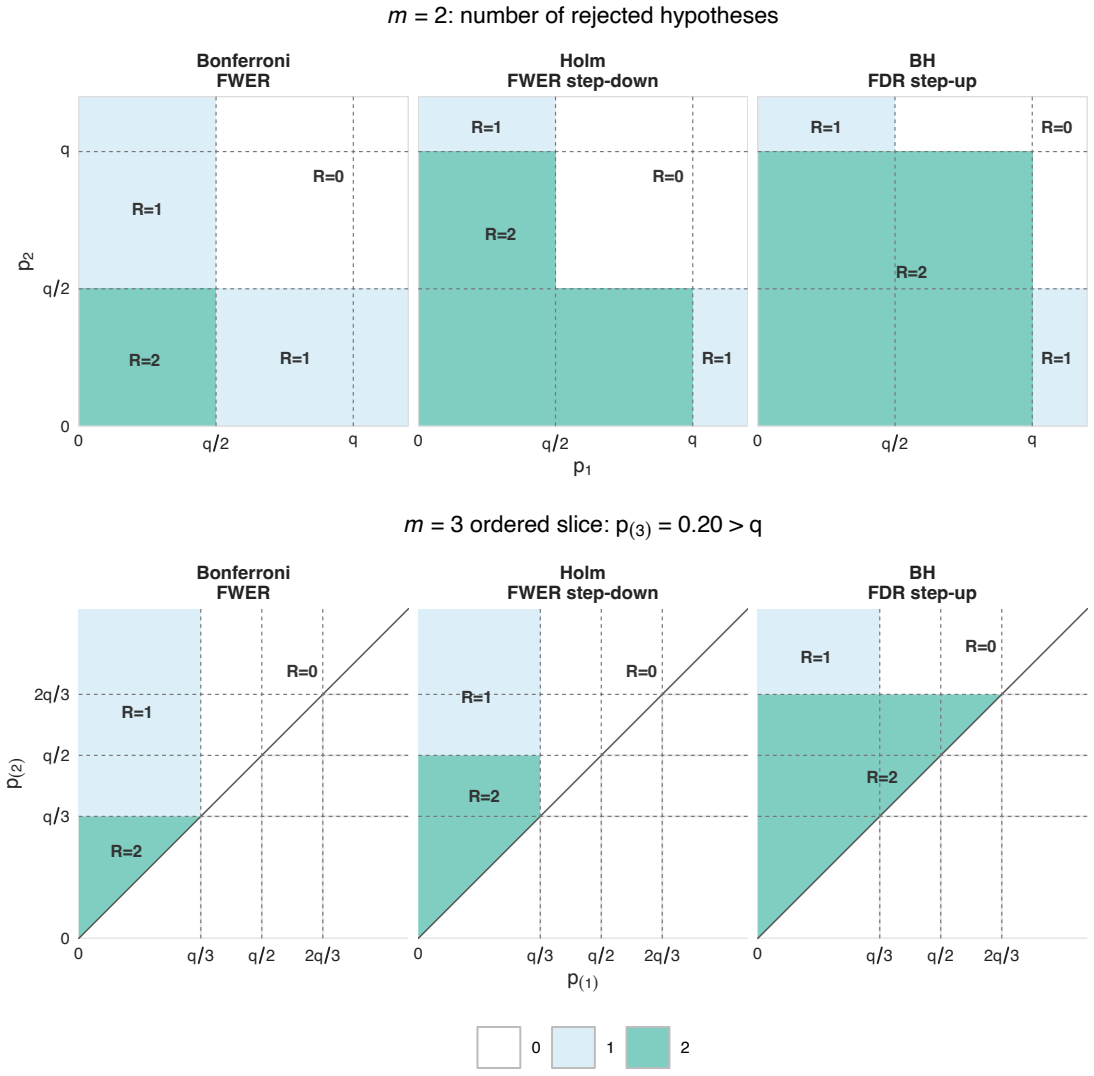

FIGURE 5.2. Rejection regions at $q = 0.10$ for $m = 2$ and $m = 3$. Color gives the number R of rejected hypotheses: white is $R = 0$, blue is $R = 1$, and green is $R = 2$. Each panel recomputes those regions for its own method. The $m = 2$ row shows the full (p_1, p_2) square near the origin. The $m = 3$ row shows the ordered slice $(p_{(1)}, p_{(2)})$ with $p_{(3)} = 0.20 > q$. Bonferroni rejects individual p-values below the first critical value. Holm requires the ordered p-values to pass the FWER step-down critical values. BH uses qi/m, so its two-rejection region is visibly larger in the ordered $m = 3$ slice.

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

$$\widehat{T} = \frac{q\widehat{k}}{m}, \qquad \widehat{k} = \max\left\{ i : \; p_{(i)} \le \frac{qi}{m} \right\},$$

again with $\widehat{k} = 0$ if the set is empty. It rejects $p_i \le \widehat{T}$ when $\widehat{k} > 0$, and rejects nothing when $\widehat{k} = 0$. Equivalently, when at least one p-value is rejected, $\widehat{T}$ is the largest positive threshold satisfying

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

The condition $R(t) > 0$ is important: without it the threshold set would include tiny t's below $p_{(1)}$, even in the no-rejection case. The numerical value $\widehat{T}$ is not generally an observed p-value: 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 $\{i : p_i \le \widehat{T}\}$ 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 empirical CDF 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

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

i.e., the empirical CDF 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 canonical threshold $\widehat{T} = q\widehat{k}/m$ is the right endpoint of the corresponding ECDF step, so it can lie to the right of the last rejected observed p-value $p_{(\widehat{k})}$. For the Benjamini–Yekutieli (BY) procedure, defined fully in Section 10, the analogous endpoint is $\widehat{T}_{\mathrm{BY}} = q\widehat{k}/(mH_m)$, where $H_m = \sum_{j=1}^{m} j^{-1}$ is the mth harmonic number. In both cases the endpoint and the observed $p_{(\widehat{k})}$ give the same rejection set whenever no observed p-values fall between them, but they are different numerical thresholds. 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 FDR 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,

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

The running minimum over $j \ge i$ is essential: adjusted p-values must be monotone in the ordered ranks. The value $Q_{(i)}$ is then assigned back as Q_i to the hypothesis that produced $p_{(i)}$.

Example 5.1

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

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

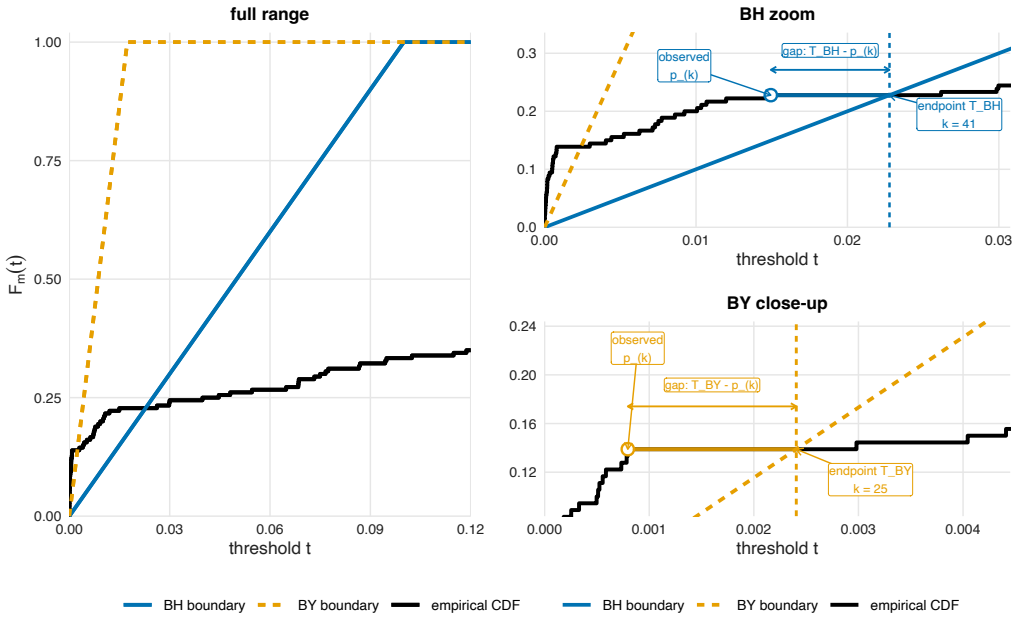

FIGURE 5.3. Empirical-CDF view of BH and BY. The ECDF is a step function. Open circles mark the observed last rejected p-values $p_{(\widehat{k})}$, while dashed vertical lines mark the selected endpoint thresholds $q\widehat{k}/m$ and $q\widehat{k}/(mH_m)$. The zoom panels explicitly mark the slack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

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 adjusted p-values Q_i (or $Q_{(i)}$ in ordered form) are sometimes also called *q-values*. They use m as a conservative upper bound on the number of true nulls. Chapter 6 replaces m by an adaptive estimate $\widehat{\pi}_0 m$, giving a Storey-adjusted version of the same formula; that version is what is conventionally meant by "q-value" in the empirical-Bayes literature.

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 false discovery proportion. 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 threshold, not pointwise domination of the realized curve.

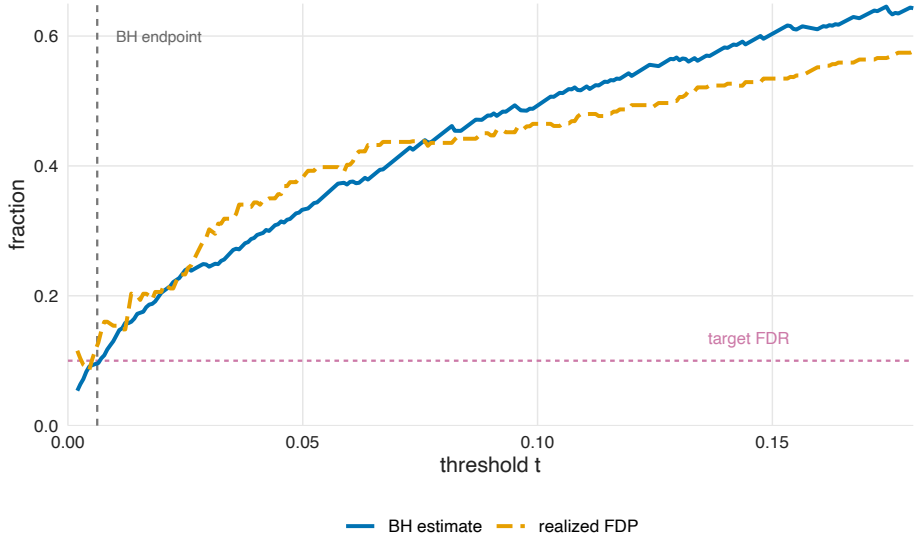

FIGURE 5.4. A single simulated path of the realized FDP and the BH estimator $mt/R(t)$. The estimator need not dominate the realized FDP pointwise; the BH theorem is an expectation statement at the selected threshold. The gray vertical line is the selected BH endpoint. Here $m = 700$, $q = 0.10$, and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

5. The Independent-Null Proof

Route: Core

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 sth ordered p-value after lowering p_i is the same as the original sth 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 $R_i = R$. In particular, $R \vee 1 = R_i$ on this event, so

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

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 $\{p_j : j \ne i\}$, the leave-one-out count R_i is fixed and $qR_i/m \in [0,1]$. Super-uniformity gives

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

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

This proof is the template for much of modern FDR theory. The important features are self-consistency of the rejection threshold and a leave-one-out argument that separates one true-null p-value from the random threshold.

6. Z-Score Thresholds

Route: Core

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

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

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. Thus

$$\widehat{k} = \max\{k : p_{(k)} \leq qk/m\},$$

with no rejections when the set is empty. If $\widehat{k} > 0$, set

$$\widehat{s} = q\widehat{k}/m, \qquad \widehat{T} = \Phi^{-1}(1 - \widehat{s}/2),$$

and reject $|Z_i| \geq \widehat{T}$. If $\widehat{k} = 0$, no z-score threshold is reported and the rejection set is empty.

Equivalently, when at least one BH rejection occurs, $\widehat{T}$ is the smallest observed-relevant t at which the z-scale FDP estimate is controlled:

$$\widehat{T} = \inf\left\{t \geq 0 : \frac{2m\{1 - \Phi(t)\}}{R(t) \vee 1} \leq q\right\},$$

but the no-rejection case is not an empty search over all $t \geq 0$: as $t \to \infty$, the numerator tends to zero. The no-rejection case is simply $\widehat{k} = 0$ in the p-value BH rule, equivalently no observed candidate threshold satisfies the BH inequality. Theorem 5.2 therefore gives FDR $\leq q$ under independent null z-scores, with no additional calculation 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 $w_i \geq 0$ 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.

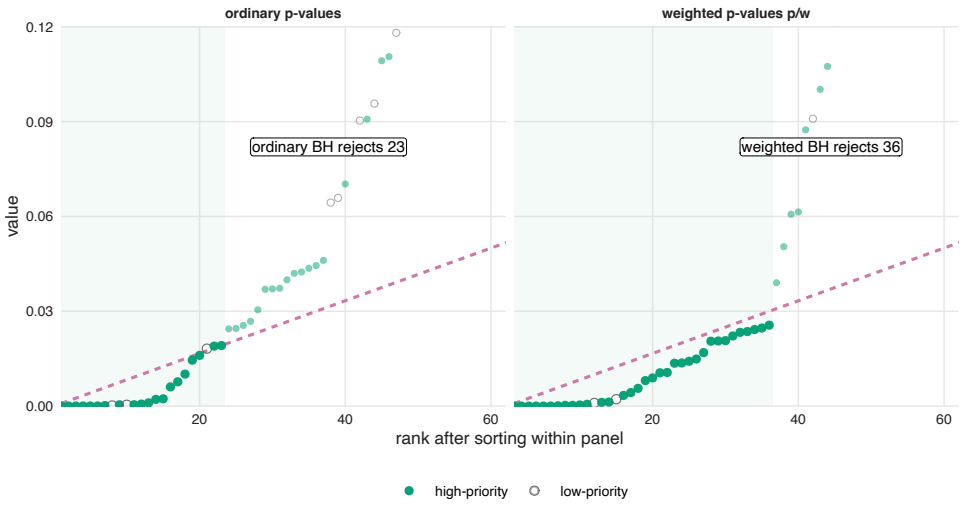

FIGURE 5.5. Weighted BH as BH on transformed p-values p_i/w_i. High-priority hypotheses receive larger weights, so their transformed p-values move downward relative to low-priority hypotheses; in this simulated example the weighted rule rejects visibly more hypotheses than ordinary BH. Here $m = 120$,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

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: IHW

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. The ideal covariate is informative about power or the prior chance of being nonnull, but under a true null it does not use the focal testing noise that produced p_i.

Weighted BH explains how to use fixed side information. Independent hypothesis weighting (IHW) asks the next question: can the weights themselves be learned from the collection of covariates and p-values? The answer is yes when the learning is arranged so that the weight assigned to a true null is not learned from that null p-value.

Ignatiadis et al. [83] implement this idea with fold splitting. The clean cross-fitted version below is a sufficient recipe for the leave-one-out proof. Under mutual independence, a weight for hypothesis i may depend on other p-values, even other p-values in the same fold, because those p-values are independent of a true-null p_i. What the finite-sample proof cannot allow is for the final weight assigned to i, or a normalizing constant that changes it, to depend on p_i itself. A simplified cross-fitted version that enforces this condition is:

(1) Partition the hypotheses into K folds without using the focal p-values; in applications the folds are usually balanced across covariate bins.

(2) For each fold k, use hypotheses outside fold k to learn a nonnegative weight function $\widehat{w}_{-k}(x)$ of the covariate. The training folds may use both their covariates and p-values to choose this function. IHW commonly learns piecewise-constant weights over covariate bins, with regularization to avoid unstable choices.

(3) For each held-out hypothesis $i \in F_k$, set a raw weight $\widetilde{w}_i = \widehat{w}_{-k}(X_i)$. This evaluates the learned function on the held-out covariate X_i, not on the held-out p-value p_i; p_i is saved for the final weighted BH step. After all folds have received raw weights, normalize globally so that

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

This normalization is part of the weight-construction step; the final normalized weights must still satisfy the leave-one-out exogeneity condition stated below.

(4) Run one weighted BH procedure on the transformed p-values p_i/w_i, with $w_i = 0$ interpreted as making hypothesis i unrejectable.

The optimization used to choose the bin weights is not the main point here. The main point is the exogeneity condition. Power considerations naturally push us to use as much p-value information as possible when learning weights. With mutually independent p-values, the weight for i can depend on $\{p_j : j \neq i\}$, including p-values from the same fold, without depending on p_i. Thus a more data-efficient implementation can be valid if it is effectively leave-one-out for every focal null. By contrast, fitting one fold-level weight rule using all p-values in F_k and assigning it back to every member of F_k usually makes w_i depend on p_i, and then this simple finite-sample proof no longer applies without an additional masking, monotonicity, or stability argument.

Proposition 5.4: Leave-one-out weighting validity

Suppose the p-values are mutually independent, each true-null p-value is uniform, and the final nonnegative weight vector $W = (w_1, \ldots, w_m)$ satisfies $\sum_i w_i = m$. Assume also that, for every true null i, the final weight vector used by the BH step can be conditioned on without using p_i; equivalently, after conditioning on the information that produced W and on the other p-values, p_i remains uniform and W is fixed. Then weighted BH applied to p_i/w_i controls FDR at level q.

Proof of Proposition 5.4

Condition on the learned weights and on all p-values except the focal true-null p-value p_i. Under the assumption above, w_i is fixed in this conditional calculation and p_i remains uniform. Let R_i be the rejection count after setting $p_i = 0$, using the same weights. On the event that i is rejected, the weighted BH leave-one-out identity gives $R = R_i$ and

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

Therefore

$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\middle| \text{ other data}\right] \leq \mathbb{E}\left[\frac{\mathbf{1}\{p_i \leq q w_i R_i/m\}}{R_i}\middle| \text{ other data}\right] \leq \frac{q w_i}{m}.$$

If $qw_iR_i/m > 1$, the same bound is trivial because then $1/R_i \le qw_i/m$. Summing over true nulls gives

$$\mathrm{FDR} \le \sum_{i\in\mathcal{I}_0} \frac{qw_i}{m} \le q,$$

since $\sum_i w_i = m$. Averaging over the data used to learn the weights completes the argument. □

This proposition is deliberately cleaner than the full IHW theorem. It captures the reason fold splitting is needed, while the actual IHW analysis also verifies that the fold construction, regularization, and normalization preserve the required leave-one-out structure. If the same p-values are used both to choose their own weights and to test themselves, small null p-values can receive large weights precisely because they are small, and the weighted-BH charge calculation no longer applies.

More adaptive covariate-threshold methods require additional ideas. In particular, procedures such as AdaPT use local-FDR-style models to propose covariate-dependent threshold surfaces, and masking or mirror symmetry to keep the adaptive search valid. Those ingredients are introduced after 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. Let $\psi_i : [0,1] \to [0,\infty)$ be a coordinate-specific strictly increasing transformation with $\psi_i(0) = 0$, so that the rejection event $\psi_i(p_i) \le t$ is equivalent to

$$p_i \le \psi_i^{-1}(t) \wedge 1,$$

where the cap at 1 handles inputs t for which the inverse exceeds the unit interval (in which case the inequality holds trivially because $p_i \le 1$). When we write the rejection threshold below we always interpret $\psi_i^{-1}(t)$ capped at 1 without further comment. Let $g(t)$ be an increasing boundary function with $g(0) = 0$, and define

$$R(t) = \sum_{i=1}^{m} \mathbf{1}\{p_i \le \psi_i^{-1}(t)\}.$$

To avoid threshold-attainment distractions, state the template on a deterministic finite grid $\mathcal{T} \subset (0,1]$. Continuous-threshold versions require the same proof plus one-sided continuity or a largest-attained-threshold convention. The generalized BH threshold is

$$\widehat{T} = \max\left(\{0\} \cup \left\{t \in \mathcal{T} : \frac{mg(t)}{R(t) \vee 1} \le q\right\}\right).$$

The procedure rejects

$$p_i \le \psi_i^{-1}(\widehat{T}).$$

Proposition 5.5: Generalized BH bound

Assume the true-null p-values are independent uniforms and are independent of the false-null p-values. Define

$$C = \frac{1}{m} \sum_{i\in\mathcal{I}_0} \sup_{t\in\mathcal{T}} \frac{\psi_i^{-1}(t)}{g(t)}.$$

Assume additionally the following leave-one-out stability property: for every true null i, if i is rejected by the original procedure and $(R_i, \widehat{T}_i)$ are the rejection count and threshold after setting $p_i = 0$, then $R_i = R$ and $p_i \le \psi_i^{-1}(\widehat{T}_i)$. Ordinary BH and fixed weighted BH have this property because they are ordinary BH procedures applied to transformed p-values. Then the generalized BH procedure satisfies

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

Proof of Proposition 5.5

The proof is the same leave-one-out argument as for BH. For a true null i, set $p_i = 0$ and let R_i and $\widehat{T}_i$ be the resulting rejection count and threshold. On the event that the original procedure rejects i, leave-one-out stability ensures that the attained grid threshold $\widehat{T}_i$ is a threshold at which the leave-one-out procedure rejects, so the self-consistency relation reads

$$\frac{m g(\widehat{T}_i)}{R_i} \le q.$$

Conditioning on all other p-values makes R_i and $\widehat{T}_i$ fixed, while p_i is uniform. By leave-one-out stability, the contribution of hypothesis i to the FDR is bounded by

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_i \le \psi_i^{-1}(\widehat{T}_i)\}}{R_i} \,\middle|\, \{p_j : j \ne i\}\right] = \frac{\psi_i^{-1}(\widehat{T}_i)}{R_i} \le \frac{q}{m} \cdot \frac{\psi_i^{-1}(\widehat{T}_i)}{g(\widehat{T}_i)} \le \frac{q}{m} \sup_{t \in \mathcal{T}} \frac{\psi_i^{-1}(t)}{g(t)}.$$

Summing over $i \in \mathcal{I}_0$ and using the definition of C gives the result. □

Ordinary BH satisfies the stability condition in Proposition 5.5 because lowering an already rejected p_i to zero only moves that value within the first R ordered positions; the transformed values beyond rank R do not change, so no larger rank can newly cross and $R_i = R$. The canonical cutoff qR/m is therefore also unchanged, which gives the required focal inequality. For fixed weighted BH, apply the same argument to the fixed transformed values p_j/w_j. A rejected hypothesis necessarily has $w_i > 0$, and lowering p_i to zero again moves only its transformed value within the rejected prefix. Hence the transformed cutoff remains qR/m, and rejection gives $p_i \le w_i qR/m = \psi_i^{-1}(\widehat{T}_i)$. If $w_i = 0$, the hypothesis cannot be rejected and the condition is vacuous.

For ordinary BH, $\psi_i^{-1}(t) = t$ and $g(t) = t$, so

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

For weighted BH with $\psi_i(p) = p/w_i$ and weights normalized so that $\sum_i w_i = m$, one has $\psi_i^{-1}(t) = w_i t$; taking $g(t) = t$ gives

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

Thus fixed weighted BH controls FDR at level q. The inequality is strict when larger weights are assigned mostly to nonnull hypotheses, which is why genuinely external prior information can improve power.

The value of this template is conceptual. It shows that BH is one member of a larger family of self-consistent thresholding rules. 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 positive regression dependence on a subset (PRDS), defined in Section 1. 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 $\mathbb{E}[\mathbf{1}\{p_i \le qR_i/m\}/R_i] = q/m$; 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 [20] 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 the true-null p-values are marginally continuous uniforms, 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

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

We reserve η 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

The displayed proof uses exact continuous uniformity of the true-null p-values. For merely super-uniform p-values, one should use the standard reshaping formulation or state the corresponding conservative interval-probability inequalities explicitly. Let R be the number of rejections. Step-up self-consistency says that if hypothesis i is rejected and $R = r$, then $p_i \le a_r$. Hence, for a true null i,

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

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,

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

Taking expectations and using the marginal uniformity of the true-null p-value gives

$$\mathbb{E}\left[\frac{\mathbf{1}\{i \text{ is rejected}\}}{R \vee 1}\right] \leq \sum_{s=1}^{m} \frac{a_s - a_{s-1}}{s}.$$

For BH run at nominal level η, $a_s = \eta s/m$, so

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

Summing over the m_0 true nulls shows that BH at nominal level η has

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

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 $\beta(0) = 0 \leq \beta(1) \leq \cdots \leq \beta(m)$, the same argument gives

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

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 [27].

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 $0 < t \leq 1/2$, define

$$\widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) = \frac{1 + \#\{i : p_i \geq 1 - t\}}{R(t) \vee 1}.$$

The plus one stabilizes the estimate. A mirror procedure chooses from the finite candidate set of positive reflected magnitudes

$$\mathcal{U}_{\mathrm{mir}} = \{\min(p_i, 1 - p_i) : 1 \leq i \leq m, \min(p_i, 1 - p_i) > 0\}.$$

Because every p-value lies in $[0, 1]$, each $\min(p_i, 1 - p_i)$ 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

$$\widehat{T}_{\mathrm{mir}} = \max\left\{t \in \mathcal{U}_{\mathrm{mir}} : \widehat{\mathrm{FDP}}_{\mathrm{mir}}(t) \leq q\right\},$$

with $\widehat{T}_{\mathrm{mir}} = 0$ if no candidate passes, and rejects $p_i \leq \widehat{T}_{\mathrm{mir}}$.

This rule is attractive because it estimates the number of false discoveries from the data rather than using the conservative upper bound mt. Under suitable joint sign-symmetry and

stopping-time conditions – for example, conditionally independent null signs given magnitudes and nonnull information, with the threshold chosen by a valid sign-revealing filtration – the procedure satisfies

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

The argument is a martingale stopping-time argument on the sign-flip variables $\mathbf{1}\{p_i \le 1/2\} - \mathbf{1}\{p_i > 1/2\}$; the same sign-symmetry logic appears in the knockoff proofs of Chapter 9, so we defer the detailed argument to Barber and Candès [8]. Three caveats matter in practice: the orientation of the mirror count is essential (counting $p_i \ge 1 - t$ rather than $p_i \le 1 - t$), the plus-one stabilizer is what gives 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.

Figure 5.6 uses nominal level $q = 0.10$. BH rejects where the conservative estimate $mt/R(t)$ first reaches this level; in the example this gives 53 rejections. The mirror rule uses the observed upper tail instead. Because the upper-tail mirror count is small at the selected threshold, $(1 + \#\{p_i \ge 1 - t\})/(R(t) \vee 1)$ stays near 0.10 until a larger threshold and the rule rejects 84 hypotheses. This is a data-adaptive power gain in this symmetric example, not a guarantee that mirror always dominates BH.

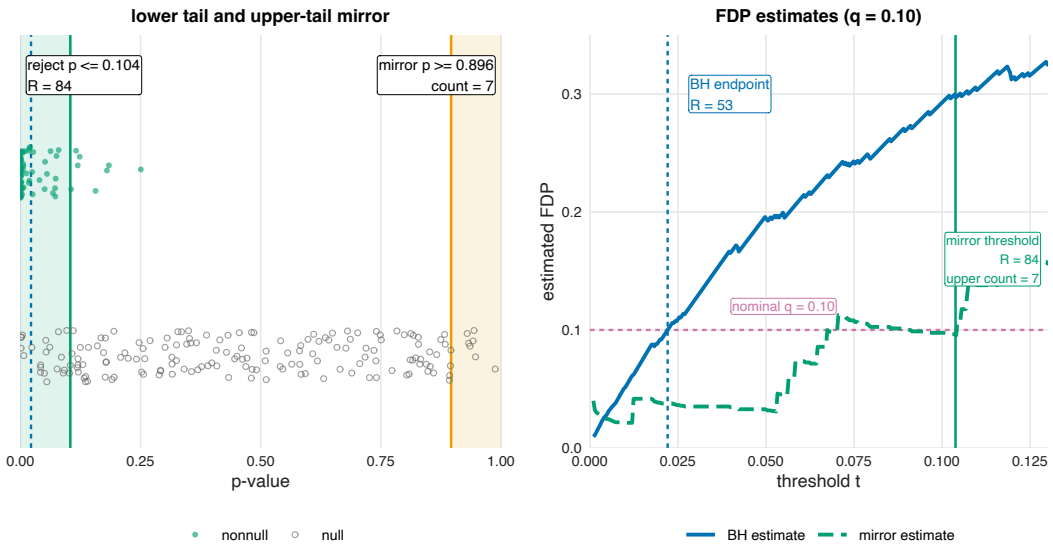

FIGURE 5.6. Mirror estimation in one simulated example at nominal level $q = 0.10$. The lower tail supplies candidate discoveries, while the upper tail $p_i \ge 1 - t$ estimates the null contribution in the lower tail. Compared with BH's conservative estimate $mt/R(t)$, the mirror estimate can allow 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.

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.

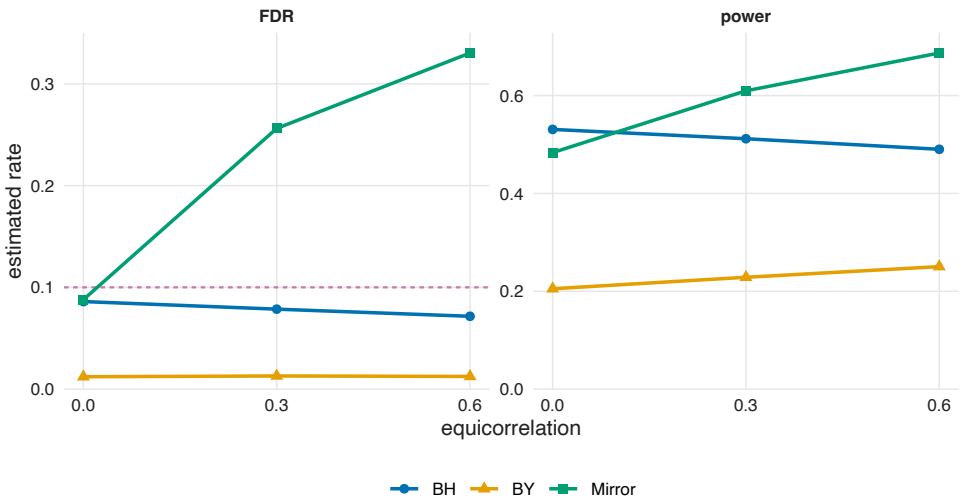

FIGURE 5.7. Simulation comparing BH, BY, and a mirror estimator. BY is robust but conservative. The mirror rule can gain power, but its validity is more sensitive to dependence and symmetry assumptions. The design uses $m = 300$ one-sided Gaussian tests, 45 signals of mean 2.45, $q = 0.10$, 1,000 replications, and equicorrelations 0, 0.3, and 0.6.

12. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Check marginal null calibration, duplicated tests, weight normalization, and dependence simulations tied to the design.
Not identified from these data	PRDS is a property of the joint null law and is rarely testable from a single p-value vector.
Sensitivity analysis	Compare BH with BY and vary filtering or weighting choices fixed independently of the tested p-values.
Conservative fallback	Use BY, or use BH only after a design-based independence/PRDS justification.

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 noise. Mirror estimators use symmetry and require more care than the formula alone suggests.

13. Bibliographic Notes

Route: Core

The BH procedure was introduced by Benjamini and Hochberg [19]. The BY arbitrary-dependence correction is due to Benjamini and Yekutieli [20]. Generalized and reshaped FDR procedures are discussed by Blanchard and Roquain [27]. The mirror estimator and its sign-flip finite-sample argument are due to Barber and Candès [8], 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 [181]. The empirical-process, adaptive, and empirical-Bayes extensions are deferred to Chapter 6.

Independent hypothesis weighting is introduced by Ignatiadis et al. [83]; 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 [9] and Yekutieli [178]; 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 (ε, δ)-differential privacy with a multiplicative inflation factor; see Dwork et al. [51] 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 ≤ 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

$$0.004,\ 0.011,\ 0.018,\ 0.042,\ 0.067,\ 0.21,\ 0.33,\ 0.46,\ 0.71,\ 0.88,$$

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.

Exercise 5.11: Threshold equivalence

Prove that the ordered-p-value definition of BH is equivalent to the threshold definition

$$\widehat{k} = \max\{i : p_{(i)} \leq qi/m\}, \qquad \widehat{T} = q\widehat{k}/m,$$

with $\widehat{k} = 0$ if the set is empty. Show also that when $\widehat{k} > 0$, this $\widehat{T}$ is the largest threshold with $R(t) > 0$ and $mt/R(t) \leq q$, while $\widehat{k} = 0$ gives no rejections.

Intermediate.

Exercise 5.12: Leave-one-out proof

Fill in the details of Theorem 5.2. In particular, prove the self-consistency relation between rejection of a true null and the leave-one-out rejection count R_i.

Exercise 5.13: Z-score BH

Starting from two-sided p-values $p_i = 2\{1 - \Phi(|Z_i|)\}$, derive the z-score threshold used when BH makes at least one rejection:

$$\widehat{T} = \inf\left\{t \geq 0 : \frac{2m\{1 - \Phi(t)\}}{R(t) \vee 1} \leq q\right\}.$$

Explain why the infimum is the right operator (rather than supremum) when larger $|Z_i|$ is stronger evidence. Also explain why the no-rejection case is $\widehat{k} = 0$ in the p-value BH rule, not an empty search over all $t \geq 0$.

Exercise 5.14: Weighted BH

Show that fixed weighted BH is ordinary BH applied to p_i/w_i. State the normalization condition on the weights and explain why data-learned weights are not automatically valid.

Exercise 5.15: IHW validity through folds

Suppose an IHW-style procedure splits the hypotheses into K folds. For each fold k, a weight function $\widehat{w}_{-k}(x)$ is learned using only $\{(p_j, X_j) : j \notin F_k\}$, assigned to held-out hypotheses by $w_i = \widehat{w}_{-k}(X_i)$ for $i \in F_k$, and normalized so that the final weights satisfy $\sum_i w_i = m$. Assume the normalization is leave-one-out-safe: when analyzing a focal true null i, the final weight vector used by the BH step is fixed after conditioning on data that exclude p_i. Condition on that information and on all p-values except p_i. Explain why p_i remains uniform and the weighted-BH leave-one-out proof applies. Under mutual independence, why would it also be safe for the weight of i to use p-values in F_k other than p_i? What can go wrong if p_i itself is used to tune the weight assigned to i, or if a global normalizing constant reintroduces dependence on p_i?

Exercise 5.16: 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 5.17: Pathwise FDP

Reproduce Figure 5.4. Explain why the BH estimator need not dominate the realized FDP for every threshold.

Exercise 5.18: Simulation study

Simulate BH, BY, and the mirror procedure under independent and equicorrelated Gaussian z-scores. Report FDR, power, and the distribution of realized FDP. 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 values such as $200, 1000, 5000$. 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 explain the trend using Theorem 6.8. The

figure-generation function `fig04_bh_by_bc_sim()` in `scripts/make_figures.R` gives one reference implementation.

Advanced.

Exercise 5.19: Mirror orientation

For the mirror estimator, explain why the numerator counts $\#\{p_i \geq 1-t\}$ rather than $\#\{p_i \leq 1-t\}$. What is the role of the plus-one term? Construct a small example with two independent null p-values showing that omitting the plus-one breaks the finite-sample bound.

Exercise 5.20: Generalized BH

Specialize the generalized threshold template to weighted BH and verify the constant C under independent true-null p-values. Then show that ordinary BH is the case $\psi_i(p) = p$, $g(t) = t$, and that the BY procedure corresponds to a different choice of g (which one?).

CHAPTER 6

Dependence, Adaptive FDR, and Empirical Bayes

Chapter 5 presented BH under independent null p-values and then introduced BY as a robust but conservative arbitrary-dependence correction. In large-scale inference, neither independence nor worst-case dependence is a satisfactory default. Gene-expression measurements are correlated through pathways. Imaging statistics are spatially dependent. Benchmark scores share training data, prompts, and evaluation pipelines. Dependence is part of the problem.

This chapter has two goals. First, it explains when dependence is benign for BH. Positive dependence, formalized through PRDS, often preserves FDR control. Second, it explains how the null fraction and posterior null probabilities can be estimated. Storey's estimator, q-values, and local FDR all refine the basic BH idea, but they answer different questions and rely on different assumptions.

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 is a condition that keeps the dependence in a favorable direction.

Because p-values are small when evidence is strong, one must be careful with monotonicity. For two p-value vectors $p = (p_1, \dots, p_m)$ and $p' = (p'_1, \dots, p'_m)$, write $p' \geq p$ when $p'_j \geq p_j$ for every coordinate. A Borel-measurable set $A \subseteq [0,1]^m$ is called increasing if

$$p \in A, \quad p' \geq p \quad \Longrightarrow \quad p' \in A.$$

The set A is the measurable geometric object; for a random p-value vector P, the corresponding event is $\{P \in A\}$. Increasing P coordinatewise can therefore only move this event from false to true.

Definition 6.1: PRDS for p-values

A p-value vector $P = (P_1, \dots, P_m)$ is positively regression dependent on a set of true nulls $\mathcal{I}_0$ if, for every Borel-measurable increasing set $A \subseteq [0,1]^m$ and every $i \in \mathcal{I}_0$,

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t)$$

is nondecreasing in t, using a regular conditional distribution.

Figure 6.1 gives the geometry behind the definition in two coordinates: once a point enters the shaded upper set, moving either p-value upward cannot make it leave. PRDS asks that conditioning a true-null coordinate at larger values makes the event $\{P \in A\}$ no less likely for every such set.

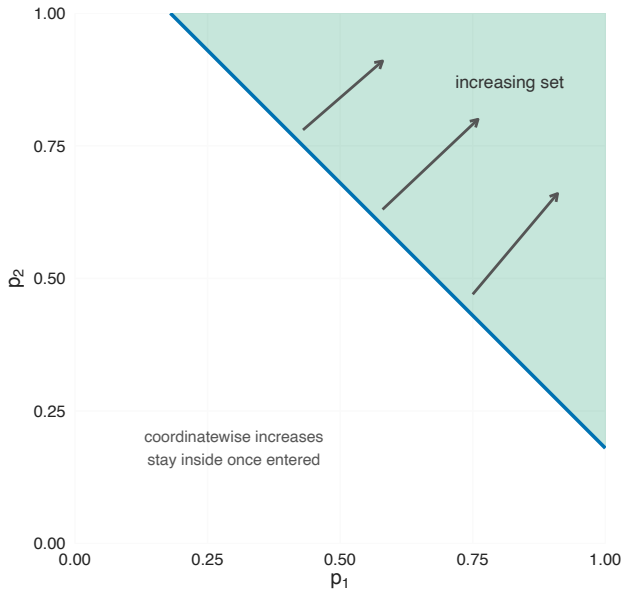

FIGURE 6.1. A schematic increasing set in p-value coordinates. Once a point is inside the shaded set, making coordinates larger keeps it inside. PRDS is a positive-dependence condition stated in terms of the events generated by these increasing sets.

Theorem 6.2: Gaussian sufficient condition for PRDS

Let $Z = (Z_1, \dots, Z_m)$ be multivariate normal with means μ_i and covariance Σ, and let

$$P_i = 1 - \Phi(Z_i)$$

be the right-tail p-value for testing $H_i : \mu_i = 0$ against positive alternatives, so $\mu_i = 0$ for true nulls. If

$$\mathrm{Cov}(Z_i, Z_j) \geq 0 \qquad \text{for every } i \in \mathcal{I}_0,\ j \in \{1, \dots, m\},$$

then the p-value vector $P = (P_1, \dots, P_m)$ is PRDS on $\mathcal{I}_0$ in the sense of Definition above.

Proof of Theorem 6.2

Fix $i \in \mathcal{I}_0$ and let A be a Borel-measurable increasing set in p-value coordinates. Its inverse image in Z-coordinates is the set

$$B = \{z \in \mathbb{R}^m :\ (1 - \Phi(z_1), \dots, 1 - \Phi(z_m)) \in A\}.$$

It is *decreasing*: increasing the Z_j's decreases the P_j's, so a set that is upward closed in p-value coordinates is downward closed in Z-coordinates. If $z' \leq z$ coordinatewise and $z \in B$, then the corresponding p-value vector at z' is at least the one at z, so z' also lies in B.

The conditional distribution of Z_{-i} given $Z_i = z$ is Gaussian with covariance not depending on z and mean

$$\mathbb{E}[Z_{-i} \mid Z_i = z] = \mu_{-i} + \Sigma_{-i,i}\Sigma_{i,i}^{-1}(z - \mu_i).$$

Under the covariance assumption, $\Sigma_{-i,i}$ is entrywise nonnegative, so this conditional mean is coordinatewise nondecreasing in z. To see the monotonicity explicitly, take $z_1 < z_2$ and represent the two conditional vectors as

$$Z_{-i}^{(r)} = m(z_r) + \varepsilon, \qquad r = 1, 2,$$

using the same centered Gaussian noise ε, where $m(z) = \mu_{-i} + \Sigma_{-i,i}\Sigma_{i,i}^{-1}(z - \mu_i)$. Then $m(z_1) \le m(z_2)$ coordinatewise. Since B is decreasing,

$$\mathbf{1}\{(z_2, m(z_2) + \varepsilon) \in B\} \le \mathbf{1}\{(z_1, m(z_1) + \varepsilon) \in B\}$$

after inserting the fixed ith coordinate into the same position in both vectors. Taking expectations of this pointwise inequality with respect to the shared noise gives

$$\begin{aligned}\mathbb{P}(P \in A \mid Z_i = z_2) &= \mathbb{E}_\varepsilon[\mathbf{1}\{(z_2, m(z_2) + \varepsilon) \in B\}] \\ &\le \mathbb{E}_\varepsilon[\mathbf{1}\{(z_1, m(z_1) + \varepsilon) \in B\}] \\ &= \mathbb{P}(P \in A \mid Z_i = z_1).\end{aligned}$$

Here the displayed vectors are understood with the ith coordinate inserted in its original position. Thus $\mathbb{P}(P \in A \mid Z_i = z)$ is nonincreasing in z.
Translating back to p-values reverses the direction once more. The map $P_i = 1 - \Phi(Z_i)$ is strictly decreasing, so conditioning on $P_i = t$ is the same as conditioning on $Z_i = \Phi^{-1}(1 - t)$, and this value of Z_i decreases as t increases. Therefore a conditional probability that is nonincreasing in Z_i becomes nondecreasing in t:

$$t \mapsto \mathbb{P}(P \in A \mid P_i = t)$$

is nondecreasing, which is the PRDS condition for the p-value vector. □

The sign manipulation above is the source of many mistakes in informal descriptions of PRDS. The two coordinate systems are equivalent, but coordinatewise transformations preserve PRDS only when their order is tracked. If each T_i is strictly increasing, transforming P_i to $T_i(P_i)$ preserves the coordinate order. If each T_i is strictly decreasing, the partial order reverses and the PRDS statement must be written in the reversed order. In the theorem, the right-tail map $P_i = 1 - \Phi(Z_i)$ is decreasing, so the nonnegative regression shift in Z-coordinates becomes the required monotonicity in p-value coordinates. Keeping this transformation-order note next to the Gaussian proof makes the two reversals explicit.

Remark 6.3: MTP2, PRDS, and Gaussian signs

Multivariate total positivity of order two (MTP2) is a stronger positive- dependence condition. A density f on a product lattice is MTP2 if

$$f(x \vee y)f(x \wedge y) \ge f(x)f(y) \qquad \text{for all } x, y,$$

where $\vee$ and $\wedge$ are coordinatewise maximum and minimum. MTP2 implies positive association and, in the appropriate coordinate order, PRDS [94, 140]. For a nonsingular Gaussian vector, MTP2 is equivalent to the precision matrix Σ^{-1} having nonpositive off-diagonal entries. This is stronger than pairwise nonnegative covariance in dimensions larger than two. Theorem 6.2 used a direct one-sided Gaussian sufficient condition; it should not be read as an MTP2 characterization.

Lemma 6.4: PRDS conditioning inequality

If P is PRDS on the true-null set $\mathcal{I}_0$, then for every Borel-measurable increasing set $A \subseteq [0,1]^m$ and every $i \in \mathcal{I}_0$,

$$t \mapsto \mathbb{P}(P \in A \mid P_i \le t)$$

is nondecreasing.

Proof of Lemma 6.4

Let $h(s) = \mathbb{P}(P \in A \mid P_i = s)$, which is nondecreasing by PRDS, and let F_i denote the marginal CDF of P_i. Then

$$\mathbb{P}(P \in A \mid P_i \le t) = \frac{1}{F_i(t)} \int_0^t h(s)\, dF_i(s).$$

For $t' > t$,

$$\mathbb{P}(P \in A \mid P_i \le t') = \frac{F_i(t)}{F_i(t')}\mathbb{P}(P \in A \mid P_i \le t) + \frac{F_i(t') - F_i(t)}{F_i(t')}\mathbb{P}(P \in A \mid t < P_i \le t').$$

The right-hand side is a convex combination of $\mathbb{P}(P \in A \mid P_i \le t)$ and an average of h over $(t, t']$. Since h is nondecreasing, the second average is at least the first, so the convex combination is at least $\mathbb{P}(P \in A \mid P_i \le t)$. Therefore $\mathbb{P}(P \in A \mid P_i \le t)$ is nondecreasing in t. □

The PRDS definition says that conditioning a true-null p-value to be less stringent should not make an increasing-set event less likely. We now connect that abstract condition to BH. The key events in its proof are $\{R \le k\}$, where R is the BH rejection count. If $p \le p'$ coordinatewise, every ordered p-value of p' is at least the corresponding ordered p-value of p. A BH crossing that disappears after increasing the p-values cannot be replaced by a new crossing, so $R(p') \le R(p)$. Consequently $\{R \le k\} = \{P \in A_k\}$ for an increasing set A_k, which is exactly the form needed to apply the preceding PRDS conditioning inequality.

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 the BH procedure at level q satisfies

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

Proof of Theorem 6.5

Let R be the number of BH rejections and fix a true null i. It is enough to show

$$\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le qR/m\}}{R \vee 1}\right] \le \frac{q}{m}.$$

Write $F_i(t) = \mathbb{P}(P_i \le t)$. Super-uniformity gives, for every $k = 1, \dots, m$,

$$\frac{F_i(qk/m)}{k} \le \frac{q}{m}.$$

Decompose by the value of R:

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{1}\{P_i \le qR/m\}}{R \vee 1}\right] &= \sum_{k=1}^{m} \frac{1}{k}\mathbb{P}\left(P_i \le \frac{qk}{m}, R = k\right) \\ &= \sum_{k=1}^{m} \frac{F_i(qk/m)}{k}\mathbb{P}\left(R = k \,\middle|\, P_i \le \frac{qk}{m}\right) \\ &\le \frac{q}{m}\sum_{k=1}^{m}\mathbb{P}\left(R = k \,\middle|\, P_i \le \frac{qk}{m}\right).\end{aligned}$$

If $F_i(qk/m) = 0$, the corresponding summand is zero and its conditional probability may be defined arbitrarily. The event $\{R \le k\}$ is increasing in the p-value coordinates: BH compares $p_{(j)}$ with qj/m, so making p-values larger only weakens crossings, hence R can only decrease. Therefore $\{R \le k\}$ is preserved under coordinatewise increase, and Lemma 6.4 applies. Writing each term as a difference of $\{R \le k\}$-probabilities,

$$\begin{aligned}&\sum_{k=1}^{m}\mathbb{P}\left(R = k \,\middle|\, P_i \le \frac{qk}{m}\right) \\ &\quad = \mathbb{P}(R \le m \mid P_i \le q) - \mathbb{P}(R \le 0 \mid P_i \le q/m) \\ &\qquad + \sum_{k=1}^{m-1}\left[\mathbb{P}\left(R \le k \,\middle|\, P_i \le \frac{qk}{m}\right) - \mathbb{P}\left(R \le k \,\middle|\, P_i \le \frac{q(k+1)}{m}\right)\right].\end{aligned}$$

The first term is at most 1. The second is zero: conditional on $P_i \le q/m$, the p-value P_i itself satisfies the smallest BH threshold $p_{(1)} \le q/m$, so the BH procedure rejects at least hypothesis i, giving $R \ge 1$ deterministically. Each bracket in the final sum is nonpositive: by Lemma 6.4, the conditional probability $\mathbb{P}(R \le k \mid P_i \le t)$ is nondecreasing in t, so increasing t from qk/m to $q(k+1)/m$ only increases the right-hand side. Hence the total is at most 1, and summing the bound q/m over the m_0 true nulls proves the claim. □

PRDS is not arbitrary dependence. Negative or adversarial dependence can break the proof. This is why BY remains useful as a safe fallback and why dependence modeling deserves attention.

Remark 6.6: Two-sided Gaussian p-values

For two-sided p-values

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

from a Gaussian vector with arbitrary covariance, the one-sided PRDS argument above no longer applies directly: the map $Z_i \mapsto |Z_i|$ is not coordinatewise monotone. This obstruction leads to two distinct questions.

For general configurations of null and nonnull means, BH need not control the FDR under arbitrary Gaussian dependence. Dobriban [46] constructs a sequence with $m = 100N$ coordinates, of which $96N$ are true nulls and $4N$ are nonnulls, such that BH at level $q = 0.01$ satisfies $\mathrm{FDR}_N > 0.0104$ for all sufficiently large N. In that construction, the two nonnull blocks enlarge the BH rejection count in the same latent-factor states where the conditional null distribution has heavier two-sided tails. Thus the general arbitrary-covariance conjecture is false.

The counterexample is not a global-null example. Under the global null, $V = R$, so $\mathrm{FDR} = \mathbb{P}(R > 0)$, and FDR control at level q is equivalent to the two-sided Gaussian Simes

bound

$$\mathbb{P}\left(\exists k:\ P_{(k)} \le \frac{kq}{m}\right) \le q.$$

The case $m = 2$ is known, and several positive-dependence or MTP2 subclasses are covered by existing comparison arguments. In dimensions $m \ge 3$, an arbitrary Gaussian correlation matrix can have negative-dependence patterns that fall outside PRDS, MTP2, and the usual Gaussian correlation inequality machinery. The counterexample above does not settle this global-null Simes question, which remains open for an arbitrary Gaussian correlation matrix.

2. Why PRDS Does Not Protect Mirror Thresholds

Route: Research frontier

The PRDS theorem above belongs to BH; it is not a generic license for every adaptive FDR threshold. Recall the mirror rule from Section 11 of Chapter 5. For

$$R(u) = \#\{i : P_i \le u\}, \qquad L(u) = \#\{i : P_i \ge 1 - u\},$$

let

$$\mathcal{U} = \{\min(P_i, 1 - P_i) : 1 \le i \le m\} \cap (0, 1/2]$$

and choose the largest $u \in \mathcal{U}$ satisfying

$$\frac{1 + L(u)}{R(u) \vee 1} \le q.$$

The rule rejects $P_i \le u$, with no rejections if no candidate passes.

This is algebraically the knockoff+ threshold applied to $W_i = 1/2 - P_i$. Indeed, with $t = 1/2 - u$,

$$P_i \le u \iff W_i \ge t, \qquad P_i \ge 1 - u \iff W_i \le -t,$$

and a candidate $u = \min(P_i, 1 - P_i)$ corresponds to $t = |W_i|$. Apart from deterministic tie conventions, the two formulas therefore give the same rejection set. This equivalence is algebraic, not a validity theorem. A sufficient condition used by valid knockoff statistics is that, conditional on all magnitudes and all nonnull scores, the signs of the nonzero null scores are mutually independent fair signs. Marginal symmetry does not imply this conditional coordinatewise property [181].

Under the global null every rejection is false, so $\mathrm{FDR} = \mathbb{P}(R > 0)$. A simple mechanism can therefore force failure. If $m > 1/q$ and all W_i's are positive, then at $t = \min_i W_i$ the knockoff+ ratio is $1/m < q$, so the procedure rejects. The next example makes this common positive movement compatible with exactly uniform PRDS p-values and a positive joint density.

Example 6.7: Uniform PRDS p-values with full support

Let $H \sim \mathrm{Bernoulli}(1/2)$. Conditional on $H = h$, let $P_1, \dots, P_m$ be independent with density f_h, where, for $0 < \epsilon < 1/2$,

$$f_0(u) = \begin{cases} 2(1 - \epsilon), & 0 < u < 1/2, \\ 2\epsilon, & 1/2 < u < 1, \end{cases} \qquad f_1(u) = f_0(1 - u).$$

Because $(f_0 + f_1)/2 = 1$, every P_i is uniform on $(0, 1)$. The joint density

$$g(p_1, \ldots, p_m) = \frac{1}{2} \prod_{j=1}^{m} f_0(p_j) + \frac{1}{2} \prod_{j=1}^{m} f_1(p_j)$$

is strictly positive on $(0, 1)^m$, and absolute continuity makes ties have probability zero. To verify PRDS, fix a Borel-measurable increasing set A and a coordinate i, and set

$$a_h(u) = \mathbb{P}\{(u, P_{-i}) \in A \mid H = h\}, \qquad w(u) = \mathbb{P}(H = 1 \mid P_i = u).$$

The conditional law under $H = 1$ stochastically dominates the law under $H = 0$. Consequently, a_0 and a_1 are nondecreasing and $a_1 \geq a_0$. A regular conditional version of w is

$$w(u) = \begin{cases} \epsilon, & u < 1/2, \\ 1/2, & u = 1/2, \\ 1 - \epsilon, & u > 1/2, \end{cases}$$

which is nondecreasing. Writing

$$h_A(u) = \mathbb{P}(P \in A \mid P_i = u) = \{1 - w(u)\}a_0(u) + w(u)a_1(u),$$

we obtain, for $u < v$,

$$\begin{aligned} h_A(v) - h_A(u) = \{1 - w(v)\}\{a_0(v) - a_0(u)\} + w(v)\{a_1(v) - a_1(u)\} \\ + \{w(v) - w(u)\}\{a_1(u) - a_0(u)\} \geq 0. \end{aligned}$$

Thus $P = (P_1, \ldots, P_m)$ is PRDS.

Now take $q = 0.1$, $m = 11$, and $\epsilon = 0.1$. Put $W_i = 1/2 - P_i$, $M_i = |W_i|$, and $S_i = \mathbf{1}\{W_i > 0\}$, and order the magnitudes as $M_{\pi(1)} > \cdots > M_{\pi(11)}$. Conditional on $H = h$, signs and magnitudes are independent, so the reordered signs remain iid Bernoulli(p_h), where $p_0 = 1 - \epsilon$ and $p_1 = \epsilon$. A prefix containing at most nine scores cannot pass because its ratio is at least $1/9 > 0.1$. The ten-score prefix passes exactly when all ten signs are positive, because then the ratio is $1/10 = q$. The eleven-score prefix can pass only when all eleven signs are positive, an event already included in the preceding one. Hence, under the global null,

$$\mathrm{FDR} = \frac{1}{2}\{(1 - \epsilon)^{10} + \epsilon^{10}\} = 0.1743392201 > 0.1.$$

The exponent ten records the convention $\leq q$. With the strict convention $< q$, the ten-score boundary no longer passes and the FDR becomes

$$\frac{1}{2}\{(1 - \epsilon)^{11} + \epsilon^{11}\} = 0.15690529805.$$

Every sign is marginally fair, but for $i \neq j$, $\mathrm{Cov}(S_i, S_j) = (1/2 - \epsilon)^2 > 0$. The latent variable, invisible in the uniform margins, aligns many null signs at once.

The failure is not limited to a specially designed mixture.

Theorem 6.8: Fixed positive equicorrelation can defeat the mirror threshold, cited result [181]

Let $q \in (0, 1)$ and $\rho \in (0, 1)$, and let $Z_0, Z_1, \ldots$ be independent standard normal variables. For each m, apply the knockoff+ threshold to the all-null scores

$$W_i^{(m)} = \sqrt{\rho}\, Z_0 + \sqrt{1 - \rho}\, Z_i, \qquad i = 1, \ldots, m.$$

Then

$$\liminf_{m\to\infty} \mathrm{FDR}_m \geq \frac{1}{2}.$$

Consequently, if $q < 1/2$, FDR control fails for all sufficiently large m. The one-sided p-values $P_i^{(m)} = 1 - \Phi(W_i^{(m)})$ are exactly uniform and PRDS, and their mirror rule has the same rejection set after the corresponding monotone change of threshold.

Proof idea. Condition on $Z_0 = z > 0$. Far enough into the tails, the conditional probability that W_i exceeds a positive threshold is much larger than the probability that it falls below the corresponding negative threshold. As m grows, the two empirical tail counts converge to these conditional probabilities, and their ratio falls below q with probability tending to one. Integrating over $z \geq \delta$ and then letting δ decrease to zero gives the lower bound. This is a liminf statement, not a claim that the FDR converges to $1/2$. Also, although ρ may be any fixed positive number, the largest eigenvalue of the equicorrelation matrix grows with m, so the model is not weakly dependent in every matrix sense.

Low-order diagnostics cannot repair the problem. There are exchangeable score vectors with identical continuous symmetric margins and zero pairwise correlations for which the FDR is arbitrarily close to one. Nor can one obtain a useful distribution-free repair merely by replacing the plus one with a more conservative deterministic function of the two current tail counts. For $q < 1/2$, any such function that is nonincreasing in the positive count and nondecreasing in the negative count is either invalid over the full-support PRDS family above or never rejects. This dichotomy does not cover randomized or nonmonotone rules, or procedures that use the whole path of counts.

These statements do not contradict the knockoff theorems. A valid knockoff filter couples the threshold with a construction whose swap exchangeability and antisymmetric statistic produce the needed conditional null sign flips; see Chapter 9. Conditional sign flips are a clean sufficient condition for this threshold, not a necessary condition for every mirror method. Masking, controlled revelation, sample splitting, conditional calibration, or a calibrated approximate-knockoff argument can place the validity burden elsewhere.

3. Empirical Processes and Reverse Martingales

Route: Optional proof

Threshold procedures can be studied as empirical-process stopping rules. For a p-value threshold t, write

$$V(t) = \#\{i \in \mathcal{I}_0 : P_i \leq t\}, \qquad R(t) = \#\{i : P_i \leq t\}.$$

BH chooses the largest t for which

$$\frac{mt}{R(t) \vee 1} \leq q.$$

Under independent true nulls, the process $V(t)$ is a binomial counting process with mean $m_0 t$. The ratio $V(t)/t$ has a reverse-martingale structure when time runs from 1 down to 0. One convenient reverse filtration is

$$\mathcal{F}_t = \sigma(\{\mathbf{1}(P_i \leq u) : i \in \mathcal{I}_0,\ u \geq t\}),$$

which records the configuration of true-null p-values in the already revealed upper part $[t, 1]$, together with the number $V(t)$ that remain below t. With respect to this decreasing-time

filtration,

$$\mathbb{E}\left[\frac{V(s)}{s}\,\middle|\,\mathcal{F}_t\right] = \frac{V(t)}{t}, \qquad 0 < s < t \le 1,$$

because the true-null p-values not yet revealed, those lying in $[0, t]$, are conditionally exchangeable and uniformly scattered in that interval [151].

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 $s < t$,

$$V(s) \mid V(t) = v \sim \text{Bin}\left(v, \frac{s}{t}\right), \qquad \mathbb{E}\left[\frac{V(s)}{s}\,\middle|\,V(t) = v\right] = \frac{v}{t}.$$

The full filtration version says that this identity remains true after also revealing which true-null p-values lie above t, and after conditioning on nonnull p-values when they are independent of the nulls. This is the technical reason Storey-style thresholds can be stopped before λ without losing the null-count calibration.

This viewpoint gives a procedure-specific design rule: a data-dependent threshold must be a stopping time for a filtration that reveals only information compatible with that procedure's null-count calibration. A filtration suitable for one estimator cannot automatically be reused for another. The next section illustrates the rule with Storey's estimator: tail information above a fixed λ is revealed first, and only then does the threshold search move downward through $[0, \lambda]$.

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. For a tuning parameter $\lambda \in [0, 1)$, define

$$\widehat{\pi}_0^{\lambda} = \frac{1 + \#\{i : P_i > \lambda\}}{(1 - \lambda)m}.$$

The plus one is a finite-sample stabilizer. The idea is that alternatives tend to produce small p-values, so p-values above λ are enriched for nulls.

Figure 6.2 shows the tuning tradeoff directly. At small λ, alternatives can contaminate the right-tail count and push the estimate upward; near one, the estimate is based on very few p-values and becomes visibly variable. The useful range lies between these two failures.

Given $\widehat{\pi}_0^{\lambda}$, an adaptive BH threshold replaces m by $m\widehat{\pi}_0^{\lambda}$:

$$\widehat{T} = \sup\left\{t \le \lambda : \frac{m\widehat{\pi}_0^{\lambda} t}{R(t) \vee 1} \le q\right\}.$$

If the set is empty, set $\widehat{T} = 0$ and make no rejections. The restriction $t \le \lambda$ is important in finite-sample proofs. The right-tail estimate $\widehat{\pi}_0^{\lambda}$ is formed from p-values above λ, and the threshold search then moves downward through $[0, \lambda]$. Thus the estimate is fixed before the reverse-time martingale is stopped. Under independence, the proof uses the binomial identity for the number of true-null p-values above λ:

$$V_0(\lambda) = \#\{i \in \mathcal{I}_0 : P_i > \lambda\} \sim \text{Bin}(m_0, 1 - \lambda).$$

The reciprocal moments of $1 + V_0(\lambda)$ are what make the plus-one version conservative.

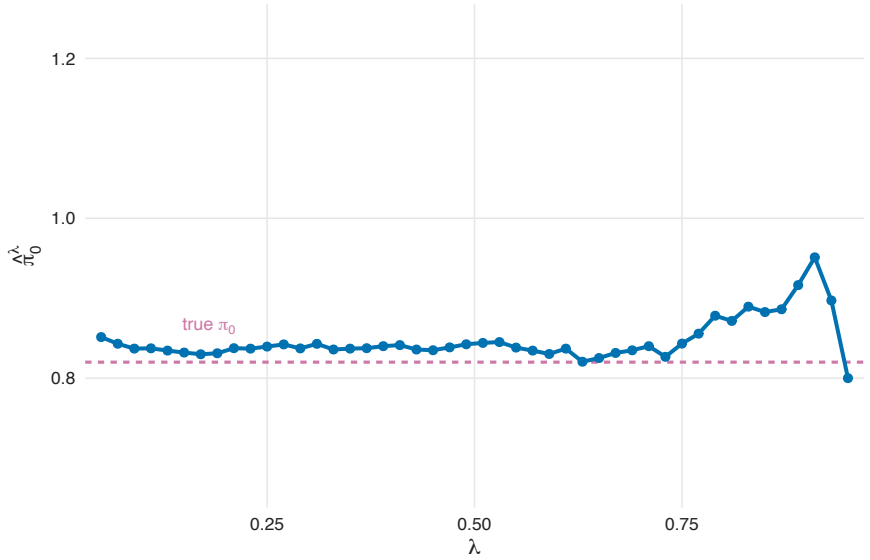

FIGURE 6.2. Storey's estimate $\widehat{\pi}_0^\lambda$ as λ varies in one simulated mixture of $m = 2500$ tests with true null fraction $\pi_0 = 0.82$, null distribution $N(0,1)$, and alternative distribution $N(2.6,1)$. Small λ can be biased by alternatives; large λ can be noisy because few p-values remain in the tail.

Theorem 6.9: Storey's finite-sample bound, simplified

Assume the true-null p-values are independent uniforms and are independent of the nonnull p-values. For fixed $\lambda \in (0,1)$, Storey's procedure above satisfies

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

Proof of Theorem 6.9

Let $V(t) = \#\{i \in \mathcal{I}_0 : P_i \le t\}$. The value λ 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 λ, the remaining true-null p-values in $[0,\lambda]$ are iid uniforms on $[0,\lambda]$. Call this conditioning the *tail information at* λ. 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 λ.

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 λ, so $\widehat{T}$ is a reverse stopping time bounded by λ. Optional stopping gives

$$\mathbb{E}\left[\mathbf{1}\{\widehat{T} > 0\}\frac{V(\widehat{T})}{\widehat{T}}\,\middle|\,\text{tail information at } \lambda\right] = \frac{V(\lambda)}{\lambda}.$$

Thus

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left[\frac{V(\widehat{T})}{R(\widehat{T}) \vee 1}\right] \\ &\leq \frac{q}{m}\mathbb{E}\left[\mathbf{1}\{\widehat{T} > 0\}\frac{V(\widehat{T})}{\widehat{\pi}_0^{\lambda}\widehat{T}}\right] \\ &= \frac{q}{m}\mathbb{E}\left[\frac{m(1-\lambda)}{1 + \#\{j : P_j > \lambda\}}\frac{V(\lambda)}{\lambda}\right].\end{aligned}$$

Since the denominator includes all p-values above λ,

$$1 + \#\{j : P_j > \lambda\} \geq 1 + m_0 - V(\lambda).$$

Here $m_0 - V(\lambda)$ is exactly the number of true-null p-values above λ. 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

$$\mathrm{FDR} \leq q(1 - \lambda^{m_0}) \leq q.$$

□

Storey's estimate is not automatically valid under arbitrary dependence or arbitrary data-dependent choices of λ. 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. The more adaptive the procedure, the more carefully its randomness must be separated from the null p-values it will test.

Figure 6.3 makes this distinction visible by plotting both FDR and power as equicorrelation increases. The power panel records the attraction of estimating π_0, while the FDR panel shows why that gain must be interpreted together with the estimator's dependence assumptions.

5. Q-Values

Route: Core

The q-value of a hypothesis is the smallest FDR level at which that hypothesis would be rejected. In practice, q-values are often computed by applying a Storey-type estimate of π_0 to BH-style adjusted p-values. If $\widehat{\pi}_0$ is fixed, the ordered q-values are

$$Q_{(i)} = \min_{j \geq i}\left\{\frac{\widehat{\pi}_0 m p_{(j)}}{j}\right\} \wedge 1.$$

As with BH-adjusted p-values, the running minimum ensures monotonicity across ordered ranks.

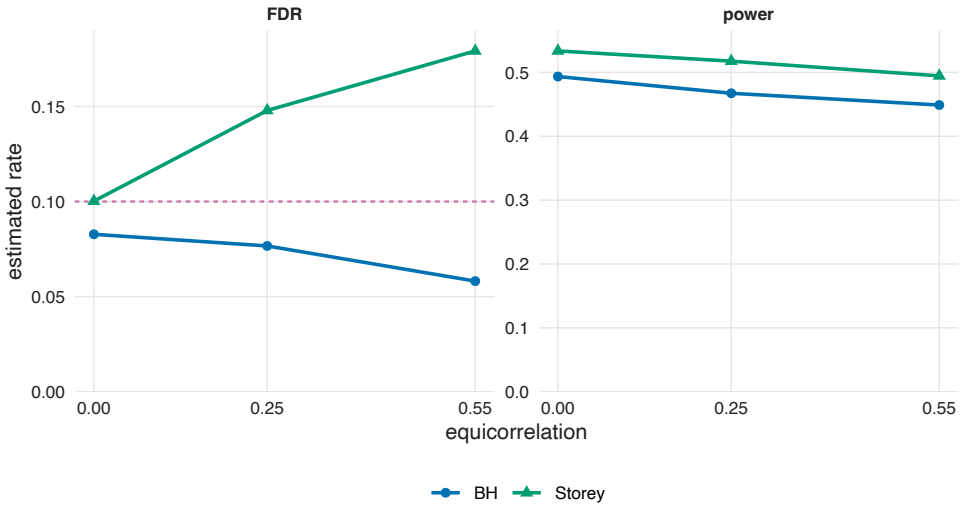

FIGURE 6.3. BH and a simple Storey adaptive procedure under equicorrelated one-sided Gaussian tests at target $q = 0.10$. Each configuration has $m = 400$ tests, including 65 nonnulls with mean 2.35; the Storey estimator uses $\lambda = 0.5$. Points summarize 900 Monte Carlo runs at correlations $\rho \in \{0, 0.25, 0.55\}$. The adaptive procedure gains power when the estimator is reliable, but this simulation also shows why dependence and adaptivity require explicit assumptions.

Writing V for the number of false rejections and R for the total number of rejections, the positive false discovery rate is

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{V}{R} \,\middle|\, R > 0\right],$$

provided $\mathbb{P}(R > 0) > 0$; Section 10 develops its posterior interpretation. A family of rejection regions is *nested* when any two regions are ordered by inclusion—equivalently, relaxing its threshold can add cases but cannot exchange one selected case for another.

Q-values are tail-area quantities: they summarize the error rate for a nested rejection region containing all hypotheses at least as significant as the current one. For a one-sided statistic, the q-value $Q(z)$ attached to z asks about a region such as $\{Z \geq z\}$, not about the single point $Z = z$. This is why q-values are closer to pFDR for nested rejection regions than to local posterior probabilities for individual hypotheses. The distinction becomes clear in the two-groups model.

6. The Two-Groups Model and Local FDR

Route: Core

The empirical-Bayes view starts by putting each hypothesis into a latent 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

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z).$$

The local false discovery rate is the posterior null probability

$$\mathrm{lfdr}(z) = \mathbb{P}(\theta = 0 \mid Z = z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1-\pi_0) f_1(z)}.$$

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 $\mathrm{lfdr}(z)$. Regions dominated by the alternative therefore correspond to low local FDR.

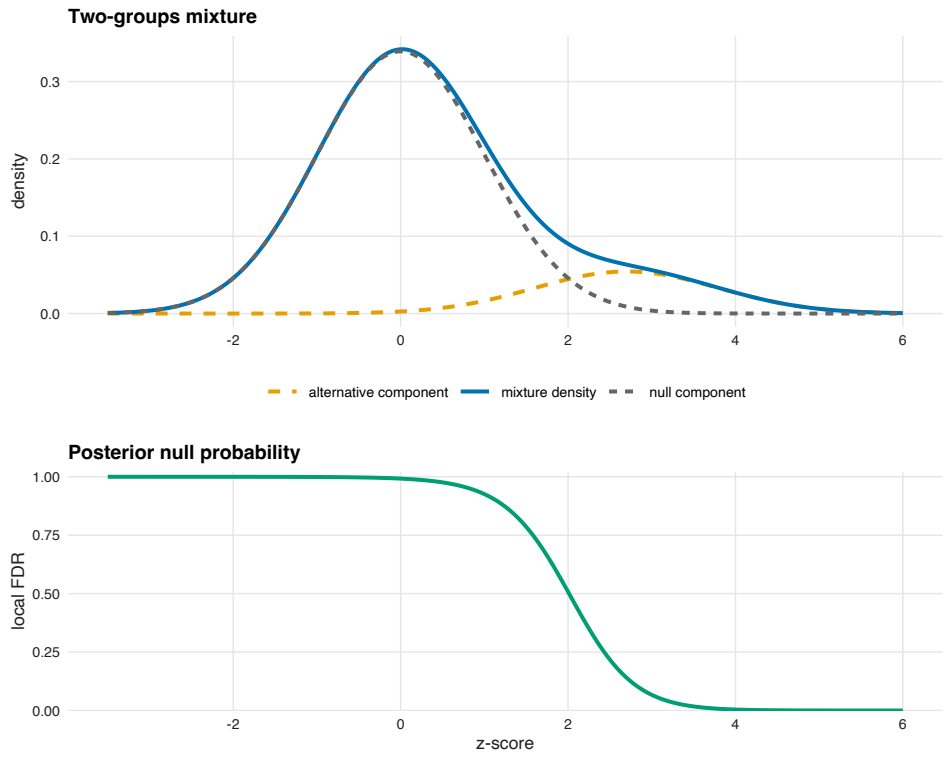

FIGURE 6.4. Two-groups model with $\pi_0 = 0.85$, $f_0 = N(0,1)$, and $f_1 = N(2.7, 1.1^2)$. The top panel shows the weighted null and alternative components, $\pi_0 f_0$ and $(1-\pi_0) f_1$, together with their mixture f. The bottom panel shows local FDR, the posterior probability that a case with z-score z is null.

Local FDR is a pointwise quantity. A q-value is a tail-area quantity. In a one-sided problem, the tail-area FDR for rejecting $Z \geq z$ is

$$\mathbb{P}(\theta = 0 \mid Z \geq z).$$

This averages local FDR values over the whole rejection tail:

$$\mathbb{P}(\theta = 0 \mid Z \geq z) = \mathbb{E}\{\mathrm{lfdr}(Z) \mid Z \geq z\}.$$

Here is the calculation. Let $I = \mathbf{1}\{\theta = 0\}$. Since $\mathrm{lfdr}(Z) = \mathbb{P}(\theta = 0 \mid Z) = \mathbb{E}[I \mid Z]$, conditioning further on the tail event and applying iterated expectation gives

$$\begin{aligned}
\mathbb{P}(\theta = 0 \mid Z \geq z) &= \mathbb{E}[I \mid Z \geq z] \\
&= \mathbb{E}\{\mathbb{E}[I \mid Z] \mid Z \geq z\} \\
&= \mathbb{E}\{\mathrm{lfdr}(Z) \mid Z \geq z\}.
\end{aligned}$$

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

$$\frac{\int_z^\infty \mathrm{lfdr}(u) f(u)\,du}{\int_z^\infty f(u)\,du} = \frac{\int_z^\infty \pi_0 f_0(u)\,du}{\int_z^\infty f(u)\,du}.$$

Thus $\mathrm{lfdr}(z)$ 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. A q-value $Q(z)$ reports the smallest tail error level at which an observation would enter a nested rejection rule. The two numbers can differ substantially, especially when the mixture density changes rapidly.

Figure 6.5 compares the two summaries on the same ranked observations. Each local-FDR point evaluates one case, whereas each q-value pools the entire prefix up to that rank; the separation between the curves is the visible cost of pointwise versus tail-area averaging.

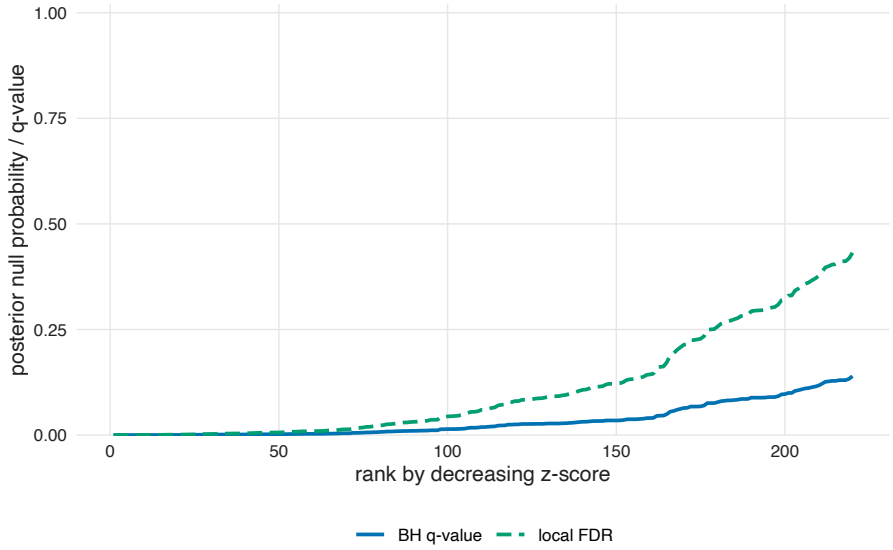

FIGURE 6.5. Local FDR and BH q-values for the 220 largest z-scores from a simulated two-groups sample with $m = 1800$, $\pi_0 = 0.85$, $f_0 = N(0,1)$, and $f_1 = N(2.7, 1.1^2)$. Local FDR ranks individual cases by posterior null probability, while q-values summarize the tail set that would be reported at each rank.

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 π_0, null density f_0, and alternative density f_1. Suppose a rule rejects observations in a region Γ. For m independent draws from the two-groups model, the expected numbers of false and true discoveries are

$$\mathrm{EFP}(\Gamma) = m\pi_0 \int_\Gamma f_0(z)\,dz, \qquad \mathrm{ETP}(\Gamma) = m(1-\pi_0)\int_\Gamma f_1(z)\,dz.$$

The corresponding marginal false discovery rate 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 $\mathbb{E}[V/(R \vee 1)]$, but it is often the natural oracle objective in empirical-Bayes decision theory.

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 spends less false-discovery budget per rejection.

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 Γ 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

Among all oracle procedures δ 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 and some boundary randomization $0 \le \rho(z) \le 1$. If $\ell(Z)$ has no atom at the boundary, the rule is simply $\Gamma_c = \{z : \ell(z) \le c\}$, with c chosen as the largest threshold for which

$$\frac{\int_{\{\ell(z) \le c\}} \ell(z) f(z)\, dz}{\int_{\{\ell(z) \le c\}} f(z)\, dz} \le q.$$

Thus the optimal oracle rejection boundary is a level surface of local FDR.

Proof of Proposition 6.10

Write integrals with respect to the mixture measure $d\nu(z) = f(z)\, dz$. For a rule δ, 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

$$v(\delta) - q r(\delta) \le 0,$$

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 $\eta \geq 0$ 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 η, 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. Thus any rule selected by this constrained optimization 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 nontrivial boundary points therefore have the same local-FDR value. This is the no-side-information version of the equal-boundary condition: the rejection boundary is a level surface of $\ell(z)$.

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 higher-local-FDR set B while not fully rejecting a lower-local-FDR set A, take smaller subsets if necessary so the available ν-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 ℓ. 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 nested family of local-FDR level sets. Within that family, adding points in increasing order of ℓ increases expected true discoveries until adding more points would violate the mFDR constraint, or until all points have been included. If $\ell(Z)$ has no boundary atom, this gives the displayed largest feasible threshold; if there is a boundary atom, ρ randomizes on that atom to obtain the same boundary level. □

The finite-sample posterior version of the same argument treats local FDR as the cost of rejecting one observed case. Conditional on the observed Z_i's, define

$$\ell_i = \mathrm{lfdr}(Z_i) = \mathbb{P}(\theta_i = 0 \mid Z_i).$$

For a rejection set A, the posterior expected number of false discoveries is

$$\mathbb{E}[V(A) \mid Z_1, \ldots, Z_m] = \sum_{i \in A} \ell_i,$$

and the posterior plug-in mFDR of A is

$$\widehat{\mathrm{mFDR}}(A) = \frac{\sum_{i \in A} \ell_i}{|A| \vee 1}.$$

Thus a discovery with smaller ℓ_i has smaller posterior false-discovery cost. The optimal oracle rule should spend its false-discovery budget on the smallest local-FDR values first.

Proposition 6.11: Oracle local-FDR step-up rule

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 that minimize the posterior expected number of false discoveries, the size-k set rejects the k smallest ℓ_i's; among all oracle prefix rules satisfying $\widehat{\mathrm{mFDR}} \le q$, the one with the most rejections is

$$\widehat{k} = \max\left\{1 \le k \le m : \frac{1}{k}\sum_{j=1}^{k} \ell_{(j)} \le q\right\},$$

with the convention $\widehat{k} = 0$ if the set is empty. It rejects the $\widehat{k}$ hypotheses with the smallest local-FDR values and has posterior plug-in mFDR at most q.

Proof of Proposition 6.11

Given the observed Z_i's, ℓ_i is the posterior probability that hypothesis i is null. 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 ℓ_i's.
Among these nested optimal size-k sets, the posterior plug-in mFDR is the prefix average $k^{-1}\sum_{j=1}^{k} \ell_{(j)}$. The displayed rule chooses the largest prefix whose average is at most q, so no other oracle prefix rule can make more rejections while satisfying the same plug-in mFDR constraint. □

In practice the ℓ_i are replaced by estimates $\widehat{\ell_i}$ from a fitted mixture model. The proposition then describes the *oracle* optimum; the realized procedure inherits the model errors of the mixture fit.

This explains why empirical-Bayes ranking can differ from p-value ranking. If the alternative density is asymmetric, heteroscedastic, or multimodal, the most promising observations are those with small posterior null probability, not necessarily those with the smallest classical p-values.

In practice, π_0, f_0, and f_1 must be estimated. Common approaches include maximum likelihood, EM algorithms, central matching for the empirical null, and flexible mixture modeling. These methods are powerful but model-dependent. Their output should be read as posterior-style evidence under the fitted model, not as a distribution-free finite-sample FDR guarantee.

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 an omics-wide screen, genes or CpG sites with higher read depth, stronger baseline expression, known functional annotation, or corroborating evidence from a previous GWAS may have a different prior chance of being nonnull or a different power curve. 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: a moderate p-value may be more compelling in a covariate stratum with high power or low null fraction.

Proposition 6.12: Oracle covariate local-FDR thresholding

Suppose the oracle local-FDR values at the observed data, $\ell_i = \mathrm{lfdr}(p_i, X_i)$, are known, and 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 ℓ_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. Among these nested oracle sets, the displayed rule makes the most rejections subject to the posterior average-local-FDR constraint.

Proof of Proposition 6.12

Given the observed data (p_i, X_i), ℓ_i is the posterior probability that hypothesis i is null. If a rejection set A is chosen after seeing the data, its rejection count is fixed at $R(A) = |A|$, while

$$V(A) = \sum_{i \in A} \mathbf{1}\{\theta_i = 0\}.$$

Taking posterior expectation conditional on the observed data gives

$$\mathbb{E}[V(A) \mid (p_i, X_i)_{i=1}^m] = \sum_{i \in A} \mathbb{P}(\theta_i = 0 \mid P_i = p_i, X_i) = \sum_{i \in A} \ell_i.$$

Therefore the posterior expected FDP, with the usual convention that the FDP is zero when A is empty, is

$$\mathbb{E}\left[\frac{V(A)}{|A| \vee 1} \,\middle|\, (p_i, X_i)_{i=1}^m\right] = \frac{1}{|A| \vee 1}\sum_{i \in A} \ell_i.$$

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 scans these nested optimal sets and chooses the largest prefix whose average is at most q.

Thus its posterior expected FDP is at most q, and no other oracle prefix rule satisfying the same average-local-FDR constraint can make more rejections. □

This is a model-based oracle statement. If $\pi_0(x)$ and $f_1(p \mid x)$ are correctly specified and known, it describes the best posterior ranking rule. In practice they are estimated, for example by smoothing $\pi_0(x)$ over the covariate axis and fitting flexible conditional alternative densities. Then the guarantee is no longer automatic. The usual justification is asymptotic: as the number of hypotheses grows, the fitted local-FDR scores must be accurate enough that the empirical average false-discovery cost is close to the oracle one. This is different from a distribution-free, finite-sample FDR proof. When finite-sample validity is needed, the model should be paired with a separate device that prevents overfitting the null p-values.

9. AdaPT: Local-FDR-Guided Masking

Route: Advanced

AdaPT adds such a validity device. It combines two ideas: a model, often local-FDR-like, proposes powerful covariate-dependent threshold surfaces; a masking and mirror argument prevents the adaptive search from using the null signs it is trying to test. In this section P_i denotes a random p-value, and p_i denotes its observed value.

For each observed p-value define the masked value and sign

$$\widetilde{p}_i = \min(p_i, 1 - p_i), \qquad s_i = \mathbf{1}\{p_i \le 1/2\}.$$

The masked value tells us how close p_i is to either tail, but it hides which tail contains the point. Under a true null with $P_i \mid X_i \sim \mathrm{Unif}[0,1]$, the random sign $\mathbf{1}\{P_i \le 1/2\}$ is a fair coin conditional on $(\min(P_i, 1 - P_i), X_i)$, as stated formally in Proposition 6.13. AdaPT lets the analyst or algorithm update a threshold surface $s_t : \mathcal{X} \to [0, 1/2]$ using the covariates, masked p-values, and signs that have already been revealed, but not the hidden signs of hypotheses still eligible for rejection.

For a fixed threshold surface s, the lower-tail rejection count and upper-tail mirror count are

$$R(s) = \#\{i : p_i \le s(X_i)\}, \qquad B(s) = \#\{i : p_i \ge 1 - s(X_i)\}.$$

The two sets use the same covariate-dependent width $s(X_i)$. The lower tail contains candidate discoveries. The upper tail contains mirror cases that are not rejected but, under a null, are as likely as lower-tail cases after conditioning on the masked value. AdaPT uses the mirror FDP estimate

$$\widehat{\mathrm{FDP}}_{\mathrm{AdaPT}}(s) = \frac{1 + B(s)}{R(s) \vee 1}.$$

The $+1$ is the same finite-sample stabilizer as in mirror procedures. If $s(x) \equiv t$ is constant, this reduces to the no-covariate mirror estimate from Chapter 5. AdaPT is more than the no-covariate mirror rule, however: the threshold surface can change with x. A local-FDR model can suggest larger thresholds in covariate regions where discoveries look more credible, while the masking rule controls which sign information the adaptive search is allowed to see.

Proposition 6.13: Conditional symmetry under masking

For a true-null hypothesis with $P \mid X \sim \mathrm{Unif}(0,1)$, the sign $\mathbf{1}\{P \le 1/2\}$ is Bernoulli(1/2) and conditionally independent of $(\min(P, 1 - P), X)$.

Proof of Proposition 6.13

Conditional on $X = x$, the null p-value is uniform. The map

$$P \mapsto (\min(P, 1-P),\ \mathbf{1}\{P \le 1/2\})$$

sends the uniform distribution on $[0,1]$ to the product of a uniform distribution on $[0,1/2]$ and a Bernoulli$(1/2)$ sign. Since this holds for every x, the sign is also conditionally independent of X. □

Theorem 6.14: AdaPT FDR control, informal simplified statement

Assume the true-null p-values are conditionally independent uniforms given their covariates, and that the threshold sequence s_t is adapted to the filtration generated by the covariates, masked p-values, and previously revealed signs. In particular, while a hypothesis is still eligible for rejection, the update rule has not seen whether its p-value lies in the lower or upper tail. If AdaPT stops at a threshold surface $\widehat{s}$ satisfying

$$\frac{1 + \#\{i : p_i \ge 1 - \widehat{s}(X_i)\}}{\#\{i : p_i \le \widehat{s}(X_i)\} \vee 1} \le q,$$

then, under the regularity conditions of Lei and Fithian [107], the rejection set $\{i : p_i \le \widehat{s}(X_i)\}$ controls FDR at level q.

Proof idea. Among null hypotheses whose masked values place them inside the current two-sided band $\widetilde{p}_i \le s_t(X_i)$, the hidden signs are fair coins. The lower side contributes false discoveries, while the upper side contributes mirror decoys. The key point is not that the threshold surface is fixed in advance; it may be chosen adaptively. The key point is that, for still-hidden hypotheses, the choice of the next surface is based only on information that is independent of their null signs. Conditional on the visible information, the unrevealed null signs remain independent fair coins. This makes the upper-tail mirror count a conservative proxy for the number of lower-tail nulls, with the $+1$ term providing finite-sample stabilization. The formal proof packages this comparison into a supermartingale and applies optional stopping when the mirror estimate falls below q. The local-FDR model is used for power, by proposing promising surfaces $s_t(x)$; the masking and mirror symmetry are what supply the finite-sample validity argument.

This separation is important. A poor local-FDR model can make AdaPT weak or inefficient, but the masking proof can still protect FDR when its assumptions hold. Conversely, if the adaptive model is allowed to peek at the hidden signs of candidate discoveries, it can steer the threshold toward lower-tail nulls without paying for the corresponding upper-tail mirrors, and the mirror comparison is no longer valid.

10. Positive FDR

Route: Core

The positive false discovery rate is

$$\mathrm{pFDR} = \mathbb{E}\left[\frac{V}{R} \,\middle|\, R > 0\right].$$

Since FDP $= 0$ when $R = 0$,

$$\mathrm{FDR} = \mathrm{pFDR}\ \mathbb{P}(R > 0).$$

Under the two-groups model, if a rejection region is Γ, then pFDR has the following posterior interpretation.

Proposition 6.15: Posterior interpretation of pFDR

Under the two-groups model, for a fixed rejection region Γ,

$$\mathrm{pFDR}(\Gamma) = \mathbb{P}(\theta = 0 \mid Z \in \Gamma).$$

This identity is one reason q-values and empirical-Bayes FDR methods are often described as posterior error probabilities for rejection regions. The phrase "for rejection regions" is essential: pFDR and q-values average over all cases inside a selected set, whereas local FDR is the posterior null probability for a single observed case.

Proof of Proposition 6.15

Condition on the event $R(\Gamma) = k > 0$. By exchangeability in the two-groups model, the expected number of nulls among the k selected observations is

$$k\,\mathbb{P}(\theta = 0 \mid Z \in \Gamma).$$

Thus

$$\mathbb{E}\left[\frac{V(\Gamma)}{R(\Gamma)} \,\middle|\, R(\Gamma) = k\right] = \mathbb{P}(\theta = 0 \mid Z \in \Gamma),$$

and averaging over $k \geq 1$ gives the result. □

For a nested family of rejection regions $\{\Gamma_u : 0 < u < 1\}$, the q-value of an observed statistic z can be defined as

$$Q(z) = \inf_{\Gamma_u : z \in \Gamma_u} \mathrm{pFDR}(\Gamma_u).$$

This is the smallest positive-FDR level at which z would enter the rejection region. In monotone one-sided models it coincides with the tail-area interpretation used earlier in this chapter.

11. Conditional Calibration

Route: Research frontier

The final idea is that known dependence can sometimes be exploited rather than ignored. Conditional calibration and dependence-adjusted BH methods construct p-values or rejection thresholds using conditional distributions given dependence-relevant statistics. The goal is to get closer to BH power while retaining valid FDR control under a structured dependence model.

The basic move is the following. Suppose the joint null distribution of the p-value vector $P = (P_1, \ldots, P_m)$ has a sufficient or approximately sufficient statistic S that captures the dependence. Write p_i for the observed value of P_i, and write s_{obs} for the observed value of S. Conditional on $S = s_{\mathrm{obs}}$, the true-null p-values may become exchangeable or may have a tractable conditional null distribution. Rather than running BH on the unconditional observed p-values p_i, one runs it on conditional null p-values

$$\tilde{p}_i = \mathbb{P}_0\Big(P_i^{\mathrm{null}} \leq p_i \,\Big|\, S = s_{\mathrm{obs}}\Big),$$

where P_i^{null} denotes a draw from the conditional null law for hypothesis i. Because $\tilde{p}_i$ is the conditional null CDF evaluated at the observation, it is uniform in the continuous exact case and super-uniform with conservative randomization or discreteness. The price is computational: S

must be modeled and the conditional distribution must be tractable. The dependence-adjusted BH procedure of Fithian and Lei [57] formalizes this idea for a broad class of dependence structures and shows that BH applied to the recalibrated p-values regains FDR control.

This topic is more advanced than the core PRDS and Storey material. The important message for now is that dependence creates a spectrum of choices: use BH when benign dependence conditions are justified, use BY when one wants robustness without structure, or model the dependence explicitly when the structure is scientifically meaningful and reliable.

12. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Inspect null-tail stability, covariate balance, masking logs, and local-FDR fit residuals.
Not identified from these data	The true-null subset, exact PRDS property, and correctness of the two-groups model are not observable.
Sensitivity analysis	Vary null-fraction tuning, empirical-Bayes families, and dependence calibrations.
Conservative fallback	Use nonadaptive BH under a justified condition or BY under arbitrary dependence.

PRDS is a positive-dependence condition, not a synonym for correlation. Its direction depends on whether the coordinates are evidence scores or p-values. BH remains valid under PRDS, while BY remains valid under arbitrary dependence at a power cost. PRDS does not, by itself, validate a bare mirror or knockoff+ threshold. Storey's estimator needs the right tail of the p-value distribution to behave like mostly null p-values and requires care under dependence or data-adaptive tuning. Local FDR and q-values are different: local FDR is pointwise and model-based; q-values summarize tail-area error for sets of discoveries.

13. Bibliographic Notes

Route: Core

The PRDS framework and BH validity under positive dependence are due to Benjamini and Yekutieli [20]; related Simes-type positive-dependence results include Sarkar [140]. Failure of mirror and knockoff+ thresholds under PRDS and other dependent score laws is studied by Zhang [181]. Storey's estimator and q-values are developed in Storey [149], Storey [150], and Storey et al. [151]. The empirical-Bayes local-FDR perspective originates with Efron [52] and is treated comprehensively in the book Efron [53]. The compound-decision formulation and oracle/adaptive local-FDR thresholding are developed by Sun and Cai [155]. Covariate-assisted local-FDR rules and their oracle optimality theory are developed by Zhang and Chen [182] and Cao et al. [37]. Related structured empirical-Bayes approaches, including local-index ideas for dependent and spatial settings, are represented by Sun et al. [156]. AdaPT is due to Lei and Fithian [107], and the broader masking/interactive FDR framework is developed by Lei et al. [108]. General FDR reshaping and dependence ideas are discussed by Blanchard and Roquain [27]. Conditional calibration and dependence-adjusted BH are represented by Fithian and Lei [57]. A frontier direction with growing practical relevance under regulatory constraints such as GDPR and the 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 (ε, δ)-differential privacy of the released rejection set. Dwork et al. [51] develops the theoretical framework.

14. Exercises

Route: Core

Basic.

Exercise 6.16: Increasing sets

Draw two increasing and two non-increasing subsets of $[0,1]^2$ in p-value coordinates. Explain which direction is relevant for the PRDS definition used in this chapter.

Exercise 6.17: Local FDR

Derive

$$\mathrm{lfdr}(z) = \frac{\pi_0 f_0(z)}{\pi_0 f_0(z) + (1-\pi_0) f_1(z)}$$

from Bayes' rule.

Exercise 6.18: pFDR identity

Show that

$$\mathrm{FDR} = \mathrm{pFDR}\ \mathbb{P}(R > 0).$$

Then, under the two-groups model, show that for a fixed rejection region Γ, $\mathrm{pFDR}(\Gamma) = \mathbb{P}(\theta = 0 \mid Z \in \Gamma)$.

Intermediate.

Exercise 6.19: Gaussian PRDS direction

Let Z be a Gaussian vector with nonnegative covariances and suppose larger Z_i is stronger evidence. Explain why monotone transformations to right-tail p-values reverse the coordinate order. State the corresponding PRDS condition in p-value coordinates.

Exercise 6.20: BH under PRDS

In the BH proof under PRDS, show that the event $\{R(t) \le k\}$ is increasing in the p-value vector. Explain how this replaces the independence step in the leave-one-out proof.

Exercise 6.21: Storey's estimator

Derive Storey's estimator

$$\widehat{\pi}_0^{\lambda} = \frac{1 + \#\{P_i > \lambda\}}{(1-\lambda)m}.$$

Discuss the bias-variance tradeoff as λ changes.

Exercise 6.22: Optimal covariate local-FDR thresholding

Let p_i be the observed value of P_i, and set $\ell_i = \mathrm{lfdr}(p_i, X_i)$. For a rejection set A, verify that the posterior expected number of false discoveries is $\sum_{i\in A} \ell_i$ and that the posterior expected FDP is

$$\frac{1}{|A| \vee 1} \sum_{i\in A} \ell_i.$$

Use an exchange argument to show that, among rejection sets of fixed size k, the set with smallest posterior expected number of false discoveries rejects the k smallest local-FDR values. Derive the oracle step-up rule that chooses the largest k for which

$$\frac{1}{k} \sum_{j=1}^{k} \ell_{(j)} \le q,$$

with $k = 0$ if no nonempty prefix satisfies the inequality. Explain why this is a model-based posterior optimality statement rather than a finite-sample frequentist FDR proof.

Computational.

Exercise 6.23: Storey under dependence

Reproduce Figure 6.3. Vary λ, the signal strength, and the correlation. Identify settings where the adaptive estimator is too aggressive.

Exercise 6.24: Q-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 heavy-tailed alternative density. Use one-sided p-values $p = 1 - \Phi(Z)$. Compute BH q-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. Explain why q-values summarize tail rejection sets, while local FDR ranks individual cases by posterior null probability.

Exercise 6.25: Two-stage BKY

Let $q \in (0,1)$ and set $q_1 = q/(1+q)$. The two-stage procedure of Benjamini et al. [22] first runs BH at level q_1, obtaining R_1 rejections. If $R_1 = 0$ or $R_1 = m$, it stops. Otherwise it runs 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 Storey's procedure under independent p-values with $\pi_0 \in \{1, 0.9, 0.7\}$. Then repeat under positively correlated Gaussian z-scores. Report empirical FDR and power, and explain why this is an adaptive method for structured settings, with independence as the basic theorem, rather than an arbitrary-dependence replacement for BY.

Advanced.

Exercise 6.26: Why masking suffices for AdaPT

Consider a single null hypothesis with $P \sim \text{Unif}[0,1]$, masked value $\widetilde{P} = \min(P, 1-P)$, and sign $S = \mathbf{1}\{P \leq 1/2\}$. Show that conditional on $\widetilde{P}$, the sign S is Bernoulli$(1/2)$ and independent of any function of $\widetilde{P}$. For a covariate-dependent threshold $s(x)$, identify the lower-tail event $P \leq s(X)$ and the upper-tail mirror event $P \geq 1 - s(X)$. Explain why an AdaPT update based only on covariates, masked p-values, and already revealed signs cannot peek at the hidden null signs. Then explain why fitting the next threshold using unmasked signs of still-eligible hypotheses would invalidate the upper-tail mirror comparison.

Exercise 6.27: Why two-sided Gaussian p-values need care

Let $Z \sim N(0, \Sigma)$ and define two-sided p-values

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

Show that $\{P_i \leq t\}$ is not monotone in Z_i, and explain why this blocks a direct application of Theorem 6.2 but proves neither validity nor failure of BH. The counterexample of Dobriban [46] uses $4N$ nonnulls to enlarge the rejection count. Explain why it disproves general FDR control without settling the global-null two-sided Gaussian Simes bound.

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 per-level FDR targets (TreeBH), and (iv) replicability across studies, where a discovery requires evidence in enough separate experiments (partial conjunction). 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 certify the individual leaves inside a selected group.

Several groupings of the same leaves.: Use the p-filter when every leaf belongs to several prespecified layers—possibly nested, possibly overlapping—and 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 partial-conjunction 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 $\mathcal{G}_1, \dots, \mathcal{G}_G$, 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^{group} is the number of *rejected null groups* (i.e., null groups in which the procedure rejected at least one hypothesis) and R^{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 $40/840 \approx 4.8\%$, which is within target. But 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. A procedure controls *both* only if it accounts for the group structure explicitly. The methods of this chapter let the analyst declare the resolutions of interest in advance and obtain calibrated FDR at each one simultaneously.

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 $p_{g,1}, \dots, p_{g,m_g}$, 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 kth-smallest p-value in group g. Group BH then applies the ordinary BH procedure at level q to $p_1^{\mathrm{Simes}}, \dots, p_G^{\mathrm{Simes}}$ 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, the latter condition holds for exact-valid PRDS p-values under the usual Simes assumptions. 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 $\mathbb{P}(p_g^{\text{Simes}} \le t) = t$. For completeness, the induction proof is the same as in Theorem 3.1: split on the largest order statistic $U_{(m_g)}$, condition on $U_{(m_g)} = s$, and rescale the remaining $m_g - 1$ uniforms to $[0,1]$. The integral gives $t(1 - t^{m_g - 1}) + t^{m_g} = t$.
Now relax to independent super-uniform nulls. Since each $p_{g,i}$ stochastically dominates a uniform random variable, Strassen's monotone coupling theorem gives a pair $(p'_{g,i}, U_i)$ with $p'_{g,i} \overset{d}{=} p_{g,i}$, $U_i \sim \text{Unif}(0,1)$, and $p'_{g,i} \ge U_i$ almost surely. Applying this coupling independently for $i = 1, \dots, m_g$ preserves the product joint law of the p-values and makes $U_1, \dots, U_{m_g}$ independent uniforms. 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^{Simes}. The PRDS example is the positive-dependence Simes theorem of Sarkar [140]; 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^{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 correlations are strong enough to violate the Simes assumptions, one needs a Bonferroni-type aggregation (or a BY-reshaped Simes) inside the group 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 $p_{(k)} \leq qk/m$. Here $p_{(4)} = 0.031 \leq 0.0333$, 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 [9] solves this.

Setup. Let the hypotheses $\{1,\dots,m\}$ carry L partition layers; layer ℓ 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 ℓ. 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-ℓ 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-ℓ 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 ℓ.

The iterative algorithm. For each layer ℓ and each group g at that layer, let $p_g^{(\ell)}$ denote a layer-ℓ group p-value – in practice, a super-uniform aggregator such as Simes applied to the within-group p-values (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-ℓ 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 [9] 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

$$\#\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{t}(\boldsymbol{k})) \ge k_\ell \qquad \text{for every layer } \ell.$$

Equivalently, 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 ℓ, and $\boldsymbol{t}$ is the largest such vector on the finite BH ladders.

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_ℓ by $\min(k_\ell, r_\ell)$ and t_ℓ by $q_\ell k_\ell / G_\ell$ for every layer, then repeat until $\boldsymbol{k}$ stabilizes.

The iteration is coordinate-wise monotone: each k_ℓ 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 [9].

Validity. The theorem below is quoted as a result from Barber and Ramdas [9]. A simple sufficient case is mutual independence of all base p-values, with true-null base p-values super-uniform, and group p-values formed by Simes or weighted Simes as in that paper. More general positive-dependence cases are possible, but they are assumptions on the full collection of null group p-values across layers, not automatic consequences of having valid p-values within each group.

Theorem 7.5: p-filter FDR control, cited result [9]

Under the assumptions of Barber and Ramdas [9] (in particular, the layer-ℓ group p-values $\{p_g^{(\ell)}\}$ are valid super-uniform combinations of the within-group p-values, and the dependence structure of the null group p-values satisfies their joint independence-or-PRDS condition), the p-filter output at levels $q_1, \ldots, q_L$ satisfies $\mathrm{FDR}^{(\ell)} \le q_\ell$ for every ℓ.

Proof idea. At a single layer ($L = 1$), the boundary (3) is exactly the BH self-consistency characterization: the largest t for which $G\,t/\#\widehat{\mathcal{R}} \le q$ and $\widehat{\mathcal{R}} = \{g : p_g \le t\}$ is the BH threshold. The BH leave-one-out proof of Chapter 5 then applies verbatim and gives $\mathrm{FDR} \le q$. For multiple layers, Barber and Ramdas [9] show, via a per-layer leave-one-out argument that mirrors the single-layer case but accounts for the cross-layer coupling through the shared hypothesis-level rejection set, that the aggregate contribution of null groups at layer ℓ to $\mathrm{FDR}^{(\ell)}$ is bounded by q_ℓ; their detailed bookkeeping under within-layer PRDS is the main technical content of that paper, and we refer the reader to it for the full argument.

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

A different and more permissive structure is a 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* (or "prefixed") if, whenever a leaf is rejected, all of its ancestors up to the root 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 claiming the containing pathway as a discovery, since

the pathway null "no leaf in the pathway is non-null" is implied to be false by the very rejection of one of its descendant genes.

Algorithm. The TreeBH procedure of Bogomolov et al. [29] implements a top-down rejection rule. Let $C(v)$ denote the children of node v, and write ρ for the formal root. We index depth from $d = 1$ at the topmost *tested* layer (the children of ρ) down to $d = D$ at the leaves; ρ 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 $d = 1, \dots, D$.

TreeBH uses valid p-values throughout the tree. Leaves carry their original p-values. For an internal node w, define its p-value by a valid combination rule applied to the p-values of its immediate children, typically Simes; recursively, this supplies valid p-values for all internal nodes under the TreeBH dependence assumptions. Using Simes over all descendant leaves is a valid simplified aggregation in some settings, but the formal TreeBH algorithm is naturally stated through immediate-child p-values.

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 ρ 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 ρ 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 concentrates the available q_{d+1} budget into branches that survived earlier screens. For a two-level hierarchy this product reduces to the familiar 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 ℓ. For a selected node u at depth ℓ, 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 ℓ. At the formal root ρ, it is the recursively averaged FDP for the whole selected tree at depth ℓ. 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 displayed algorithm is a pedagogical summary of the procedure in Bogomolov et al. [29]. The exact path-specific levels, valid tree p-values, the paper's dependence assumptions, and the simple-selection-rule condition are all part of the cited theorem. The simple-selection-rule condition is substantive: selected child families must come from the prescribed BH-on-valid-child-p-values step so that the selected-family adjustment applies.

Theorem 7.6: TreeBH recursive selected-family FDR control, cited result

Apply the TreeBH procedure of Bogomolov et al. [29] with the exact path-specific levels above, valid tree p-values, the paper's dependence assumptions, and the required simple selection rule at each tested family. Then the TreeBH output satisfies

$$s\,\mathrm{FDR}^{\ell} \le q_\ell, \qquad \ell = 1, \dots, D.$$

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. [29] – 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

$$q_{3,P} = q_3 \cdot \frac{6}{50} \cdot \frac{R_{\mathrm{genes}}(P)}{G_{\mathrm{genes}}(P)}.$$

The variant threshold is therefore pathway-specific. The global ratio $0.10 \cdot R_{\mathrm{genes,total}}/800$, 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 $s\,\mathrm{FDR}^{\ell}$ 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:

$$\text{all variants}$$
$$\Downarrow$$
$$p_{\text{variant}} \le t_{\text{variant}},\quad p^{\text{Simes}}_{\text{gene}} \le t_{\text{gene}},\quad p^{\text{Simes}}_{\text{pathway}} \le t_{\text{pathway}}$$
$$\Downarrow$$
$$\text{one leaf set}$$

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:

$$\text{pathways}$$
$$\Downarrow$$
$$\text{genes within selected pathways}$$
$$\Downarrow$$
$$\text{variants within selected genes}$$

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 Simes gene and pathway p-values can be inputs to both procedures, but that does not make the methods interchangeable. Their 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: each rejected leaf has a structurally certified parent at every 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 *partial conjunction null* negates replicability. Writing $\mathcal{N}_i$ for the set of true-null per-study nulls of feature i,

$$H_i^{(r/K)} : \quad |\mathcal{N}_i| \ge K - r + 1,$$

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 [28] (with foundational antecedents in Benjamini and Heller's earlier work on partial conjunction 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 $p_{i,(1)}\le\cdots\le p_{i,(K)}$ be their order statistics. The r-out-of-K Bonferroni-style partial conjunction 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): $p_i^{(1/K)}=Kp_{i,(1)}$, 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 bounded below by the smallest null p-value, hence super-uniform.

Theorem 7.11: Super-uniformity of the Bonferroni PC p-value

Suppose that, under the partial conjunction 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

$$p_{i,(r)} \geq \min_{k \in S} p_{i,k}$$

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 $\mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \leq s) \leq (K - r + 1)\, s$ without any independence assumption, because the union bound is dependence-free. Hence $\mathbb{P}(p_i^{(r/K)} \leq t) \leq \mathbb{P}(\min_{k \in \mathcal{N}_i^\star} p_{i,k} \leq t/(K - r + 1)) \leq t$, establishing the claim. □

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*

$$p_i^{(r/K),\text{Simes}} = \left[\min_{r \leq j \leq K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1} \right] \wedge 1$$

  sharpens Bonferroni under conditions ensuring the Simes inequality for the relevant true-null per-study p-values, such as independence or suitable PRDS [140]; 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 χ^2_4 calibration similarly rejects with probability $e^{-9.488/4} \approx 0.093$ under $p_1 = p_2 = U$. The advanced exercises work through these calculations.

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 [76] 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 partial-conjunction 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, $\max(0.03, 0.02) = 0.03$, is not small enough to certify replication at the joint BH threshold of 0.025 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 false-discovery proportion 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 [88] (ICH E9 addendum on estimands), the EMA *Multiplicity Issues in Clinical Trials* guideline [54], and the FDA *Multiple Endpoints in Clinical Trials* guidance [161]. 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 false-discovery proportion across many findings. Treating these documents as endorsements of hierarchical FDR would overstate their scope.

7. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Audit the hierarchy, group overlap, aggregation p-values, and whether the scientific claim matches the controlled layer.
Not identified from these data	The hierarchy's scientific correctness and every cross-layer dependence condition are not learned from the final p-values.
Sensitivity analysis	Repeat with alternative prespecified groupings and report which discoveries depend on the hierarchy.
Conservative fallback	Use flat BY or layer-specific Bonferroni/Holm when structured assumptions are doubtful.

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 uses a stronger multilayer joint assumption; a simple sufficient case is mutual independence of all base p-values, with true-null base p-values super-uniform and Simes or weighted-Simes group p-values formed as in the p-filter theorem.

TreeBH uses valid p-values at internal nodes, usually built recursively by applying Simes to the immediate child p-values. Thus it requires the dependence assumptions that make those

node-level p-values valid and that make the simple-selection-rule theorem apply; this is what validates the recursive selected-family FDR bound of Theorem 7.6. "Arbitrary within-node dependence" would break ordinary Simes and therefore the TreeBH chain unless one replaces Simes by a Bonferroni or BY-reshaped node-level aggregator.

Partial conjunction 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 [140], 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 [178]. The modern tree-structured procedure with multi-resolution FDR control is from Bogomolov et al. [29]. The p-filter and its multilayer generalization are due to Barber and Ramdas [9]. Adaptive group-aware procedures for partially ordered side information are from Li and Barber [110]. The original partial conjunction framework is Benjamini and Heller [18]; the replicability extension to multiple features and two-study FDR control is developed by Bogomolov and Heller [28] and Heller and Bogomolov [76]. The connection to regulatory practice in clinical trials is documented in International Council for Harmonisation [88], European Medicines Agency [54], and U.S. Food and Drug Administration [161].

9. Exercises

Route: Core

Basic.

Exercise 7.16: Group BH by hand

Six hypotheses are grouped as $\{1, 2, 3\}$ and $\{4, 5, 6\}$ with p-values $(0.028, 0.028, 0.028)$ and $(0.031, 0.40, 0.50)$. Compute the Simes representative for each group and apply BH at $q = 0.05$ on the two representatives. Compare with flat BH on all six p-values. Verify the arithmetic of Example 7.4: flat BH rejects the four smallest p-values (three from Gene 1 plus the smallest p-value of Gene 2), while group BH rejects no group.

Exercise 7.17: Partial conjunction p-value

Two studies report p-values

$$(p_{1,1}, p_{2,1}, p_{3,1}, p_{4,1}) = (0.001, 0.002, 0.03, 0.45), \qquad (p_{1,2}, p_{2,2}, p_{3,2}, p_{4,2}) = (0.0005, 0.5, 0.02, 0.03)$$

for four features. Compute the 2-out-of-2 Bonferroni partial conjunction p-values and apply BH at $q = 0.05$. Compare with BH applied separately within each study and with the naive intersection of the two single-study rejection sets. Verify the arithmetic of Example 7.15: PC rejects only feature 1, while the naive intersection is $\{1, 3\}$.

Exercise 7.18: Tree-respecting rejection

Consider a rooted tree with two internal nodes A and B, where A has leaves $\{1, 2\}$ and B has leaves $\{3, 4\}$. Suppose a tree-aware procedure rejects the internal node A but not B, while flat BH on the four leaves would reject leaves $\{1, 3\}$. Under the rule that a leaf may be reported only if every ancestor on its path has been rejected, identify the tree-respecting leaf rejection set and the flat-BH-only rejection.

Intermediate.

Exercise 7.19: Simes within groups

Fill in the proof of Theorem 7.2. In particular, verify the equality

$$\mathbb{P}\left(\bigcup_{k=1}^{m_g} \{U_{(k)} \le tk/m_g\}\right) = t$$

for independent uniform $U_{(1)} \le \cdots \le U_{(m_g)}$ by induction on m_g, or by the area argument used in the original Simes [146] paper.

Exercise 7.20: 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

$$p_i^{(r/K),\text{Simes}} = \left[\min_{r \le j \le K} \frac{(K - r + 1)\, p_{i,(j)}}{j - r + 1}\right] \wedge 1.$$

Show that this is super-uniform under $H_i^{(r/K)}$, for example when the per-study p-values are independent (or suitable PRDS [140]) and the true-null p-values are super-uniform. Hint: argue that the least favorable case for validity occurs when the $r - 1$ non-null p-values are pushed to zero; in that least-favorable configuration, the $K - r + 1$ largest order statistics $p_{(r)}, \ldots, p_{(K)}$ coincide with the order statistics of the $K - r + 1$ null p-values, and Simes's inequality applied to those null order statistics gives the bound. Then show algebraically that the Simes-style PC p-value is uniformly no larger than the Bonferroni-style PC p-value at every realization: the Simes formula is the minimum over a set whose $j = r$ term equals the Bonferroni value $(K - r + 1)p_{i,(r)}$, so the minimum is $\le$ the Bonferroni value at every realization.

Exercise 7.21: 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 to group BH of Section 2.

Computational.

Exercise 7.22: 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 $Z_i \sim \mathcal{N}(\mu_i, 1)$, 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. For TreeBH, compute internal-node p-values recursively from immediate children using Simes; for the p-filter, compute layer group p-values using the same Simes aggregation convention. 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 ≥ 200 replicates and report variant-, gene-, and pathway-level FDR (or recursive selected-family FDR for TreeBH) and power for each procedure. (An extended version of this exercise scales the hierarchy up to genome-wide proportions of $\sim 10^5$ variants; we keep the scale modest here for pedagogical clarity.)

Exercise 7.23: Replication study

Generate three independent batches of $m = 1000$ p-values each. In each study, the per-feature p-value is $p = \Phi(-Z)$ with $Z \sim \mathcal{N}(\mu, 1)$; set $\mu = 0$ for null features and $\mu = 2.5$ for signal features. In the joint design, 100 features are signals in at least two of the three studies ("2-out-of-3 replicated truths"), 100 are signals in exactly one study, and 800 are null in all three. Compare three procedures targeting $r = 2$-out-of-$K = 3$ replication: (a) BH at $q = 0.05$ applied to each study separately, then declaring a feature replicated if it is rejected in at least two of the three study-specific BH analyses, (b) BH applied to the 2-out-of-3 Bonferroni PC p-values, and (c) BH applied to the 2-out-of-3 Simes-style PC p-values (which differ from Bonferroni PC because $r < K$). Average over ≥ 200 replicates and report the realized replication FDR and power, counting false discoveries as selected features with fewer than two non-null studies and true discoveries as selected features among the 100 replicated-truth features.

Exercise 7.24: Tree depth and power

Generate a three-level balanced tree with 64 leaves (a branching factor of 4 at each level, so depth 3). Leaf p-values are $p_i = \Phi(-Z_i)$ with $Z_i \sim \mathcal{N}(\mu_i, 1)$: $\mu_i = 0$ for nulls and $\mu_i = 2.5$ for signals. Place all signal in one top-level branch, so every leaf below one child of the formal root is a signal and all other leaves are null. Internal-node p-values are formed recursively from immediate children using Simes; in this balanced simulation, one may alternatively compare with a fixed descendant-leaf aggregation rule if it is declared in advance and remains valid. Apply TreeBH with top-tested-layer target $q_1 = 0.05$ and $q_d = 0.10$ for lower tested depths $d \geq 2$, and record leaf-level power. Now repeat with a deeper tree of the same total size: 64 leaves arranged as six tested levels with branching factor 2, keeping all other simulation parameters fixed and using the same per-depth target convention. Average over ≥ 200 replicates and discuss how the additional depth affects leaf-level power and the recursive selected-family FDR.

Advanced.

Exercise 7.25: 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 at the leaf level with TreeBH using $(q_1, q_2, q_3, q_4) = (0.05, 0.10, 0.10, 0.10)$ over at least 500 Monte

Carlo replicates. Report leaf-level power and explain the interpretation gained when a rejected leaf is accompanied by rejected ancestors.

Exercise 7.26: Stouffer at the global null can fail under positive dependence

For two studies with one-sided z-scores $Z_k = \Phi^{-1}(1 - p_k)$, $k = 1, 2$, the *Stouffer global-null combination* (i.e., the $r = 1$, $K = 2$ Stouffer PC) is

$$p^{(1/2),\text{Stouffer}} = 1 - \Phi\left(\frac{Z_1 + Z_2}{\sqrt{2}}\right).$$

Under the global null (both $p_k \sim \text{Unif}(0, 1)$) with *independent* studies, show that $p^{(1/2),\text{Stouffer}}$ is exactly uniform. Then suppose p_1 and p_2 are *perfectly positively dependent*, $p_1 = p_2 = U \sim \text{Unif}(0, 1)$ (so $Z_1 = Z_2 = Z$ with $Z \sim \mathcal{N}(0, 1)$). Show that the Stouffer combination evaluates to $1 - \Phi(\sqrt{2}\, Z)$, and that at the nominal $q = 0.05$ cutoff, $\mathbb{P}(p^{(1/2),\text{Stouffer}} \leq 0.05) = \mathbb{P}(Z \geq 1.645/\sqrt{2}) = \mathbb{P}(Z \geq 1.163) \approx 0.122$, so the Stouffer global-null p-value is anti-conservative by more than a factor of 2. Discuss why this kind of failure does not happen for the Bonferroni global-null p-value $p^{(1/2),\text{Bonf}} = 2 \min(p_1, p_2) \wedge 1$, which remains super-uniform under any dependence by the union bound, and relate the contrast to the dependence-robustness theme of Heller and Bogomolov [76].

Exercise 7.27: Sharp PC with verified side information

Heller and Bogomolov [76] 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. Derive the corresponding "conditional" PC p-value – the Bonferroni union bound applied only to the studies in S_i – and show that, under independence across features, BH applied to this conditional PC p-value controls FDR at level q without further inflation. Comment on what could go wrong if S_i is inferred from the same p-values, or if the side information is only an unverified belief: explain why such a procedure cannot in general be justified for FDR control 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.

Advanced extensions are developed in Appendix A. Section 4 of Appendix A explains why weak convergence alone does not preserve the e-value expectation budget and develops uniform-integrability, truncation, and tail-bound remedies for asymptotic e-values. Section 5 of Appendix A treats predictable online FDR allocation, including the roles of the filtration, time order, delayed outcomes, alpha-wealth, LORD, SAFFRON, and e-LOND. Section 6 of Appendix A shows how inverting an e-process through $C_t = \{\theta : M_t(\theta) < 1/\alpha\}$ yields simultaneous-in-time confidence sequences, together with mixture and stitched boundaries.

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 Θ_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_θ, 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 tth 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 τ 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 $P \in \mathcal{H}_0$. For $0 < u \leq 1$,

$$\mathbb{P}_P(p_E \leq u) = \mathbb{P}_P(E \geq 1/u) \leq u\,\mathbb{E}_P[E] \leq u,$$

where the inequality is Markov's inequality. The case $u = 0$ is immediate. 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 $p_E = 1$.

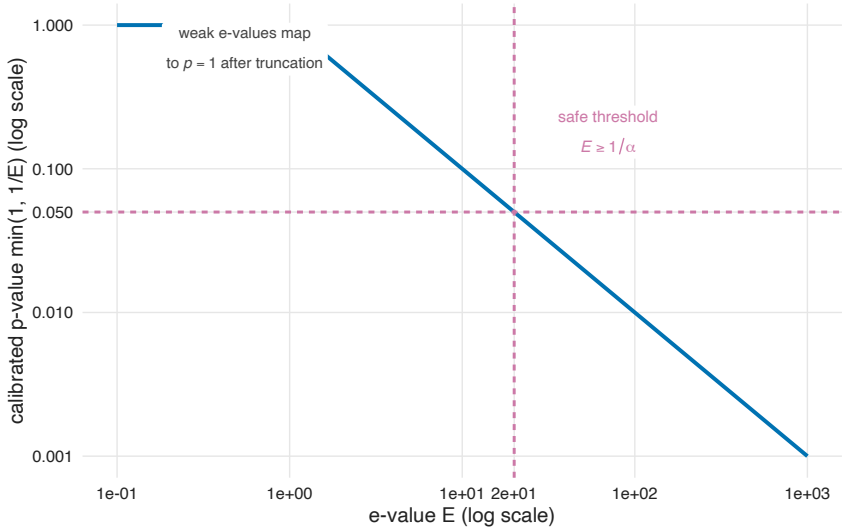

FIGURE 8.1. Calibration from e-values to p-values on log-log axes. The conversion $p_E = \min\{1, 1/E\}$ is always valid by Markov's inequality. Weak e-values with $E < 1$ all map to $p_E = 1$, so calibration is valid but can lose information.

The 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, $f(0) = \infty$, 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. 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 [170] and Chapter 2 of Ramdas and Wang [130]. □

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 ∞ 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 the mixture calibrator, define the endpoint by continuity: $f_{\mathrm{mix}}(1) = 1/2$. The second example is the average of f_κ over $\kappa \in (0,1)$, useful when one does not want to commit to a single κ. The function equal to $2(1-p)$ for $p > 0$ and to ∞ 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.

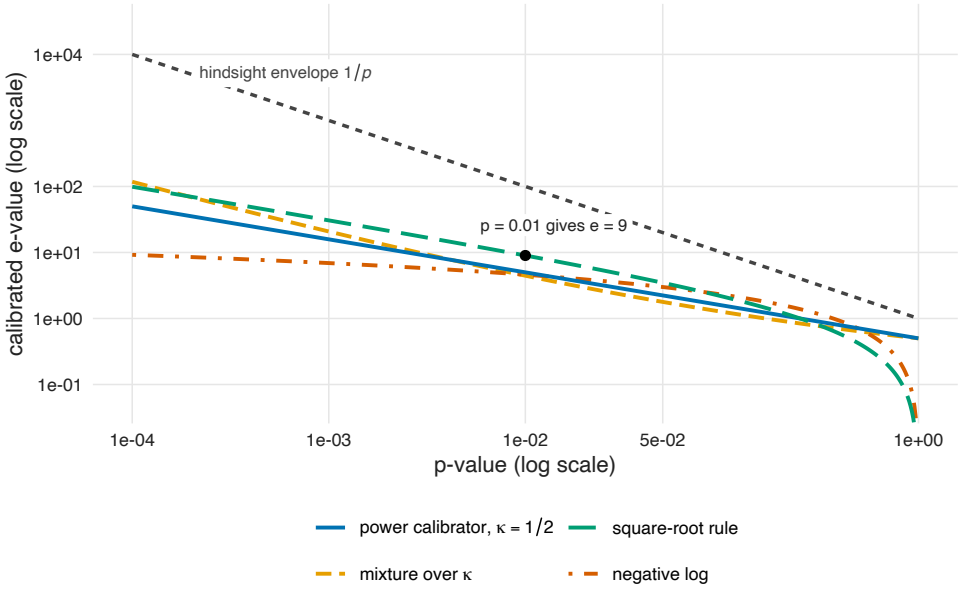

FIGURE 8.2. Four admissible calibrators on log-log axes, with the hindsight envelope $1/p$ as a dashed line. Every calibrator lies below the envelope, and no calibrator dominates another: the mixture is largest for very small p-values, the square-root rule near $p = 0.01$, the negative logarithm for 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 $f(p) = p^{-1/2} - 1$.

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,

$$\int_0^1 f(u)\,du \ge \int_0^{p_0} f(u)\,du \ge p_0 f(p_0) > 1.$$

Equality $f(p_0) = 1/p_0$ at a fixed point p_0 forces $f(u) = \mathbf{1}\{u \le p_0\}/p_0$ 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

$$f_r(u) = \frac{\mathbf{1}\{u \le r\}}{r}$$

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 [130, 170]. 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 α. No p-to-e calibrator has this property: by the bound $f(p) \leq 1/p$, converting p to $f(p)$ and then testing $f(p) \geq 1/\alpha$ is strictly more conservative than testing $p \leq \alpha$, except for the all-or-nothing calibrator at its own threshold. 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. Finally, calibration is exhaustive: on rich enough sample spaces, every e-value can be written as $f(p)$ for some calibrator f and some exactly uniform p-value p, one satisfying $\mathbb{P}_P(p \leq u) = u$ for all $u \in [0, 1]$ under the null [130].

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 $\mathbf{1}\{p \leq \alpha\}/\alpha$; for the latter, $E = 10$ only records that $p \leq 0.1$. Second, rejecting when $E \geq 1/\alpha$ gives a valid predeclared level-α e-test by Markov's inequality. This operational threshold is not a calibration theorem or a numerical conversion between p-values and e-values. 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 $E' \geq E$ 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 [130]. 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 μ 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 Θ_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 θ.

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 $\mathcal{H}_0$ 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.

Numeraire and reverse projection. For a composite null and a fixed alternative, the growth-optimal e-value is the numeraire; when an appropriate reverse information projection exists, it is a likelihood ratio against that projected null law. This is an optimality statement, not an additional requirement for ordinary e-value validity. Existence, uniqueness, support details, and the reverse-projection derivation are developed in Appendix A.

Universal inference: validity without regularity conditions. The guess-and-verify recipe requires a good guess, and for many composite nulls no tractable guess is known. 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 [174].

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 θ, 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 θ, so, P_θ-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_θ. 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 θ, 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, and it is dominated by the numeraire for the same null and alternative [100]. 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 [50, 174].

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 p_{G_1}/p_{G_0} with a prior on the null is an e-value precisely when Q_{G_0} is the reverse information projection of Q_{G_1} onto the null, in which case it is the numeraire. In this sense, the numeraire is the unique Bayes factor that is also an e-value [100].
- 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-α 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.

The logarithm of M_t explains what a good e-value is trying to optimize. If $X_1, X_2, \ldots$ are i.i.d. from an alternative Q, 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 $\mathbb{E}_{P_0}[E_1] \leq 1$. 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.

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.

Definition 8.9: E-power

The *e-power* of an e-value E against an alternative Q is $\mathbb{E}_Q[\log E]$, which may be $-\infty$ or $+\infty$.

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 $E_1, E_2, \dots$ 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 Kullback-Leibler divergence:

$$\mathbb{E}_Q[\log E^*] = D(Q \,\|\, P_0).$$

Proof of Theorem 8.11

Proposition A.3, applied with the point mass at the single null distribution, shows that E^* is the numeraire: $\mathbb{E}_Q[E/E^*] = \mathbb{E}_{P_0}[E] \le 1$ for every e-value E, and Jensen's inequality converts this into log-optimality. The e-power is $\mathbb{E}_Q[\log\{g(X)/p_0(X)\}] = D(Q \,\|\, P_0)$ by the definition of the divergence. For uniqueness, we treat the case where the relevant expectations are finite; the general case follows from the uniqueness of the numeraire [100]. Let E be any log-optimal e-value. Log-optimality of E tested against E^* gives $\mathbb{E}_Q[\log(E^*/E)] \le 0$, and log-optimality of E^* tested against E gives $\mathbb{E}_Q[\log(E/E^*)] \le 0$; the two expectations sum to zero, so both vanish. Then

$$0 = \mathbb{E}_Q\left[\log \frac{E}{E^*}\right] \le \log \mathbb{E}_Q\left[\frac{E}{E^*}\right] \le 0,$$

and equality in Jensen's inequality for the strictly concave logarithm forces E/E^* to be Q-almost surely constant; the constant is one because the middle expectation equals one. □

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 $-\infty$ [130]. 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 $\lambda \in [0, 1]$ [130]; 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 μ_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 ν 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 θ 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 ν 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 grows under either sign of the true mean.

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 $X_1, \dots, X_{s-1}$. 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. Both recipes achieve the oracle growth rate in the limit: under mild conditions, $n^{-1}\mathbb{E}_Q[\log E_n] \to D(Q \,\|\, P_0)$ for every alternative Q in the model, for the mixture method whenever the prior puts mass around the true parameter, and for the plug-in whenever the estimates are consistent [69, 175]. 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.

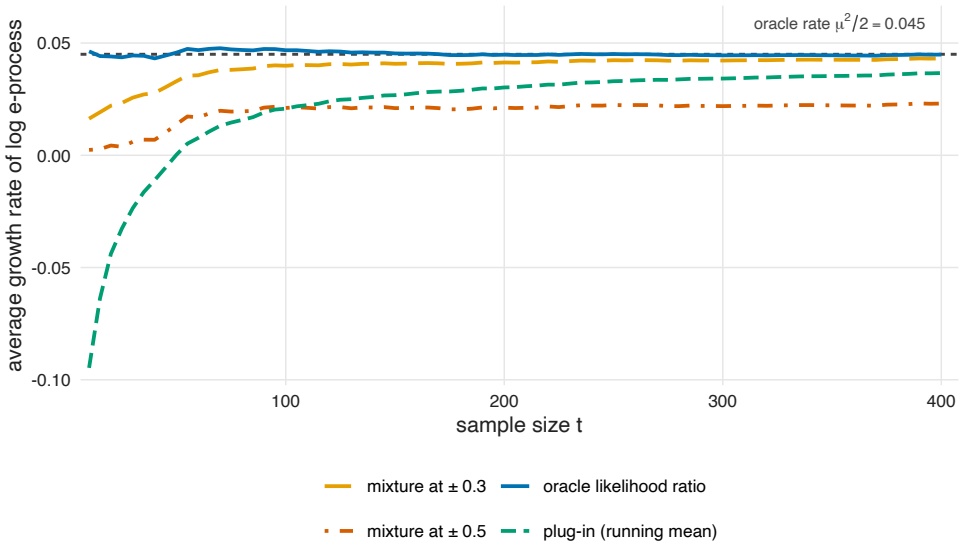

FIGURE 8.4. Average growth rate $t^{-1}\mathbb{E}_Q[\log M_t]$ of four e-processes for $H_0 : \mu = 0$, estimated from 300 simulated $N(0.3, 1)$ sequences of length 400. The oracle likelihood ratio that knows $\mu = 0.3$ sits at the rate $\mu^2/2 = 0.045$; the predictable plug-in with the running-mean estimate climbs toward the same rate; the two-point mixture at ± 0.3 approaches it; and the mixture at ± 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, Markov's inequality is an equality only for an e-value supported on the two points 0 and $1/\alpha$ with mean one; for likelihood-ratio e-values, the realized type-I error of the rule $M_n \geq 1/\alpha$ at a fixed n is typically far below α. 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 α [129]. Third, no expressibility is lost by working on the e-value scale: any level-α 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 [130].

Post-hoc validity: choosing the level after the data. Classical testing fixes the level α before the data are seen, and the guarantee evaporates if α 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-α test, with $\phi_\alpha = 1$ meaning rejection, and suppose ϕ_α is nondecreasing as α 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

$$\mathbb{E}_P\left[\sup_{\alpha\in(0,1)} \frac{\phi_\alpha}{\alpha}\right] \leq 1 \qquad \text{for every } P \in \mathcal{H}_0.$$

The ratio ϕ_α/α is the Type I error budget spent if one rejects at level α; the supremum charges the analyst for the most aggressive level they could choose after seeing the data. For a fixed α, the definition recovers the usual guarantee, since $\mathbb{P}_P(\phi_\alpha = 1) = \mathbb{E}_P[\phi_\alpha] \leq \alpha$.

Proposition 8.13: E-value tests are post-hoc valid

Let E be an e-value for $\mathcal{H}_0$ and let $\phi_\alpha = \mathbf{1}\{E \geq 1/\alpha\}$. Then

$$\sup_{\alpha\in(0,1)} \frac{\phi_\alpha}{\alpha} \leq E \quad \text{pointwise},$$

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 α equals E, attained at $\alpha = 1/E$. In the remaining case $E = \infty$, the ratio is $1/\alpha$ for every α, 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 [68, 98]. The practical reading deserves emphasis. Observing $E = 25$ licenses the statement "rejected at level 0.05" even if the level 0.05 was settled on after seeing E, and equally licenses "rejected at level 0.04." Observing $p = 0.0001$ licenses rejection only at levels fixed in advance. For this reason, the calibrated p-value $1/E$ carries a strictly stronger guarantee than an ordinary p-value of the same numerical size, and comparing the two at face value understates the e-value. A multiple-testing version of the same phenomenon 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. Suppose E_t is a conditional e-value,

$$\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le 1,$$

and let $\lambda_t \in [0,1]$ be $\mathcal{F}_{t-1}$-measurable. Thus λ_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

$$\mathbb{E}[1 - \lambda_t + \lambda_t E_t \,|\, \mathcal{F}_{t-1}] = 1 - \lambda_t + \lambda_t \mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le 1.$$

Consequently,

$$M_t = \prod_{s=1}^{t} \{1 - \lambda_s + \lambda_s E_s\}$$

satisfies the explicit conditional calculation

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}\mathbb{E}[1 - \lambda_t + \lambda_t E_t \mid \mathcal{F}_{t-1}] \le M_{t-1}.$$

It is therefore a nonnegative test supermartingale, and

$$\mathbb{P}_0\left(\sup_{t \ge 0} M_t \ge 1/\alpha\right) \le \alpha.$$

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 $\mathbb{E}[E_i] \le 1$ for every true null i. For multiple testing, the proof only needs the aggregate null-evidence condition

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m. \tag{A}$$

Vectors $(E_1, \dots, E_m)$ 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 $m_0 \le m$. So do deterministically weighted e-values $w_i E_i$ with $w_i \ge 0$ and $\sum_i w_i \le m$. 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. [85]; the condition itself powers the FDR theorem of Wang and Ramdas [173].

Definition 8.14: The e-BH procedure

Order the evidence scores from largest to smallest:

$$E_{(1)} \ge E_{(2)} \ge \cdots \ge E_{(m)}.$$

At FDR level q, define

$$\widehat{k} = \max\left\{k \in \{1, \dots, m\} : E_{(k)} \ge \frac{m}{qk}\right\},$$

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 $E_1, \dots, E_m$ are nonnegative and satisfy (A), then e-BH controls FDR:

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

In particular, if each true null $i \in \mathcal{I}_0$ satisfies $\mathbb{E}[E_i] \le 1$, then $\mathrm{FDR} \le (m_0/m)q \le q$.

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

$$E_i \ge E_{(R)} \ge \frac{m}{qR}.$$

Therefore, path by path,

$$\frac{\mathbf{1}\{i \in \mathcal{R}\}}{R} \le \frac{qE_i}{m}.$$

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 $E_1, \ldots, E_m$, 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.

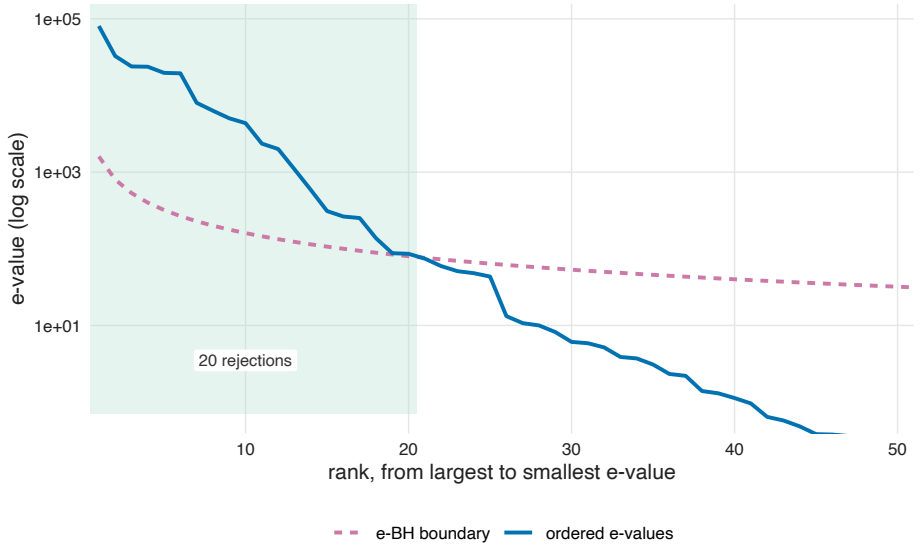

FIGURE 8.5. The e-BH rule orders e-values from largest to smallest and compares the first 50 ranks of a simulation with $m = 160$, 36 nonnulls, $n = 12$, working effect $\theta = 0.9$, and target $q = 0.10$ against the boundary $m/(qk)$. The largest rank $\widehat{k}$ at which the ordered e-value crosses its boundary determines the cutoff; all top $\widehat{k}$ hypotheses are rejected, including any earlier ranks whose values lie below their own rankwise boundaries.

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, under only coordinatewise null expectation budgets and natural symmetry conditions, self-consistent e-BH is essentially universal. The formal statement and proof are deferred to Appendix A.

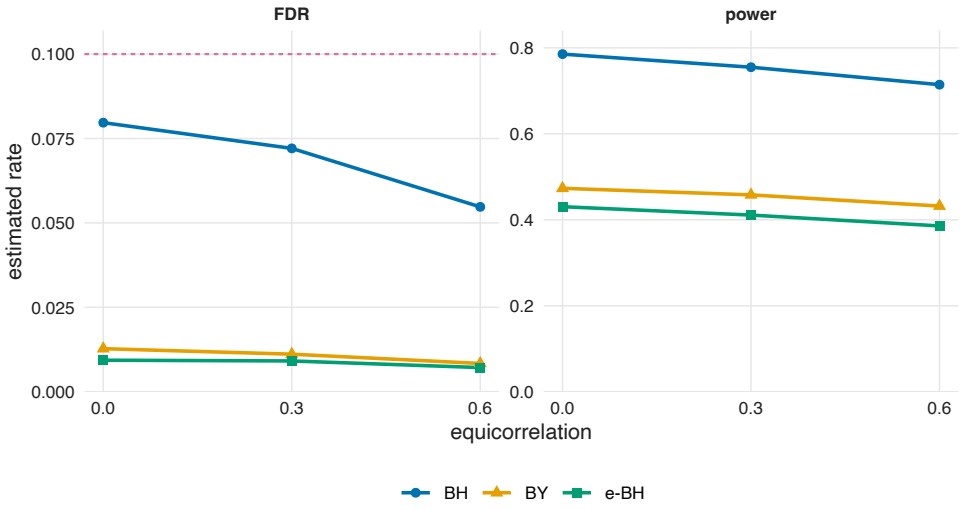

FIGURE 8.6. Simulation with $m = 400$ hypotheses, 80 nonnulls, equicorrelated Gaussian test statistics based on $n = 20$ observations with working effect $\theta = 0.65$, 650 Monte Carlo runs, correlations $\rho \in \{0, 0.3, 0.6\}$, and $q = 0.10$. The e-values are Gaussian likelihood ratios $\exp\{\theta\sqrt{n}\,Z - n\theta^2/2\}$. The lesson is not that e-BH is always more powerful; it is that e-BH's proof uses an e-value moment condition rather than a p-value dependence condition.

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:

$$\begin{array}{c}\text{threshold rule}\\ \Downarrow\\ \text{induced evidence vector}\\ \Downarrow\\ \text{aggregate null-evidence proof}\\ \Downarrow\\ \text{e-BH FDR control}\end{array}$$

The point is not to rename old procedures. The point is to identify the exact budget that must be controlled when designing new procedures.

BH as the first example. Recall from Chapter 5 that BH can be written in threshold form. Let $p_1, \dots, p_m$ 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

$$\widehat{t}_{\mathrm{BH}} = \sup\left\{0 < s \le 1 : \frac{ms}{R(s) \vee 1} \le q\right\}.$$

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 [26] 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_{BY} is nonincreasing, the kth largest e-value corresponds to the kth 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 kth 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 calculation reframes the harmonic penalty. The factor $H_m \approx \log m$ is not an unavoidable tax on dependence; it is exactly the cost of forcing p-values through a calibrator, the worst-case exchange rate between the probability scale and the expectation scale. Evidence born on the e-value scale enters e-BH with no correction under the same arbitrary dependence. When only

p-values are available, however, the calibration route is essentially the best possible: BY is the p-value optimum of this construction [130, 173].

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

$$\widehat{t} = \sup\left\{ s \in \mathcal{T} : \frac{\widehat{M}(s)}{R(s) \vee 1} \leq q \right\}$$

and rejects all i with $R_i(\widehat{t}) = 1$. We assume throughout that the supremum is attained, as it is for the right-continuous step functions arising from p-value thresholds.

The induced evidence formula requires a positive false-discovery budget when there is at least one rejection. We therefore assume

$$R(\widehat{t}) > 0 \quad \Longrightarrow \quad \widehat{M}(\widehat{t}) > 0.$$

If a candidate rule can have $\widehat{M}(\widehat{t}) = 0$ and $R(\widehat{t}) > 0$, 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

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \leq m,$$

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 ψ_i and define

$$Z_i = \psi_i(p_i).$$

Assume ψ_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, ψ_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 $\widehat{M}(s)$. One simple choice is $mg(s)$, where the curve g is chosen before testing. Another choice uses the mirror idea from Chapter 5: estimate the left-tail null count by looking at very large p-values. Under a well-calibrated null, p-values near 0 and p-values near 1 are both rare tail events. Thus the right tail can sometimes be used as a conservative guide to how many null p-values might fall into the left tail. 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 λ, 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 [85, 173]. 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 $\widehat{M}(s)$ [111, 112]. 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 $\lambda_\ell \ge 0$ with $\sum_\ell \lambda_\ell \le 1$,

$$\mathbb{E}_P\left[\sum_{\ell=1}^{L} \lambda_\ell E_\ell\right] = \sum_{\ell=1}^{L} \lambda_\ell \, \mathbb{E}_P[E_\ell] \le 1$$

by linearity, no matter how the E_ℓ 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 $\mathbb{E}[\prod_\ell E_\ell] = \prod_\ell \mathbb{E}[E_\ell] \le 1$, 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 λ_0 on the constant one, and the admissible e-merging functions are exactly these $F_{\boldsymbol{\lambda}}$ [130, 170]. For symmetric merging rules, the arithmetic mean has a related but weaker property called *essential domination*: if a symmetric e-merging function G satisfies $G(\boldsymbol{e}) > 1$, then the arithmetic mean is at least $G(\boldsymbol{e})$. This is not pointwise domination below one. The geometric and harmonic means are pointwise bounded by the arithmetic mean.

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 ℓ, define

$$R_\ell = |\mathcal{R}_\ell|, \qquad V_\ell = |\mathcal{R}_\ell \cap \mathcal{I}_0|, \qquad \mathrm{FDR}_\ell = \mathbb{E}\left[\frac{V_\ell}{R_\ell \vee 1}\right].$$

Controlling groupwise FDR at level q means $\mathrm{FDR}_\ell \leq q$ for every group ℓ. 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 $\mathrm{FDR}^{\mathrm{pool}} \leq q$. 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_ℓ 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 ℓ 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 ℓ'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 $w_{\ell i} \ge 0$. If

$$\sum_{\ell=1}^{L} \max_{1\le i\le m} w_{\ell i} \le 1,$$

then the combined vector satisfies

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] \le m.$$

Consequently, e-BH applied to $E_1,\dots,E_m$ controls FDR at level q.

Proof of Proposition 8.20

Using linearity of expectation and nonnegative weights,

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L}\sum_{i\in\mathcal{I}_0} w_{\ell i}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L}\left(\max_i w_{\ell i}\right)\sum_{i\in\mathcal{I}_0}\mathbb{E}[E_i^{(\ell)}] \\ &\le \sum_{\ell=1}^{L}\left(\max_i w_{\ell i}\right) m \le m.\end{aligned}$$

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 datasets. Now suppose the hypotheses are split across disjoint families

$$\mathcal{G}_1, \dots, \mathcal{G}_L, \qquad |\mathcal{G}_\ell| = m_\ell, \qquad \sum_{\ell=1}^{L} m_\ell = m.$$

Family ℓ produces evidence $E_i^{(\ell)}$ only for $i \in \mathcal{G}_\ell$, and assume that within that family

$$\mathbb{E}\left[\sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} E_i^{(\ell)}\right] \le m_\ell.$$

This displayed bound has an immediate consequence: if family ℓ runs e-BH only on $\{E_i^{(\ell)} : i \in \mathcal{G}_\ell\}$, then Theorem 8.15 with $m = m_\ell$ gives $\mathrm{FDR}_\ell \le q$. 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 $w_{\ell i} \ge 0$. If

$$\sum_{\ell=1}^{L} m_\ell \max_{i \in \mathcal{G}_\ell} w_{\ell i} \le m,$$

then the pooled vector satisfies

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] \le m.$$

Thus e-BH controls $\mathrm{FDR}^{\mathrm{pool}}$ for the pooled rejection list.

Proof of Proposition 8.21

$$\begin{aligned}
\mathbb{E}\left[\sum_{i \in \mathcal{I}_0} E_i\right] &= \sum_{\ell=1}^{L} \sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} w_{\ell i} \mathbb{E}[E_i^{(\ell)}] \\
&\le \sum_{\ell=1}^{L} \left(\max_{i \in \mathcal{G}_\ell} w_{\ell i}\right) \sum_{i \in \mathcal{G}_\ell \cap \mathcal{I}_0} \mathbb{E}[E_i^{(\ell)}] \\
&\le \sum_{\ell=1}^{L} m_\ell \max_{i \in \mathcal{G}_\ell} w_{\ell i} \le m.
\end{aligned}$$

□

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 FDR_ℓ. 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.

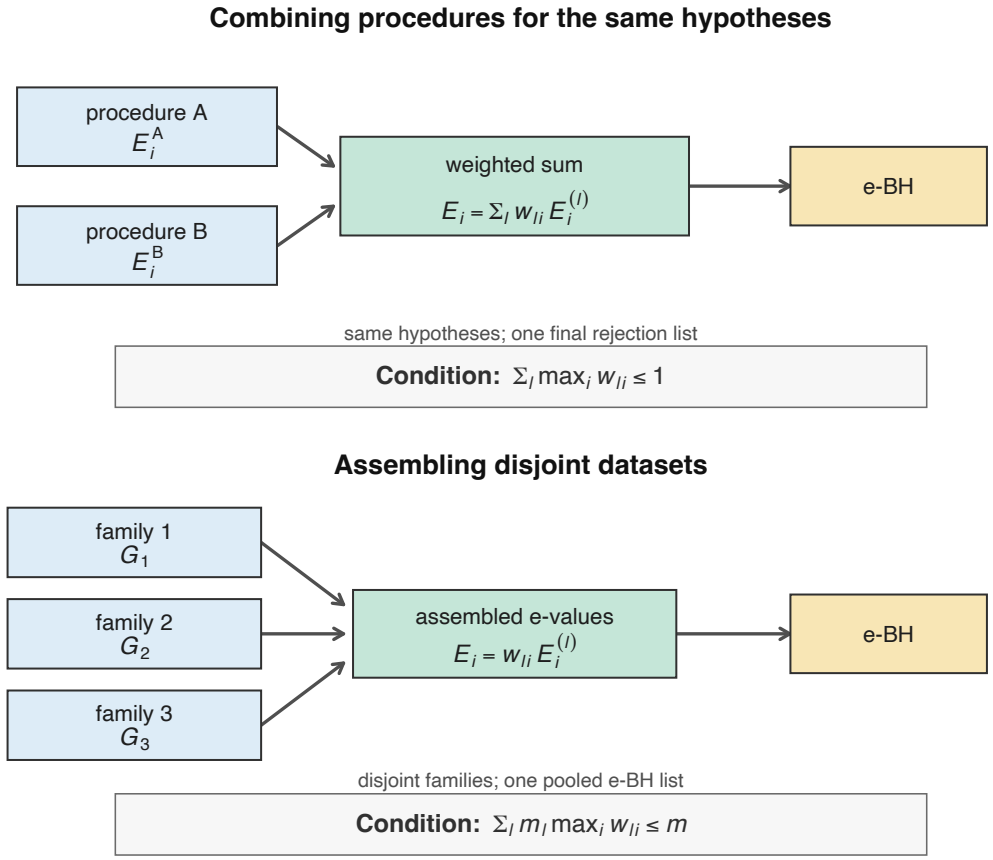

FIGURE 8.7. Combining evidence from several procedures on the same hypotheses and assembling evidence from disjoint datasets. In assembling, the main output is one pooled e-BH list; family-level aggregate bounds also justify separate family lists when reported. The max-weight conditions protect against unknown true-null placement.

E-values as weights in p-value procedures. A hybrid setting arises when both evidence scales are present for the same hypotheses: p-values $p_1, \dots, p_m$ from confirmatory data, and e-values $E_1, \dots, E_m$ 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 $p_1^*, \dots, p_m^*$. 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 $(m_0/m)q$ [84]. Two features distinguish it from the weighted BH rules of Chapter 5. The weights need no normalization: classical weighted BH requires the weights to average to one, while here the e-value validity $\mathbb{E}[E_i] \le 1$ for true nulls substitutes for normalization, so the

weight vector may be arbitrarily enthusiastic about the alternatives. And nothing is assumed about the dependence within the e-value vector itself. One caution: the result is a theorem, not a plug-in observation. Dividing PRDS p-values by independent random weights does not obviously preserve PRDS, so the classical weighted-BH guarantee does not directly apply; the proof conditions on the e-values and controls the null contributions with $\mathbb{E}[E_i] \leq 1$.

9. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Recompute conditional expectation identities, replay stopping rules, and audit whether every adaptive choice is predictable.
Not identified from these data	The null conditional law and the absence of unlogged peeking cannot be verified from the terminal e-value alone.
Sensitivity analysis	Cap bets, vary working alternatives, and compare split, swapped, and averaged constructions.
Conservative fallback	Use a fixed-horizon valid test or an all-or-nothing e-value from a valid p-value.

The e-value condition is a null expectation bound. For simple nulls, likelihood ratios give canonical examples. For composite nulls, the bound must hold uniformly over the whole null model. A Bayes factor averaged over a null prior is not enough unless it also satisfies the uniform expectation bound; when the null mixture is the reverse information projection of the alternative, the resulting likelihood ratio is not only valid but log-optimal. Conversions between the two evidence scales are asymmetric: p-values become e-values only through a calibrator fixed in advance, at a real cost in magnitude, while an e-value always yields the p-value $\min\{1, 1/E\}$, which additionally survives levels chosen after seeing the data.

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. Products of conditional e-values, which are nonnegative supermartingales under the null, are the canonical construction, but the general definition asks only for the stopped-value property, and some important e-processes are not supermartingales.

E-BH controls FDR under arbitrary dependence because its proof reduces FDP to aggregate null evidence before taking expectations. The aggregate condition defines compound e-values, and it characterizes FDR control completely: every FDR procedure is e-BH applied to some compound e-values. Dependence can still affect power and construction of good e-values, but it does not enter the e-BH proof once the aggregate evidence bound is established, and the same pathwise argument yields the budget-form post-hoc guarantee in Corollary 8.16. When only asymptotically normal statistics are available, approximate e-values inherit an inflated but explicit FDR guarantee; the per-coordinate (ε, δ) bound becomes $(m_0/m)(1+\varepsilon)q + m_0\delta$.

Threshold-to-eBH arguments are design tools. They do not make a procedure valid by algebra alone. The hard step is always the same: prove that the induced evidence vector has acceptable aggregate null expectation.

Online p-value procedures rely on predictable levels and conditional super-uniformity. Online e-value procedures replace conditional super-uniformity by conditional e-value validity. Confidence sequences are the interval analogue of e-processes: invert a family of anytime-valid tests, meaning

tests whose level-α guarantee holds at every stopping time, and obtain simultaneous coverage over time.

Before applying these results, specify the null model, the filtration, and whether the target guarantee is fixed-time, anytime-valid, pooled, or groupwise. Verify the relevant expectation bound over the entire null: an individual or aggregate bound for e-BH, and a conditional bound for a sequential product. Calibrators, weights, and online testing levels must be fixed or predictable as required by their theorems. For approximate e-values, state whether the convergence is pointwise or uniform and report the resulting slack, including the factor $m\delta$ when only a bound in terms of the total number of hypotheses is available.

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 [130]. E-values, calibration, and merging are developed by Vovk and Wang [170]; the calibrator concept goes back to Vovk [164], and the term e-power and its theory are developed in Chapter 3 of Ramdas and Wang [130]. Safe testing and optional continuation are treated by Grünwald et al. [69]; the betting interpretation is emphasized by Shafer [142] and connected to the broader game-theoretic view by Ramdas et al. [134]. The numeraire, its unconditional existence and uniqueness, and the general theory of the reverse information projection are due to Larsson et al. [100], with information-theoretic precursors in Li [113] and Csiszár and Tusnády [41]. Universal inference is due to Wasserman et al. [174]; the power cost of splitting is studied by Dunn et al. [50]. The general definition of e-processes and their role in anytime-valid inference are developed by Howard et al. [82], Ramdas et al. [132], and Grünwald et al. [69]; the strict gap between e-processes and test supermartingales, illustrated by exchangeability testing, is established by Ramdas et al. [133]. Post-hoc validity and the equivalence between post-hoc valid testing and e-values are developed by Grünwald [68] and Koning [98]; the randomized Markov inequality is studied by Ramdas and Manole [129].

E-BH and its arbitrary-dependence FDR proof are due to Wang and Ramdas [173], including the observation that BY is e-BH on calibrated p-values. Compound e-values are named and systematically studied by Ignatiadis et al. [85], who prove the universality of e-BH; e-values as unnormalized weights in p-value procedures are analyzed by Ignatiadis et al. [84]; derandomization through averaged compound e-values is developed by Ren and Barber [135]. The super-uniformity lemma behind the PRDS remark is due to Blanchard and Roquain [26]. The threshold-to-eBH viewpoint and the aggregation and assembling ideas are based on Li and Zhang [112] and Li and Zhang [111]. Randomized and closed variants of e-BH, which uniformly improve it in certain regimes, are surveyed in Chapter 9 of Ramdas and Wang [130]. Approximate and asymptotic e-values, including the universal t-test example, are developed in Chapter 10 of Ramdas and Wang [130] and in Section 6.2 of Wang and Ramdas [173].

Online FDR control begins with alpha-investing [59] and includes LORD [91] and SAFFRON [131]. Online e-value rules such as e-LOND are developed by Xu and Ramdas [176]. Confidence sequences in the modern nonasymptotic form are developed by Howard et al. [82]; betting confidence sequences for bounded means are developed by Waudby-Smith and Ramdas [175]. E-confidence intervals and false coverage rate control through e-values, the e-BY procedure, are due to Xu et al. [177].

11. Exercises

Route: Core

Basic.

Exercise 8.22: Markov calibration

Let E be an e-value. With the convention $1/0 = +\infty$, prove that $p_E = \min\{1, 1/E\}$ is a valid p-value. Then give an example showing why $1/p$ is not generally an e-value when $p \sim \mathrm{Unif}(0,1)$.

Exercise 8.23: Gaussian likelihood-ratio e-value

Work through Example 8.6. Derive the likelihood ratio from the normal densities, verify that its null expectation is one, and compute the expected log e-value under $\mu = \mu_1$.

Exercise 8.24: Calibrator family

Using the integral criterion of 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. Compute the e-value that $f_{1/2}$ assigns to $p = 0.01$. Then prove that every calibrator satisfies $f(p) \le 1/p$ for all $p \in (0,1]$.

Intermediate.

Exercise 8.25: Conditional products

Let E_t be nonnegative and $\mathcal{F}_t$-measurable with $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] \le 1$. Prove that $M_t = \prod_{s=1}^{t} E_s$ is a nonnegative supermartingale. Identify exactly where measurability is used.

Exercise 8.26: Ville's inequality

Prove Ville's inequality for a finite deterministic horizon K. Then pass to the infinite horizon by monotone convergence of the crossing events.

Exercise 8.27: E-BH proof

Reproduce the proof of Theorem 8.15. Explain why the proof does not require independence among the e-values.

Exercise 8.28: BH as e-BH

Let $\widehat{t}_{\mathrm{BH}}$ be the BH threshold and define

$$E_i^{\mathrm{BH}} = \widehat{t}_{\mathrm{BH}}^{-1} \mathbf{1}\{p_i \le \widehat{t}_{\mathrm{BH}}\}$$

when $\widehat{t}_{\mathrm{BH}} > 0$, with all induced e-values equal to zero otherwise. Show that e-BH applied to these induced e-values gives the BH rejection set. Which part of the argument is algebraic, and which part is a validity proof?

Exercise 8.29: Combining procedures

Suppose two procedures for the same family of m hypotheses produce nonnegative evidence vectors $\{E_i^{(1)}\}$ and $\{E_i^{(2)}\}$, each satisfying

$$\mathbb{E}\left[\sum_{i\in\mathcal{I}_0} E_i^{(\ell)}\right] \le m, \qquad \ell = 1, 2.$$

Choose fixed, nonnegative, nonconstant weights w_{1i}, w_{2i} satisfying

$$\max_i w_{1i} + \max_i w_{2i} \le 1.$$

For $E_i = w_{1i}E_i^{(1)} + w_{2i}E_i^{(2)}$, verify the aggregate bound $\mathbb{E}[\sum_{i\in\mathcal{I}_0} E_i] \le m$. Construct an example where replacing the max-weight condition by an average-weight condition would overweight the true nulls.

Exercise 8.30: Universal inference

Prove Proposition 8.7, identifying exactly where the independence of D_0 and D_1 is used. Then show that the average of the two split e-values obtained by swapping the roles of D_0 and D_1 is again an e-value.

Exercise 8.31: Post-hoc validity

Prove that $\sup_{\alpha\in(0,1)} \mathbf{1}\{E \ge 1/\alpha\}/\alpha \le E$ pointwise, as in Proposition 8.13. Deduce that for every data-dependent level $\widehat{\alpha} \in (0, 1)$, $\mathbb{E}_P[\mathbf{1}\{E \ge 1/\widehat{\alpha}\}/\widehat{\alpha}] \le 1$ under every null P.

Computational.

Exercise 8.32: Optional stopping simulation

Reproduce Figure A.1. Simulate null Gaussian data, repeatedly check a one-sided fixed-sample p-value, and compare the probability of ever crossing 0.05 with the probability that a Gaussian likelihood-ratio e-process crosses 20.

Exercise 8.33: Dependence simulation

Implement BH, BY, and e-BH for equicorrelated one-sided Gaussian tests. Vary the correlation, signal strength, and number of nonnulls. Report FDR and power, and explain why e-BH's validity statement differs from BH's PRDS validity statement.

Advanced.

Exercise 8.34: Confidence sequence by inversion

For each θ_0, let $M_t^{\theta_0}$ be a nonnegative supermartingale under $\mathbb{P}_{\theta_0}$ with $M_0^{\theta_0} = 1$. Prove Theorem A.19. Then explain why simultaneous coverage implies coverage at any stopping time τ.

Exercise 8.35: Universality of e-BH

Prove Proposition A.11 and Theorem A.12. Why is it legitimate to assign zero evidence to the hypotheses the original procedure did not reject, and where would the argument break if a non-rejected hypothesis were assigned a positive induced e-value larger than the e-BH boundary?

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 conditional randomization test (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 negative controls: if a real feature beats its knockoff, that is evidence for the real feature; if knockoffs often beat real features, those negative wins estimate how many selected real features may be false discoveries.

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 a 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.

The conditional formulation is stricter than a marginal association statement. 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 conditional randomization test of Candès et al. [36] 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.

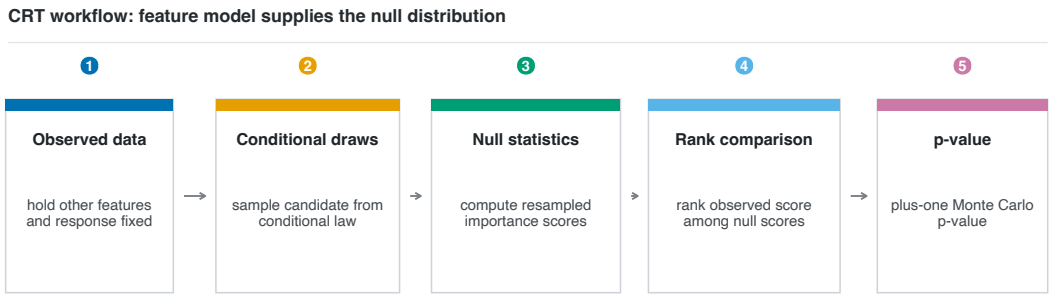

FIGURE 9.1. Conditional randomization test. The conditional feature law, not a model for $Y \mid X$, supplies the null distribution. The final p-value uses the plus-one Monte Carlo correction.

Given B 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 p-value is written for large values of T_j. For two-sided or directional statistics, one replaces the comparison by the appropriate tail event.

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, and that $H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$ holds. Conditional on (X_{-j}, Y), the observed and resampled statistics $T_j^{\text{obs}}, T_j^{(1)}, \ldots, T_j^{(B)}$ 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 $X_j^{(1)}, \ldots, X_j^{(B)}$ come from the same conditional law by construction. Therefore the collection

$$X_j,\ X_j^{(1)}, \ldots, X_j^{(B)}$$

is iid conditional on (X_{-j}, Y), and the induced statistics $T_j^{\text{obs}}, T_j^{(1)}, \ldots, T_j^{(B)}$ 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^{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 B must be large enough that the grid of possible p-values can reach the smallest BH critical 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 ε 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.

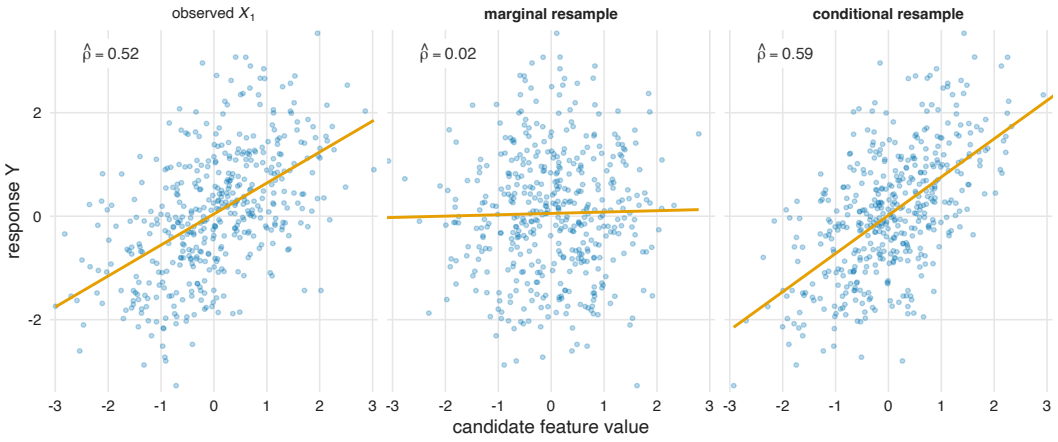

FIGURE 9.2. A conditionally null feature can be marginally correlated with the response. Each panel contains $n = 450$ observations, the data-generating correlation is $\rho = 0.65$, and the labels show empirical $\widehat{\rho}$ values in the plotted samples. Marginal resampling breaks the dependence between X_1 and X_2, whereas conditional resampling preserves it.

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.

4. Fixed-X Knockoffs

Route: Core

We next move from testing one feature at a time to constructing negative controls for all features simultaneously. The original knockoff construction of Barber and Candès [8] 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 negative control 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 $P^\top G P = \mathrm{diag}(2\Sigma - D,\, D)$, 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 Σ 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

$$\min_{0 \le s_j \le 1} \sum_{j=1}^{p} |1 - s_j| \qquad \text{subject to} \qquad 2\Sigma - \mathrm{diag}(s) \succeq 0.$$

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 vector W must satisfy two structural properties.

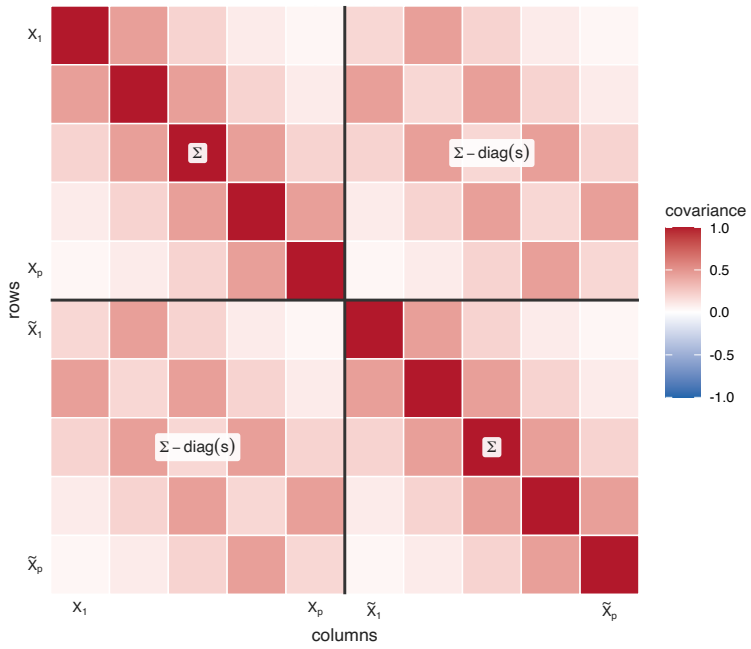

FIGURE 9.3. The knockoff block structure. Original and knockoff variables have matching within-block dependence; the diagonal gap diag(s) makes each knockoff a less-than-perfect copy of its original feature. In the fixed-X setting (§4) this is a Gram matrix; in the model-X setting (§7) it is a population covariance.

Definition 9.5: Sufficiency and antisymmetry

A fixed-X knockoff statistic vector $W = (W_1, \dots, W_p)$ is sufficient if it depends on the data only through

$$[X, \tilde{X}]^\top [X, \tilde{X}], \qquad [X, \tilde{X}]^\top y.$$

It is antisymmetric if, for each coordinate j, swapping X_j and $\tilde{X}_j$ 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 $[X, \tilde{X}]$, and let $\tilde{Z}_j$ be the analogous entry time for $\tilde{X}_j$. Concretely, the lasso solves

$$\min_b \frac{1}{2}\|y - [X, \tilde{X}]b\|_2^2 + \lambda \|b\|_1$$

as λ decreases from a very large value. At very large λ, all coefficients are zero. A column with stronger predictive information tends to become nonzero earlier, meaning at a larger value of λ. Thus Z_j and $\tilde{Z}_j$ record which member of the pair $(X_j, \tilde{X}_j)$ enters first, and how early it enters. Define the signed-max statistic

$$W_j = \max\{Z_j, \tilde{Z}_j\} \operatorname{sign}(Z_j - \tilde{Z}_j).$$

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.

Let

$$\mathcal{W} = \{|W_j| : \ |W_j| > 0, \ j = 1, \dots, p\}.$$

The knockoff+ threshold at target FDR level q is

$$T_+ = \min \left\{ t \in \mathcal{W} : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q \right\},$$

with $T_+ = \infty$ if the set is empty. The selected variables are

$$\widehat{\mathcal{S}} = \{j : W_j \ge T_+\}.$$

A boundary observation: the numerator in the threshold inequality is at least 1, so for the ratio to fall below q we need $\#\{j : W_j \ge t\} \ge 1/q$ 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 is the finite-sample analogue of the no-discovery regime in BH at very stringent FDR targets. The version without the plus one,

$$T = \min \left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q \right\},$$

is often more powerful but controls a slightly modified FDR criterion. The plus-one numerator is what gives ordinary FDR control.

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.

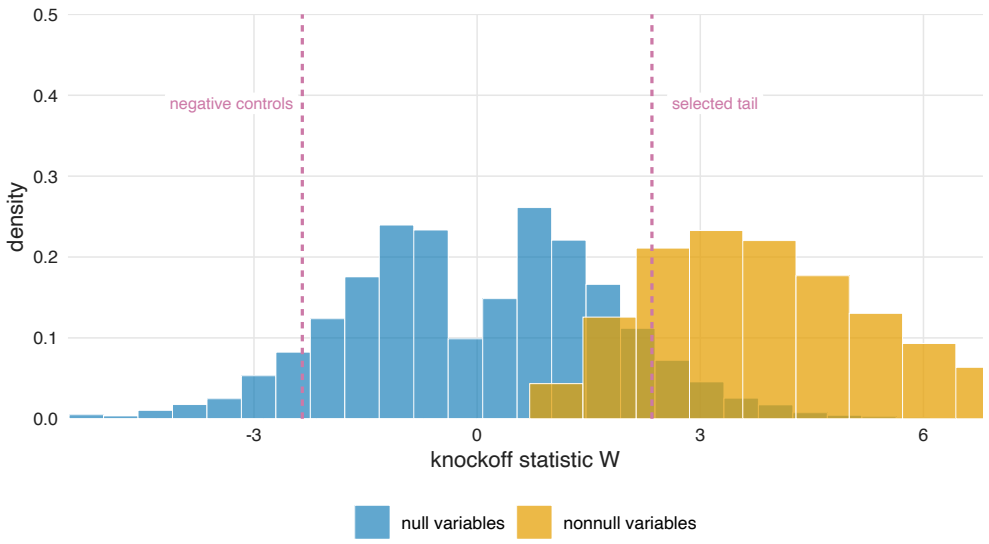

FIGURE 9.4. Knockoff signs. For true null variables, positive and negative W_j's are symmetric; negative knockoff wins therefore estimate the number of false positive original wins. Nonnull variables should shift mass toward the positive tail. This synthetic display uses 6000 null and 900 nonnull statistics; the dashed lines at ± 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

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\{j : \beta_j = 0,\ j \in \widehat{\mathcal{S}}\}}{|\widehat{\mathcal{S}}| \vee 1}\right] \le q.$$

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

$$\mathcal{C} = \sigma(\{|W_j| : 1 \le j \le p\}, \{W_j : j \notin H_0\}),$$

then the signs $\mathrm{sign}(W_j)$, $j \in H_0$ with $|W_j| > 0$, are independent fair coins conditional on $\mathcal{C}$. This is the fixed-X sign-flip result of Barber and Candès [8, Lemma 1]; the preceding paragraph gives the mechanism behind the lemma in this setting.

Stopped ratio. For a threshold t, write

$$V^+(t) = \#\{j : \beta_j = 0,\ W_j \ge t\}, \qquad V^-(t) = \#\{j : \beta_j = 0,\ W_j \le -t\}.$$

After conditioning on $\mathcal{C}$, the knockoff+ rule checks candidate thresholds from the least extreme magnitude to the most extreme. Each failed check raises the threshold and removes the least extreme active statistic; the nonnull statistics and all magnitudes are already fixed. The reverse-filtration stopping-time calculation given below therefore gives

$$\mathbb{E}\left[\frac{V^+(T_+)}{1 + V^-(T_+)} \,\middle|\, \mathcal{C}\right] \le 1,$$

and hence the same inequality holds after averaging over $\mathcal{C}$.

FDR bound. The observable knockoff+ stopping rule and $V^-(T_+) \le \#\{j : W_j \le -T_+\}$ give the pointwise bound

$$\begin{aligned}
\mathrm{FDP} &= \frac{V^+(T_+)}{|\widehat{\mathcal{S}}| \vee 1} \\
&\le \frac{1 + \#\{j : W_j \le -T_+\}}{|\widehat{\mathcal{S}}| \vee 1} \cdot \frac{V^+(T_+)}{1 + V^-(T_+)} \\
&\le q \cdot \frac{V^+(T_+)}{1 + V^-(T_+)}.
\end{aligned}$$

Taking expectations and applying the stopped sign-count calculation yields 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,

$$|W_{j_1}| \geq \cdots \geq |W_{j_m}| > 0,$$

and write $\xi_r = \operatorname{sign}(W_{j_r})$. By sign symmetry, the ξ_r's are iid fair coins. Define the active positive and negative counts and their ratio by

$$P_k = \#\{r \leq k : \xi_r = +1\}, \qquad Q_k = \#\{r \leq k : \xi_r = -1\}, \qquad M_k = \frac{P_k}{1+Q_k}, \qquad 0 \leq k \leq m.$$

The relevant reverse filtration is

$$\mathcal{F}_k = \sigma(\mathcal{C},\, P_k,\, Q_k,\, \xi_{k+1}, \ldots, \xi_m)\,, \qquad k = m, m-1, \ldots, 0.$$

As k decreases, information increases: $\mathcal{F}_k \subseteq \mathcal{F}_{k-1}$. Conditional on $\mathcal{F}_k$, 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 $M_{k-1} = k-1 \leq k = M_k$. Thus $(M_k, \mathcal{F}_k)$, read in the order $m, m-1, \ldots, 0$, is a supermartingale.

It remains to verify that the knockoff threshold is a valid stopping rule for this reverse scan. Order all candidate magnitudes from smallest to largest. At the first candidate, every nonzero statistic is active; after a failed check, increase the threshold and remove the least extreme active statistic or tied block. Conditional on $\mathcal{C}$, the nonnull contributions to the two tail counts are fixed, while the active null contributions are exactly P_k and Q_k. Therefore the decision whether the observable FDP estimate has crossed below q is measurable with the current reverse filtration. If magnitude ties occur, refine each tied block by an ordering independent of the signs and permit stopping only at the end of the block. Nonnull removals are deterministic no-change steps for M_k. Compress the candidate scan to its null-removal clock, retaining every intervening deterministic nonnull removal and threshold check. Let ρ be the number of null removals completed when the first passing candidate is encountered (or m if no candidate passes). Because the nonnull contributions and the locations of all candidate magnitudes are fixed by $\mathcal{C}$, whether the scan has stopped by the rth null removal is measurable with $\mathcal{G}_r = \mathcal{F}_{m-r}$. Thus ρ is a bounded stopping time for $(\mathcal{G}_r)_{r=0}^m$.

Let $\kappa = m - \rho$ be the number of active nulls at the stopped threshold. Then

$$M_\kappa = \frac{V^+(T_+)}{1+V^-(T_+)}.$$

Finite optional stopping, now applied only after boundedness and the stopping property have been checked, gives

$$\mathbb{E}(M_\kappa \mid \mathcal{C}) \leq \mathbb{E}(M_m \mid \mathcal{C}).$$

Finally,

$$M_m = \frac{B}{1+m-B}, \qquad B \mid \mathcal{C} \sim \text{Binomial}(m, 1/2), \qquad \mathbb{E}(M_m \mid \mathcal{C}) = 1 - 2^{-m} \leq 1.$$

This proves the stopped-ratio bound used above.

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 [36].

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 jth 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. Knockoffs are coordinatewise negative controls, not arbitrary fake features.

Lemma 9.8: Null-swap property

Let $\widetilde{X}$ be a model-X knockoff copy of X. If

$$Y \perp\!\!\!\perp X_S \mid X_{-S},$$

then

$$([X, \widetilde{X}], Y) \overset{d}{=} ([X, \widetilde{X}]_{\mathrm{swap}(S)}, Y).$$

Proof of Lemma 9.8

Two ingredients are needed: pairwise exchangeability of $[X, \widetilde{X}]$ and the response-independence condition $\widetilde{X} \perp\!\!\!\perp Y \mid X$ from the model-X knockoff definition.
Pairwise exchangeability gives the equality in distribution for $[X, \widetilde{X}]$ before looking at Y. It remains to see that the conditional response law is unchanged by a null swap. For a realized pair $(x, \widetilde{x})$, let x^{sw} be the feature vector obtained after the swap, so $x_j^{\mathrm{sw}} = \widetilde{x}_j$ for $j \in S$ and $x_j^{\mathrm{sw}} = x_j$ otherwise; define $\widetilde{x}^{\mathrm{sw}}$ analogously. Then

$$\begin{aligned}\mathcal{L}\Big(Y \mid [X, \widetilde{X}]_{\mathrm{swap}(S)} = (x, \widetilde{x})\Big) &= \mathcal{L}\Big(Y \mid X = x^{\mathrm{sw}}, \widetilde{X} = \widetilde{x}^{\mathrm{sw}}\Big) \\ &= \mathcal{L}(Y \mid X = x^{\mathrm{sw}}) \\ &= \mathcal{L}(Y \mid X_{-S} = x_{-S}) \\ &= \mathcal{L}(Y \mid X = x) = \mathcal{L}\Big(Y \mid X = x, \widetilde{X} = \widetilde{x}\Big).\end{aligned}$$

The second and last equalities use $\widetilde{X} \perp\!\!\!\perp Y \mid X$. The middle equalities use $Y \perp\!\!\!\perp X_S \mid X_{-S}$, together with the fact that swapping coordinates in S leaves X_{-S} unchanged. Combining this conditional response invariance with pairwise exchangeability of the features gives the joint equality. □

The lemma assumes a joint null for the whole swapped set S. In many regular feature models this joint null follows from the coordinatewise nulls, but there is a small point worth making explicit. Suppose, for example, that $S = \{a, b\}$ and write $C = X_{-\{a,b\}}$. Define the conditional response kernel

$$K_c(x_a, x_b) = \mathcal{L}(Y \in \cdot \mid C = c, X_a = x_a, X_b = x_b).$$

The individual null $Y \perp\!\!\!\perp X_a \mid X_{-a}$ says that, for fixed (c, x_b), this kernel does not change as x_a varies. The individual null $Y \perp\!\!\!\perp X_b \mid X_{-b}$ says that, for fixed (c, x_a), it does not change as x_b varies. If the feature law has positive density—or positive mass in the discrete case—on a product support, so that the conditional law of $(X_a, X_b) \mid C = c$ also has product support for almost every c, then these comparisons can be made across the whole rectangle of possible (x_a, x_b) values. Hence

$$K_c(x_a, x_b) = K_c(x_a', x_b) = K_c(x_a', x_b')$$

for any two pairs (x_a, x_b) and (x_a', x_b') in the support. Thus the conditional law of Y depends only on $C = X_{-\{a,b\}}$, which is exactly $Y \perp\!\!\!\perp (X_a, X_b) \mid X_{-\{a,b\}}$. This two-coordinate argument is the intersection property for conditional independence; iterating it turns $\{Y \perp\!\!\!\perp X_j \mid X_{-j} : j \in S\}$ into $Y \perp\!\!\!\perp X_S \mid X_{-S}$. Degenerate feature laws, such as perfect collinearity, can break the rectangle of comparisons; in such cases the subset swap statement should be read as requiring the displayed joint null directly.

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

$$(Z, \widetilde{Z}) = (Z_1, \ldots, Z_p, \widetilde{Z}_1, \ldots, \widetilde{Z}_p).$$

The score map is swap-equivariant if, for every coordinate set $S \subseteq \{1, \ldots, p\}$, 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

$$(W_j : j \in \mathcal{I}_0) \overset{d}{=} (\varepsilon_j W_j : j \in \mathcal{I}_0)$$

for every deterministic sign vector $\varepsilon_j \in \{-1, 1\}$. Consequently, conditional on the magnitudes $\{|W_j| : j \in \mathcal{I}_0\}$ and on 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. The observable knockoff+ FDP estimate is therefore

$$\widehat{\mathrm{FDP}}(t) = \frac{1 + \#\{j : W_j \leq -t\}}{R(t) \vee 1}.$$

Let

$$\mathcal{W} = \{|W_j| : \ |W_j| > 0, \ j = 1, \ldots, p\}$$

and define

$$\tau_q = \min\{t \in \mathcal{W} : \widehat{\mathrm{FDP}}(t) \leq q\},$$

with $\tau_q = \infty$ if the set is empty. The model-X knockoff+ selected set is

$$\widehat{\mathcal{S}} = \{j : W_j \geq \tau_q\}.$$

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. Assume also that the conditional law is regular enough that individual nulls combine into joint subset nulls, i.e. for every subset S of true nulls, $Y \perp\!\!\!\perp X_S \mid X_{-S}$. 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,

$$\mathrm{FDP} = \frac{V^+(\tau_q)}{R(\tau_q) \vee 1}.$$

Since $V^-(\tau_q) \le \#\{j : W_j \le -\tau_q\}$, the definition of τ_q gives the pointwise bound

$$\begin{aligned}
\mathrm{FDP} &= \frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \frac{1 + V^-(\tau_q)}{R(\tau_q) \vee 1} \\
&\le \frac{V^+(\tau_q)}{1 + V^-(\tau_q)} \frac{1 + \#\{j : W_j \le -\tau_q\}}{R(\tau_q) \vee 1} \\
&\le q \, \frac{V^+(\tau_q)}{1 + V^-(\tau_q)}.
\end{aligned}$$

It remains to show that the expectation of the final ratio is at most one. Condition on

$$\mathcal{C} = \sigma(\{|W_j| : 1 \le j \le p\}, \{W_j : j \notin \mathcal{I}_0\}).$$

By Lemma 9.9, conditional on $\mathcal{C}$, the signs of the nonzero null statistics are independent fair signs. The threshold τ_q is found by checking magnitudes from least to most extreme, removing the least extreme active statistic after each failed check. With the nonnull statistics and all magnitudes fixed, the scan 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 $\mathrm{FDR} \le q$. □

Remark 9.11: The threshold is not the filter

The two-count ratio defining τ_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 [181]. 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. [36], when

$$X \sim N(0, \Sigma),$$

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 Σ. 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 control 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.

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,

$$\min_{0 \le s_j \le 1} \sum_{j=1}^{p} |1 - s_j| \qquad \text{subject to} \qquad 2\Sigma - \mathrm{diag}(s) \succeq 0,$$

tries to make each original/knockoff pair as different as the covariance structure allows.

8. Sequential Conditional Independent Pairs (SCIP) and Structured Feature Laws

Route: Advanced

Gaussian knockoffs are only one example. SCIP 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 $\widetilde{X} \stackrel{d}{=} X$ already has the required conditional exchangeability.

For a general known feature distribution, the sequential conditional independent pairs (SCIP) construction of Candès et al. [36] 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

$$\mathbb{P}(\widetilde{X}_j = \widetilde{x}_j \mid X_{1:p} = x_{1:p}, \widetilde{X}_{1:j-1} = \widetilde{x}_{1:j-1}) = \frac{P_{j-1}(x_{-j}, \widetilde{x}_j, \widetilde{x}_{1:j-1})}{\sum_u P_{j-1}(x_{-j}, u, \widetilde{x}_{1:j-1})}.$$

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 $\{-1, 1\}^p$,

$$\mathbb{P}(X = x) \propto \exp\left\{ \sum_{k<\ell} \omega_{k\ell} x_k x_\ell + \sum_{k=1}^{p} h_k x_k \right\},$$

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 π_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. [141] carry this out for hidden Markov models in genome-wide association studies, where linkage disequilibrium 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 cross-validated predictive model, this can be prohibitive. Several variants reduce this cost: Liu et al. [115] 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. Read the left panel as a power check for that particular design and the right panel as an evaluation-count comparison: it illustrates amortization, not a universal wall-clock ranking.

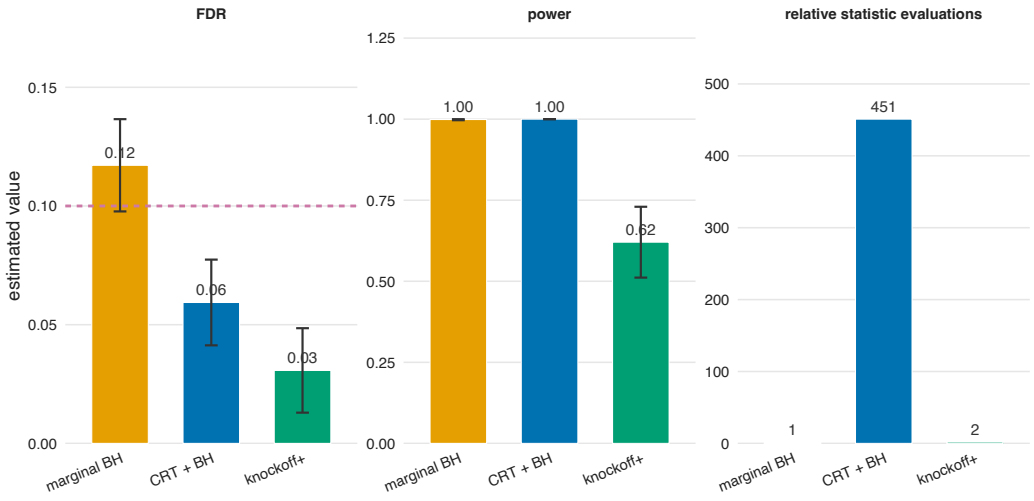

FIGURE 9.5. A Gaussian model-X simulation with $n = 220$, $p = 40$, 14 signals of size 1, AR(1) correlation $\rho = 0.5$, and FDR target $q = 0.10$. The plotted FDR and power averages use 64 independent data sets; their vertical bars are pointwise 95 percent normal Monte Carlo intervals. CRT uses $B = 450$ 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.

Figure 9.5 should not be read as a universal ranking of methods. Power depends heavily on the signal structure, the feature correlations, the statistic, and the number of hypotheses. The stable lesson is computational: CRT is the exact one-feature-at-a-time benchmark, while knockoffs amortize conditional information across the whole family.

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 (141), a copula model, a normalizing flow, or a deep generative model trained to match the swap-invariance condition (137), and then sample knockoffs from the fitted distribution. Approximate knockoffs should be treated as a modeling step, not as a theorem.

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 θ 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, but the FDR theorem applies only to the extent that the learned generator gives sufficiently accurate pairwise exchangeability.

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 diagnostics do not by themselves prove FDR control, but they reveal the failure modes that matter: 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. [10] give a quantitative robustness result: the FDR is bounded by q plus a term that depends on a Kullback–Leibler (KL) divergence between the true and approximating feature distributions, so small modeling errors translate into small FDR inflation.

10. Multilayer Knockoffs

Route: Research frontier

The model-X knockoff filter also extends to structured families of variables. Katsevich and Sabatti [95] 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 [95].

Why another layer changes the target. Let $R \subseteq \{1, \dots, p\}$ be a selected set of variables. A layer ℓ 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-ℓ group containing variable j. The variable set R induces the layer-ℓ 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 $V_{\mathrm{ind}}(R)$ be the number of selected null variables and let $V_{\mathrm{grp}}(R)$ be the number of selected null groups. The corresponding realized false discovery proportions are

$$\mathrm{FDP}_{\mathrm{ind}}(R) = \frac{V_{\mathrm{ind}}(R)}{|R| \vee 1}, \qquad \mathrm{FDP}_{\mathrm{grp}}(R) = \frac{V_{\mathrm{grp}}(R)}{|\mathcal{R}_{\mathrm{grp}}(R)| \vee 1}.$$

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

$$\mathrm{FDP}_{\mathrm{ind}}(R) = \frac{2}{3}.$$

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-ℓ group g is a *group null* if all variables in that group are null at the finest level:

$$g \in \mathcal{N}_\ell \quad \Longleftrightarrow \quad H_j : Y \perp\!\!\!\perp X_j \mid X_{-j} \text{ is true for every } j \in \mathcal{G}_g^{(\ell)}.$$

Equivalently, under the same positivity/intersection regularity discussed after Lemma 9.8, this says

$$Y \perp\!\!\!\perp X_{\mathcal{G}_g^{(\ell)}} \mid X_{-\mathcal{G}_g^{(\ell)}}.$$

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 multilayer knockoff filter requires layer-ℓ 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. This property is not automatic for an arbitrary aggregate of variable-level statistics. Simply summing the W_j's inside a group, for example, need not produce a valid group statistic, because the resulting sign may not flip cleanly when a null group is swapped. 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 [95].

Joint FDP estimate and joint thresholds. 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-ℓ discoveries are

$$\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau}) \;=\; \mathcal{R}_\ell(\widehat{R}(\boldsymbol{\tau})).$$

Let $\alpha_1, \dots, \alpha_L$ 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 $c = c_{\mathrm{kn}}$, with $c_{\mathrm{kn}} = 1.93$ in the theorem below, gives the formal worst-case guarantee. The layer-ℓ FDP estimate has the form

$$\widehat{\mathrm{FDP}}^{(\ell)}(\boldsymbol{\tau}) = \frac{c\{1 + \#\{g : W_g^{(\ell)} \leq -\tau_\ell\}\}}{|\widehat{\mathcal{R}}^{(\ell)}(\boldsymbol{\tau})| \vee 1}.$$

The numerator is layer-specific: negative group wins at layer ℓ 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 $\widehat{\mathrm{FDP}}^{(\ell)}(\boldsymbol{\tau}^*) \leq \alpha_\ell$ 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 [95] show that this threshold vector is well defined and can be found by an iterative search over the threshold grid. 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 [95]

Assume the model-X exchangeability conditions of this chapter and suppose each layer-ℓ group statistic $W_g^{(\ell)}$ satisfies the formal flip-sign and compatibility conditions of Katsevich and Sabatti [95]. For the MKF(c)+ procedure of that paper,

$$\mathrm{FDR}^{(\ell)}(\widehat{R}) \;\leq\; \frac{c_{\mathrm{kn}}}{c}\,\alpha_\ell, \qquad c_{\mathrm{kn}} = 1.93, \qquad \text{for every layer } \ell.$$

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 $\mathrm{FDR}^{(\ell)} \leq \alpha_\ell$.

Proof idea. Fix a layer ℓ, and let

$$V_\ell^+(t) = \#\{g \in \mathcal{N}_\ell : W_g^{(\ell)} \geq t\}, \qquad V_\ell^-(t) = \#\{g \in \mathcal{N}_\ell : W_g^{(\ell)} \leq -t\}.$$

At the selected threshold $\boldsymbol{\tau}^*$, the number of false layer-ℓ 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,

$$\mathrm{FDP}^{(\ell)}(\widehat{R}) \leq \frac{\alpha_\ell}{c}\,\frac{V_\ell^+(\tau_\ell^*)}{1 + V_\ell^-(\tau_\ell^*)}.$$

The difficulty is that τ_ℓ^* 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-ℓ threshold:

$$\mathbb{E}\left[\frac{V_\ell^+(\tau_\ell^*)}{1 + V_\ell^-(\tau_\ell^*)}\right] \leq \mathbb{E}\left[\sup_{t \geq 0} \frac{V_\ell^+(t)}{1 + V_\ell^-(t)}\right].$$

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-ℓ null signs matter. Conditional on magnitudes and nonnull group statistics, order the 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 [95] is at most $c_{\mathrm{kn}} = 1.93$,

giving the theorem. Using $c = c_{\text{kn}}$ converts this constant-factor bound into target-level FDR control.

Simulations in Katsevich and Sabatti [95] suggest that the uncorrected version $c = 1$ often controls FDR in practice. The constant c_{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 9.25 gives an explicit counterexample.

When variables within a group are tightly correlated (linkage-disequilibrium 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

Assumption diagnostic	
Checkable	Use held-out feature-model checks, swap diagnostics, Gram-matrix verification, and negative-control outcomes.
Not identified from these data	Exact conditional feature laws and conditional independence nulls are not generally testable from the same finite sample.
Sensitivity analysis	Repeat across plausible feature models and report robust/approximate-knockoff sensitivity.
Conservative fallback	Use fixed-X inference when its linear model is credible, or report prediction rather than conditional discovery.

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.

The CRT is exact when the conditional law $\mathcal{L}(X_j \mid X_{-j})$ is correct and the resampling is performed conditionally on the observed X_{-j}. The statistic can be arbitrarily complicated, but the conditional resampling law cannot be wrong without changing the null distribution.

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. Approximate knockoffs, fit by graphical models, normalizing flows, or deep generators, are a modeling choice, not a theorem; robustness to approximation is studied by Barber et al. [10] and related generator-based constructions by Romano et al. [137].

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 [95]. 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 [8]. Model-X knockoffs, the CRT viewpoint, Gaussian model-X construction, and SCIP were developed by Candès et al. [36]. Robustness to approximate knockoffs is studied by Barber et al. [10]. 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 [181]. Metropolized knockoff sampling provides one route to exact or controlled sampling for complex distributions [16]. Hidden Markov model knockoffs for genome-wide association studies are due to Sesia et al. [141], and deep generative knockoffs are developed by Romano et al. [137]. Computationally efficient CRT variants, including the distillation CRT used to amortize statistic evaluations across features, are studied by Liu et al. [115]. A frontier extension to privacy-constrained inference is the differentially private model-X knockoff filter of Pournaderi and Xiang [128], 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 [95].

13. Exercises

Route: Core

Basic.

Exercise 9.14: The CRT p-value

Write the CRT algorithm for testing $H_j : Y \perp\!\!\!\perp X_j \mid X_{-j}$. Prove that the plus-one Monte Carlo p-value is super-uniform when the conditional law $X_j \mid X_{-j}$ is known exactly.

Exercise 9.15: Conditional resampling

In Example 9.3, compute $\mathrm{Cov}(X_1, Y)$ and verify that $Y \perp\!\!\!\perp X_1 \mid X_2$. Derive the conditional law $X_1 \mid X_2 = x_2$, and explain why marginal resampling of X_1 is invalid.

Exercise 9.16: Pairwise exchangeability

State pairwise exchangeability for a model-X knockoff copy $\widetilde{X}$. For $p = 2$, write out explicitly the four swap identities corresponding to $S = \varnothing, \{1\}, \{2\}, \{1, 2\}$.

Intermediate.

Exercise 9.17: Fixed-X Gram verification

In the notation of Proposition 9.4, let $\Sigma = X^\top X$, $D = \mathrm{diag}(s)$, and

$$\widetilde{X} = X(I - \Sigma^{-1}D) + UC, \qquad C^\top C = 2D - D\Sigma^{-1}D,$$

where X has full column rank, $U^\top X = 0$, and $U^\top U = I_p$. Verify directly that

$$X^\top \widetilde{X} = \Sigma - D, \qquad \widetilde{X}^\top \widetilde{X} = \Sigma.$$

Exercise 9.18: Feasibility of s

Show that the condition $2\Sigma - D \succeq 0$ is equivalent to the positive semidefiniteness of $2D - D\Sigma^{-1}D$ when D is positive definite. How should the argument be modified when some $s_j = 0$?

Exercise 9.19: Signed-max statistic

Let Z_j and $\widetilde{Z}_j$ be lasso entry times for the original and knockoff variables. Show that

$$W_j = \max\{Z_j, \widetilde{Z}_j\}\,\mathrm{sign}(Z_j - \widetilde{Z}_j)$$

is antisymmetric under swapping X_j and $\widetilde{X}_j$.

Exercise 9.20: 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 ∞ if no candidate value satisfies the threshold inequality. Explain why the plus-one version can make no discoveries when the number of strong positive statistics is too small relative to $1/q$.

Exercise 9.21: Gaussian model-X conditionals

Assume $\Sigma \succ 0$. Starting from the block covariance

$$\begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix},$$

derive the conditional distribution

$$\widetilde{X} \mid X = x \sim N\left((I - D\Sigma^{-1})x,\ 2D - D\Sigma^{-1}D\right).$$

Use the Gaussian conditioning formula and identify the conditional mean and conditional covariance separately.

Computational.

Exercise 9.22: 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.

Exercise 9.23: Gaussian knockoff implementation

Write a function that generates Gaussian model-X knockoffs for $X \sim N(0, \Sigma)$. For a concrete check, take 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. Then simulate a sparse linear model, apply the knockoff+ threshold at $q = 0.1$, and report empirical FDR and power.

Exercise 9.24: Fixed-X versus model-X targets under misspecification

Generate data with a strongly nonlinear response, say $Y = \sin(2X_k) + \varepsilon$, and a null feature X_j strongly correlated with the nonnull X_k. Apply (a) fixed-X knockoffs in a linear model and (b) Gaussian model-X knockoffs. Treat the fixed-X analysis here as a working linear-model analysis under misspecification, not as an invocation of the fixed-X FDR theorem for the nonlinear data-generating mechanism. Compare the rejection sets and explain why the fixed-X null ($\beta_j = 0$ in the working linear projection) and the model-X null $Y \perp X_j \mid X_{-j}$ can disagree when the response model is wrong. Verify that under a correctly specified Gaussian linear model with random design, the two nulls coincide for every coordinate.

Advanced.

Exercise 9.25: Multilayer knockoff intersection – a fallacy

This is a stylized rejection-set counterexample; constructing actual knockoff statistics with this exact behavior is not required. Specify a probability model in which two separately valid layer-wise FDR procedures control their marginal FDRs at level α, but the intersection of the variable-level rejection set with the lifted group-level rejection set has expected variable-level FDP exceeding α. A workable design: 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 π with $\alpha < \pi < 2\alpha$, e.g. $\pi = 0.08$. Arrange the joint distribution of the variable-level and group-level procedures so that exactly one of the following two events occurs in every realization:

- Event A (with probability π): 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 marginal layer-1 FDR, the marginal layer-2 FDR, and the variable-level FDR of the intersection. Verify that the two separate procedures are each calibrated at level α, while the intersected output is not. Use the example to explain why the multilayer knockoff filter chooses one joint threshold vector rather than running separate filters and intersecting their lifted outputs.

Exercise 9.26: 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 $[X, \widetilde{X}]_{\mathrm{swap}(\{j\})}$ 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 cross-validation. 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.** The rank argument, full and split conformal prediction, conformal p-values, and conformalized quantile regression (CQR) give the main finite-sample prediction and testing tools.
- **Advanced route.** Jackknife+, aggregation, and conformal e-prediction develop more data-efficient or sequential variants after the core construction is in place.
- **Appendix B route.** Appendix B treats conditional and selection-conditional coverage (Section 3), online conformal ranks and betting (Section 5), weighting and drift (Section 6), and conformal risk control (CRC) together with Learn–Then–Test (LTT) (Section 7).

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. 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.

Throughout this chapter, the core assumption is exchangeability:

$$(Z_1, \dots, Z_{n+1}) \stackrel{d}{=} (Z_{\pi(1)}, \dots, Z_{\pi(n+1)})$$

for every permutation π. Independent identically distributed data imply exchangeability, but exchangeability is the property used by the proofs.

It is helpful to keep the main guarantees separate:

- Split conformal and CQR give marginal coverage from exchangeable calibration and test scores; CQR uses a better score to adapt interval width.
- Under exchangeability and a symmetric fitting rule, jackknife+ gives a $1-2\alpha$ guarantee without a stability assumption. Majority voting gives the same guarantee from valid base sets without assuming that those sets are independent.
- Weighted conformal changes the calibration weights when the test law is known to differ from the calibration law.
- Conformal p-values, combined with BH, control FDR for many outlier tests when the test points are mutually independent and independent of the shared reference sample, and the true-null score distribution is continuous; under these conditions the conformal p-values are PRDS.
- Online full conformal prediction with a symmetric score and almost surely distinct scores produces exactly independent p-values, and those p-values power an anytime-valid test of exchangeability itself.
- CRC and LTT replace miscoverage by broader risk guarantees, using either monotone losses or valid calibration tests.

The rest of the chapter explains what each line means and, just as important, what it does not mean.

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 rests on a general principle that is worth isolating, because it is the engine behind several later results in this chapter. 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*

$$\widehat{P}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Z_i},$$

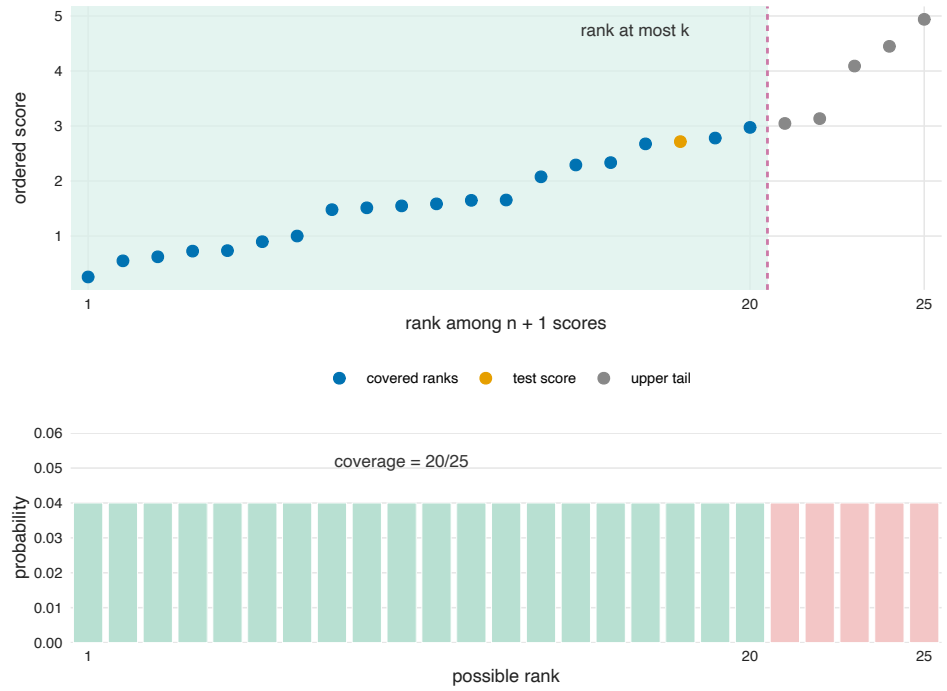

FIGURE 10.1. The conformal rank argument. Under exchangeability, the test score rank among the $n+1$ scores is uniform. The finite-sample correction uses $k = \lceil (1-\alpha)(n+1) \rceil$, not simply the empirical $1-\alpha$ quantile of the first n scores. The cutoff is denoted c_n. The displayed example takes $n = 24$ and $\alpha = 0.2$, giving $k = 20$ and coverage exactly $20/25$.

where δ_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 σ of $\{1, \dots, m\}$; the empirical distribution, the order statistics of real-valued data, and any sample average are examples.

Lemma 10.1: Conditioning on a symmetric statistic

Let $(Z_1, \dots, Z_m)$ be exchangeable and let f be a permutation-symmetric function of the data. Then the conditional law of $(Z_1, \dots, Z_m)$ given $f(Z_1, \dots, Z_m)$ is, almost surely, an exchangeable distribution. In particular, conditional on the empirical distribution $\widehat{P}_m$, each coordinate is a draw from $\widehat{P}_m$:

$$Z_i \mid \widehat{P}_m \;\sim\; \widehat{P}_m, \qquad i = 1, \dots, m,$$

so that each Z_i is uniformly distributed over the observed values when those values are distinct [4].

Proof of Lemma 10.1

Write $Z = (Z_1, \dots, Z_m)$ and, for a permutation σ, write $Z_\sigma = (Z_{\sigma(1)}, \dots, Z_{\sigma(m)})$. For the first claim, it suffices to show that, for every σ 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_σ 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 continuous 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 kth 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 [172]. For regression, a common choice is

$$S(x, y; \widehat{f}) = |y - \widehat{f}(x)|.$$

Smaller scores mean better conformity. In full conformal prediction, for each candidate value y at the test covariate X_{n+1}, one augments the data with (X_{n+1}, y), fits or refits the prediction rule in a symmetric way, and computes scores

$$R_1(y), \ldots, R_n(y), R_{n+1}(y).$$

The conformal p-value for the candidate response is

$$p_y = \frac{1 + \sum_{i=1}^{n} \mathbf{1}\{R_i(y) \geq R_{n+1}(y)\}}{n+1},$$

for nonconformity scores where larger is worse. The full conformal prediction set is

$$C_n(x) = \{y : p_y > \alpha\}.$$

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. [106], 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 $\widehat{f}_{n_1}$ using only D_1. For $i \in D_2$, compute calibration residuals

$$R_i = |Y_i - \widehat{f}_{n_1}(X_i)|.$$

Let

$$k = \lceil (1-\alpha)(n_2+1) \rceil$$

and let

$$\widehat{c} = R_{(k)}$$

be the kth 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 $\widehat{c}$. 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 $45/50 = 0.9$.

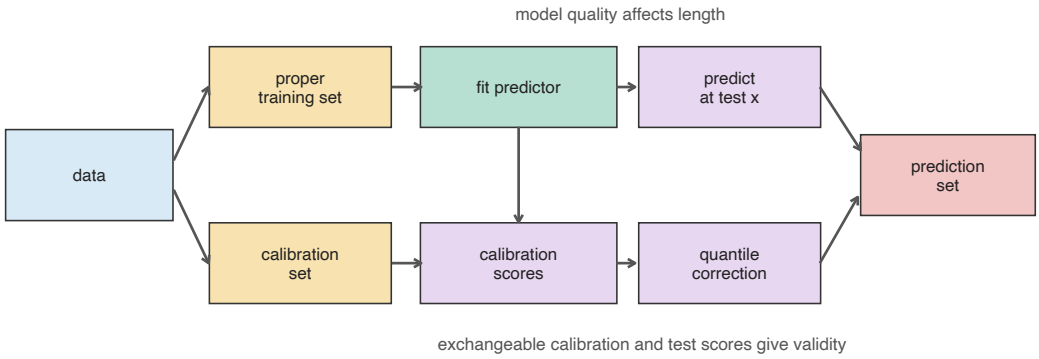

FIGURE 10.2. Split conformal prediction. The fitted predictor is trained on the proper training set; validity comes from the exchangeability of the calibration scores and the future test score conditional on that fitted predictor. Their corrected empirical cutoff is $\widehat{c}$.

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

Split conformal guarantees marginal coverage over a future exchangeable test point and, conditionally, over the proper training split. It does not generally give distribution-free pointwise coverage at every covariate value. Prespecified group coverage, approximate conditional coverage over sets of minimum mass, and model-assisted local calibration are possible relaxations. The Beta-law training-conditional analysis, impossibility theorem, and relaxation bounds are developed in Appendix B.

6. Randomization and Ties

Route: Core

The ordinary conformal set is conservative because the coverage can only move on the grid $\{1/(n_2 + 1), \ldots, 1\}$, and ties make it more conservative. Auxiliary randomization can produce exact coverage.

Let W have distribution function F, and let $U \sim \mathrm{Unif}(0,1)$ independent of 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 atoms, the uniform randomization spreads the 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. The randomized conformal set can be written as

$$C_n^*(x) = \Big\{ y : \frac{1}{n_2+1} \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) < S(x,y)\} + \frac{U}{n_2+1} \left(1 + \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) = S(x,y)\} \right) \le 1 - \alpha \Big\}.$$

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 [4, 172].

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_τ 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_τ; hence

$$K < m\tau \le K + L.$$

Next, locate S_m relative to c_τ. There are three cases.
If $S_m < c_\tau$, then every score $\le S_m$ is strictly below c_τ, 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 using Lemma 10.1. Given the multiset, the quantities c_τ, 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.

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

$$Z_1, \dots, Z_n \sim P$$

be a reference sample, and let S be a conformity score where larger values mean better conformity. 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 $Z_{n+1} \sim P$ 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 not merely analogous to a classical randomization p-value; it is one. 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 σ 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_{perm} is super-uniform: $\mathbb{P}(p_{\text{perm}} \le t) \le t$ for every $t \in [0,1]$ [4].

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

$$p_{\text{perm}} = p(z).$$

Thus the conformal p-value is exactly the permutation-test p-value, based on the rank of the test score, for the null hypothesis that z is exchangeable with the reference sample; in particular, it is super-uniform under that null, with or without ties.

Proof of Proposition 10.5

For this statistic, the indicator inside the sum is $\mathbf{1}\{-S(Z_{\sigma(n+1)}) \ge -S(Z_{n+1})\} = \mathbf{1}\{S(Z_{\sigma(n+1)}) \le S(Z_{n+1})\}$, which depends on σ only through the index $i = \sigma(n+1)$ sent to position $n+1$. Each $i \in \{1, \dots, n+1\}$ arises from exactly $n!$ permutations. Grouping the $(n+1)!$ permutations by i,

$$p_{\text{perm}} = \frac{1}{(n+1)!} \sum_{i=1}^{n+1} n!\, \mathbf{1}\{S(Z_i) \le S(Z_{n+1})\} = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{S(Z_i) \le S(Z_{n+1})\}.$$

The term $i = n+1$ contributes 1, and the remaining n terms count the reference scores that are at most $S(z)$. Hence $p_{\text{perm}} = \{1 + \#\{1 \le i \le n : S(Z_i) \le S(z)\}\}/(n+1) = p(z)$. □

This identification places conformal p-values in a family the book meets in several places: p-values obtained by comparing a statistic against its orbit under a group of transformations that leaves the null distribution of the data invariant. The conditional randomization test of Chapter 9 has the same structure, with independently resampled copies of a covariate playing the role of the permuted data. In both cases, validity does not come from a model for the data-generating process; it comes from the super-uniformity of the rank of the observed statistic within an exchangeable collection. Conformal validity is one more instance of the super-uniformity of rank statistics under a group invariance [4, 172]. When the group is too large to enumerate, the permutation sum is approximated by Monte Carlo sampling of random 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 [127].

For the absolute-residual score

$$S(x, y) = -|y - \widehat{\mu}(x)|,$$

where $\widehat{\mu}$ is trained on data independent of the reference sample, the inequality $S(Z_i) \le S(z)$ is $-|Y_i - \widehat{\mu}(X_i)| \le -|y - \widehat{\mu}(x)|$, which flips under the negation to $|y - \widehat{\mu}(x)| \le |Y_i - \widehat{\mu}(X_i)|$. The p-value therefore becomes

$$p(x, y) = \frac{1 + \#\{1 \le i \le n : |y - \widehat{\mu}(x)| \le |Y_i - \widehat{\mu}(X_i)|\}}{n + 1}.$$

The split conformal exclusion event is essentially a small conformal p-value: if the candidate response lies outside the conformal interval, then its residual is larger than the calibrated residual quantile, and the corresponding conformal p-value is at most α, up to the finite grid.

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

$$p_j = p(Z_{n+j}), \qquad j = 1, \dots, m.$$

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. [17].

Simultaneous coverage for many predictions. 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 $S'_1, \dots, S'_m$ be the future test scores, independent of the calibration scores, and let $F = F_{D_1}$ be their common score CDF. Conditional on $\widehat{c}$ and D_1,

$$\mathbb{P}(S'_1 \le \widehat{c}, \dots, S'_m \le \widehat{c} \mid \widehat{c}, D_1) = F(\widehat{c})^m.$$

Therefore, conditional on D_1,

$$\mathbb{P}(S'_1 \le \widehat{c}, \dots, S'_m \le \widehat{c} \mid D_1) = \mathbb{E}[F(\widehat{c})^m \mid D_1] \ge \{\mathbb{E}[F(\widehat{c}) \mid D_1]\}^m \ge (1 - \alpha)^m,$$

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 the joint coverage of the m prediction sets converges to $(1 - \alpha)^m$ as $n_2 \to \infty$ [4]. 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 $\mathbb{E}[F(\widehat{c})^m \mid D_1] \to (1 - \alpha)^m$. Thus familywise control

$$\mathbb{P}\{\text{all } m \text{ prediction sets cover}\} \ge 1 - \alpha_{\text{FWER}}$$

requires a Bonferroni-scale choice

$$\alpha \le 1 - (1 - \alpha_{\mathrm{FWER}})^{1/m} \approx \frac{\alpha_{\mathrm{FWER}}}{m}.$$

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 [65, 167].

Theorem 10.6: PRDS property of conformal p-values

Let $Z_1, \dots, Z_n$ be independent draws from P. Assume $S(Z)$ is continuous for $Z \sim P$, and assume that $Z_{n+1}, \dots, Z_{n+m}$ 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 monotone-rank proof of this theorem is given in Appendix B.

This theorem is the bridge back to Chapter 6. Conformal p-values are not independent, but they are super-uniform—their 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: BH at FDR target q obeys FDR $\le m_0 q/m \le q$. We reserve α 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 $\{1/(n+1), \dots, 1\}$, 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 is too wide in easy regions and too narrow in hard regions.

Quantile regression estimates conditional quantiles. For a conditional distribution

$$F(y \mid X = x) = \mathbb{P}(Y \le y \mid X = x),$$

the τth 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.

The pinball loss for estimating this quantile is

$$\rho_\tau(y, b) = \begin{cases}\tau(y - b), & y > b, \\ (1 - \tau)(b - y), & y \le b.\end{cases}$$

Classical quantile regression estimates $\xi_\tau(x)$ by minimizing an average pinball loss, possibly with regularization.

Conformalized quantile regression (CQR) of Romano et al. [136] wraps estimated lower and upper quantiles with conformal calibration. Split the data into D_1 and D_2. 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. For each $i \in D_2$, define the calibration score

$$A_i = \max\{\widehat{\xi}_{\alpha/2}(X_i) - Y_i,\ Y_i - \widehat{\xi}_{1-\alpha/2}(X_i)\}.$$

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.

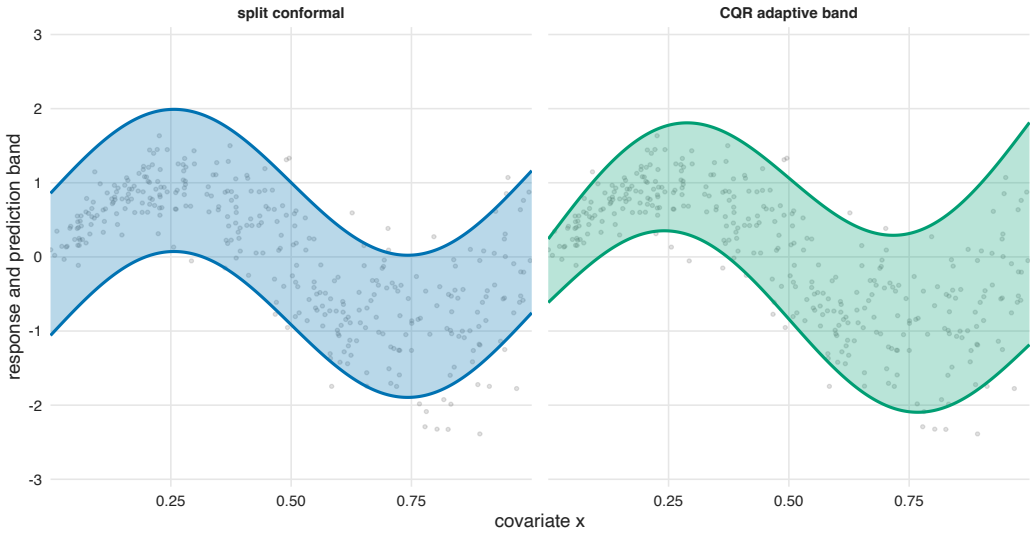

FIGURE 10.3. Split conformal and CQR adaptive bands in a heteroscedastic regression simulation with 240 training observations, 180 calibration observations, 350 independently plotted observations, and $\alpha = 0.1$. Split conformal uses a constant residual quantile; CQR calibrates a covariate-dependent 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, 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.

Let $\widehat{f}_{-i}$ be the predictor trained on all observations except (X_i, Y_i). Define the leave-one-out residual

$$R_i^{\mathrm{LOO}} = |Y_i - \widehat{f}_{-i}(X_i)|.$$

The ordinary jackknife interval uses the full-data predictor $\widehat{f}$ and the empirical quantile of the leave-one-out residuals:

$$[\widehat{f}(X_{n+1}) - \widehat{c}_{\mathrm{LOO}},\ \widehat{f}(X_{n+1}) + \widehat{c}_{\mathrm{LOO}}].$$

This 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+ [11] changes the endpoints. For 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 half-line $[L_{\mathrm{JK+}}, \infty)$ when only the upper boundary triggers, the half-line $(-\infty, U_{\mathrm{JK+}}]$ when only the lower triggers, and the whole line $\mathbb{R}$ when both trigger.

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 algorithmic symmetry of the fitting rule, the jackknife+ interval satisfies

$$\mathbb{P}\{Y_{n+1} \in [L_{\mathrm{JK+}}, U_{\mathrm{JK+}}]\} \geq 1 - 2\alpha.$$

Why the factor two appears. The jackknife+ coverage theorem above is proved by a deterministic tournament count on the augmented sample and exchangeability of the new vertex. The complete counting lemma, cyclic example, and sharpness discussion are deferred to Appendix B.

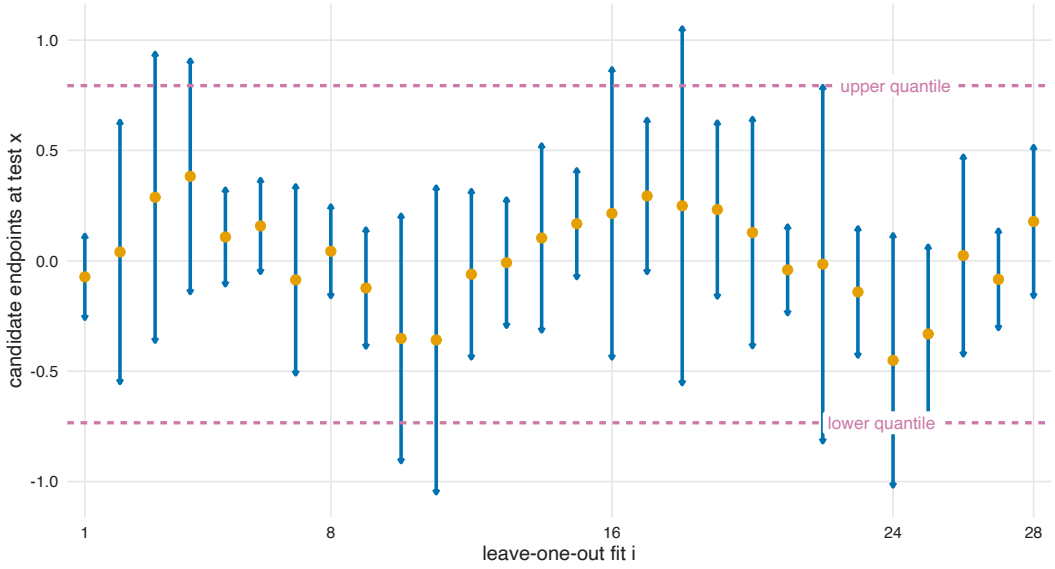

FIGURE 10.4. Jackknife+ endpoints. Each leave-one-out fit gives a lower and upper candidate endpoint at the test covariate. The final interval is formed from quantiles of these endpoint collections and is denoted $[L_{\mathrm{JK}+}, U_{\mathrm{JK}+}]$. The display uses $n = 28$ and $\alpha = 0.2$ so the endpoint quantiles are visibly interior.

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 $1 - \alpha$. 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 2α is the same kind of robust but conservative price that appears in jackknife+.

If an independent calibration sample is available after aggregation, one can calibrate the voting threshold or confidence level using the CRC theorem in Section 7 of Appendix B; this restores a direct $1 - \alpha$ marginal coverage statement for the aggregated rule. The majority-vote guarantee and post-aggregation calibration are useful when several prediction sets are available but their dependence is hard to model [40, 63].

10. Conformal e-prediction

Route: Research frontier

A different generalization inverts e-value tests rather than p-value tests [64]. For every candidate response $y \in \mathcal{Y}$, let $E_n(y) \geq 0$ be a statistic computed from the calibration data, the test covariate, and the candidate y. The required validity condition is

$$\mathbb{E}[E_n(Y_{n+1})] \leq 1.$$

In words, the statistic evaluated at the true response 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

$$C_a(X_{n+1}) = \{y \in \mathcal{Y} : E_n(y) < 1/a\}.$$

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 $\{E_n(Y_{n+1}) \geq 1/a\}$, so Markov's inequality gives probability at most a. For the second statement, the pointwise inequality

$$\frac{\mathbf{1}\{E_n(Y_{n+1}) \geq 1/\widehat{a}\}}{\widehat{a}} \leq E_n(Y_{n+1})$$

holds: whenever the indicator is one, its defining inequality gives $1/\widehat{a} \leq E_n(Y_{n+1})$, 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. [64] 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

Assumption diagnostic

Checkable	Plot calibration scores over time and groups, inspect ties, and audit that tuning precedes calibration.
Not identified from these data	Future exchangeability and pointwise conditional coverage cannot be certified by marginal calibration data alone.
Sensitivity analysis	Report rolling-window and subgroup coverage, weighting sensitivity, and drift stress tests.
Conservative fallback	Use wider weighted/robust sets, recalibrate frequently, or restrict the claim to the observed deployment regime.

Exchangeability is the engine of conformal validity. Once the calibration scores and the test score are exchangeable, the proof is a rank proof. The prediction algorithm may be complicated or misspecified; this affects interval length and adaptivity, not the marginal coverage guarantee.

The guarantee is marginal over the next draw. It does not say that

$$\mathbb{P}\{Y_{n+1} \in C_n(x) \mid X_{n+1} = x\} \geq 1 - \alpha$$

for every x. Distribution-free conditional coverage at every covariate value is impossible without weakening the requirement or adding structure. Conditioning on the proper training split, conditioning on a realized calibration set, conditioning on a selected subset, and conditioning on $X = x$ are different statements.

Randomization can remove finite-grid conservatism and handle ties exactly, but it does not repair nonexchangeability. Jackknife+ needs exchangeability and a symmetric fitting rule but no stability assumption; majority-vote aggregation needs valid base sets but no assumption on their mutual dependence. Their robust finite-sample guarantees pay a 2α price, and the tournament argument shows that the generic tournament bound itself is tight up to one player. CRC controls average loss, not necessarily every individual loss. LTT can certify non-monotone choices, but Bonferroni-type certification and BH-type certification answer different error-rate questions.

How conformal p-values may be combined depends on how they were built. Online full conformal retraining with a symmetric score and distinct scores gives exactly independent p-values. A single shared calibration set gives PRDS under the independent-test and continuous-score assumptions of Theorem 10.6; under arbitrary dependence one should fall back on dependence-robust corrections or on e-values. The independence in the online setting requires almost surely distinct scores; with ties, fresh independent smoothing is needed for exact independence. The conformal supermartingale tests exchangeability anytime, but a rejection says only that exchangeability fails somewhere in the stream, not where or how.

12. Bibliographic Notes

Route: Core

The modern conformal prediction framework is developed in Vovk et al. [172] and surveyed by Shafer and Vovk [143]. Distribution-free predictive inference for regression and split conformal methods are treated by Lei et al. [106]. Conformalized quantile regression is due to Romano et al. [136]. Jackknife+ and related cross-validation methods are developed by Barber et al. [11]. Conformal prediction under covariate shift through density-ratio reweighting is developed by Tibshirani et al. [160]; the broader treatment of conformal prediction beyond exchangeability is in Barber et al. [13]. 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. [17]. An accessible pedagogical introduction covering the breadth of the field is Angelopoulos and Bates [1]. The book-length theoretical treatment in Angelopoulos et al. [4] 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 [166]; the failure of training-conditional coverage for full conformal prediction is established by Bian and Barber [25]. Group-wise calibration under the name Mondrian conformal prediction originates with Vovk et al. [171]. Distribution-free prediction bands and the limits of test-conditional coverage are developed in Lei and Wasserman [105] and Barber et al. [12]. Simultaneous or transductive conformal prediction for multiple test points is treated by Vovk [167] and Gazin et al. [65]. Selection-conditional conformal coverage and related selective conformal p-value methods are developed by Jin and Candès [92], Bao et al. [7], and Jin and Ren [93]. Aggregation

of conformal prediction sets by majority vote is studied by Cherubin [40] and Gasparin and Ramdas [63].

Conformal risk control extends the conformal program from miscoverage to arbitrary monotone bounded losses; the core construction is from Angelopoulos et al. [3], and risk-controlling prediction sets in the dual form are from Bates et al. [15]. The Learn-Then-Test 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. [64]; 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 [165, 172]. 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 [134, 172]. Adaptive conformal inference updates the working miscoverage level online so that the realized coverage tracks the target even under distribution shift [66]. 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 [14, 104]; the positive-dependence analysis itself extends from independent to exchangeable test points in Marandon et al. [118]. The universality of conformal prediction, which says that any symmetric procedure with distribution-free marginal coverage under exchangeability coincides with full conformal prediction for some score function, so that searching over score functions loses no generality among symmetric methods, appears in Vovk et al. [172]; a textbook treatment is given by Angelopoulos et al. [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 [4, 61, 78]. 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. [72] and Vovk and Petej [168].

13. Exercises

Route: Core

Basic.

Exercise 10.10: Rank proof

Let $Y_1, \dots, Y_{n+1}$ be exchangeable and let

$$k = \lceil (1-\alpha)(n+1) \rceil.$$

Let $Y_{(k)}$ be the kth order statistic, counting ties with multiplicity. Prove that $\mathbb{P}\{Y_{n+1} \le Y_{(k)}\} \ge 1-\alpha$. Then show that, with no ties, the probability is exactly $k/(n+1)$. Alternatively, show that independent uniform tie-breaking makes the rank exactly uniform even when ties occur.

Exercise 10.11: Split conformal interval

Write the split conformal algorithm for absolute residual scores. Identify which data are used to train the predictor and which data are used to compute the residual quantile.

Exercise 10.12: Conformal p-value

(a) Suppose S is a conformity score, where larger means better conformity. Write the conformal p-value for a new point z. Then give the corresponding formula when S is a nonconformity score, where larger means worse.

(b) Now suppose $Z_1, \dots, Z_{n+1}$ are i.i.d. from P and $S(Z)$ has a continuous distribution under P. Using the conformity-score formula from part (a), prove that $p(Z_{n+1})$ is exactly uniform on $\{1/(n+1), \dots, 1\}$.

(c) Deduce from part (b) that the p-value is super-uniform: $\mathbb{P}\{p(Z_{n+1}) \le t\} \le t$ for every $t \in [0,1]$.

Intermediate.

Exercise 10.13: General scores

Let $S(x, y; \widehat{f})$ be any nonconformity score computed after training $\widehat{f}$ on D_1. Prove split conformal coverage for

$$C_n(x) = \{y : S(x, y; \widehat{f}) \le R_{(k)}\}.$$

Here $R_{(k)}$ is the kth order statistic of the calibration scores with $k = \lceil (1-\alpha)(n_2+1) \rceil$; if $k > n_2$, use $R_{(k)} = \infty$.

Exercise 10.14: Randomized ranks

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 $F^*(W) \sim \mathrm{Unif}(0,1)$. Then translate this result into exact randomized conformal coverage for discrete scores.

Exercise 10.15: CQR score

Derive the CQR score

$$A_i = \max\{\widehat{\xi}_{\alpha/2}(X_i) - Y_i,\ Y_i - \widehat{\xi}_{1-\alpha/2}(X_i)\}.$$

Show that $Y_{n+1} \in \widetilde{C}_n(X_{n+1})$ is equivalent to $A_{n+1} \le \widehat{c}_{\mathrm{CQR}}$, where A_{n+1} is the same score evaluated at the test point.

Exercise 10.16: Many test points

Suppose one split-conformal threshold is used for m test points that are iid from the calibration distribution and independent of the training and calibration data. Condition on the complete training and calibration data, including the fitted score function and threshold, and let Q denote the resulting conditional coverage probability for one fresh test point. Show that the conditional probability that all m prediction sets cover is Q^m. Use the marginal guarantee $\mathbb{E}Q \ge 1-\alpha$ and Jensen's inequality to prove the lower bound $(1-\alpha)^m$, and solve for the value of α needed to achieve a target familywise error rate α_{FWER}.

Computational.

Exercise 10.17: Heteroscedastic simulation

Reproduce Figure 10.3. Simulate heteroscedastic data and compare the average length and empirical coverage of absolute-residual split conformal and CQR-style conformal intervals.

Exercise 10.18: 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 10.19: Covariate shift

Reproduce Figure B.3. Vary the amount of test covariate shift and report empirical coverage. Then try a weighted conformal correction when the density ratio is known.

Advanced.

Exercise 10.20: Beyond marginal coverage

Give an example where split conformal has valid marginal coverage but poor conditional coverage for a subset of the covariate space. Explain why this does not contradict Theorem 10.3.

Exercise 10.21: 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 the corresponding calibration scores. An analyst lets J be the index $j \in \{1, 2\}$ with the smaller average calibration score $n_2^{-1} \sum_{i \in D_2} R_i^{(j)}$, and then runs split conformal prediction with the predictor $\widehat{f}_J$, calibrating the threshold on the *same* set D_2. Explain why the $1 - \alpha$ guarantee of Theorem 10.3 no longer follows and can in fact fail; support the conclusion either by a general argument or by an explicit example or small simulation. Identify the exact step of the coverage proof that breaks: conditional on the selection $J = j$, are the calibration scores $\{R_i^{(j)} : i \in D_2\}$ and the test score $R_{n+1}^{(j)} = |Y_{n+1} - \widehat{f}_j(X_{n+1})|$ still exchangeable? Note that the selection event depends on the calibration scores but not on the test score, so it favors the model whose calibration scores happen to be small. Finally, propose two fixes: a further split, in which the model is selected on one fold and the threshold is calibrated on a fresh fold; and a full-conformal construction that, for each candidate response y, treats the calibration points and (X_{n+1}, y) symmetrically in both model selection and calibration, and then inverts over y, as required for the response-dependent selection rules discussed in Section 3 of Appendix B.

Exercise 10.22: Conformal p-values and BH

State precisely the dependence structure between the test points that Theorem 10.6 actually requires. Then prove that applying BH to the conformal p-values for m such test points controls FDR at level q, and explain why this property does *not* extend to arbitrary joint dependence among the test points: identify what goes wrong in the PRDS proof if the test points are allowed to be adversarially dependent.

Exercise 10.23: Weighted split conformal under known density ratio

Let the calibration data be iid from P_{cal}, let the test point be independent with law P_{test}, and assume $P_{\mathrm{test}} \ll P_{\mathrm{cal}}$ with known Radon–Nikodym derivative $w(x, y) = dP_{\mathrm{test}}/dP_{\mathrm{cal}}(x, y)$. Using

one fixed score function for calibration and test observations, derive the weighted prediction set

$$\widehat{C}_n^w(x) = \{y : S(x,y) \leq \widehat{c}_{x,y}^w(\alpha)\}$$

where $\widehat{c}_{x,y}^w(\alpha)$ is the weighted $(1-\alpha)$ empirical quantile of the discrete distribution that places mass

$$\frac{w(X_i, Y_i)}{\sum_{j \in D_2} w(X_j, Y_j) + w(x,y)}$$

on each calibration score $S(X_i, Y_i)$, and mass

$$\frac{w(x,y)}{\sum_{j \in D_2} w(X_j, Y_j) + w(x,y)}$$

at $+\infty$. Prove the resulting marginal coverage guarantee under these absolute-continuity, independence, and known-ratio assumptions.

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 ℓ_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 studies a different inferential target. We do not ask whether a variable survived model selection, and we do not pretend that the selected model is fixed. Instead, for a pre-specified coordinate j, or for a low-dimensional collection of coordinates, we ask for confidence intervals and tests for the coefficient β_j in a high-dimensional linear model. 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.

1. Model and Target

Route: Core

Table 11.1 is a reading guide for the target, not a menu of interchangeable estimands. 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	β in $E[y \mid X] = X\beta$	Coefficients of the specified mean vector on the fixed design.
Random X, correct model	Population β in $E[Y \mid X] = X^\top\beta$	A structural or conditional-mean coefficient under the sampling law for rows.
Misspecified mean	$\beta^* = \arg\min_b \|\mu - Xb\|_2^2$ for fixed X, or $\arg\min_b E(Y - X^\top b)^2$ for random X	A least-squares projection target; robust variance is generally required.

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 ith row of X is $x_i^\top$, and the jth column is X_j. The empirical Gram matrix is

$$\widehat{\Sigma} = \frac{X^\top X}{n}.$$

When rows of X are random with covariance matrix Σ, the precision matrix is denoted by

$$\Gamma = \Sigma^{-1}.$$

Many papers use Θ for the precision matrix; in this book Γ 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 β_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. The three most common constraints are

$$\|b\|_0 = \sum_{j=1}^{p} \mathbf{1}\{b_j \neq 0\}, \qquad \|b\|_1 = \sum_{j=1}^{p} |b_j|, \qquad \|b\|_2^2 = \sum_{j=1}^{p} b_j^2.$$

Best subset selection uses the ℓ_0 constraint and searches directly over sparse supports. It is statistically natural but computationally difficult. Ridge regression uses an ℓ_2 penalty and is stable under collinearity, but it does not usually produce exact zeros. The Lasso [158] uses an ℓ_1 penalty:

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}. \tag{5}$$

The ℓ_1 ball is convex and has corners, so the optimization problem is tractable and the solution is often sparse.

Figure 11.1 isolates the source of the inferential problem: once a coordinate crosses the threshold, the Lasso still shifts it toward zero by λ, so sparsity and shrinkage arrive together.

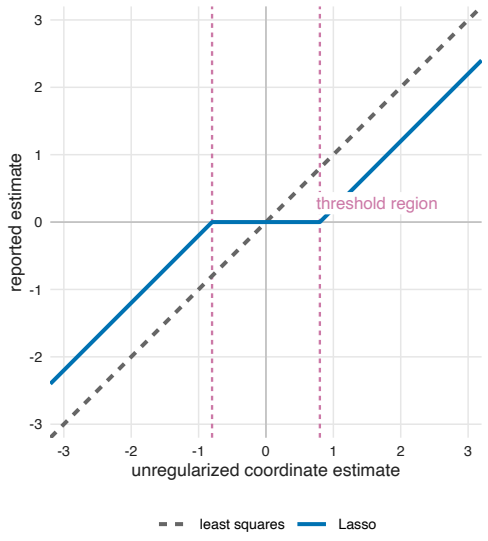

FIGURE 11.1. Lasso shrinkage in the orthonormal one-coordinate problem. The soft-thresholding rule removes small coordinates and shifts large coordinates toward zero. This bias remains visible on the scale used for ordinary confidence intervals.

The same shrinkage that helps prediction makes direct inference difficult. In the orthonormal one-coordinate problem, the Lasso estimate is

$$\widehat{\beta}_j = \operatorname{sign}(z_j)(|z_j| - \lambda)_+,$$

where $z_j = X_j^\top y/n$ is the least-squares coordinate. This soft-thresholding map is not centered at the truth. Therefore $\widehat{\beta}_j$ is not approximately normal around β_j unless the penalty is asymptotically negligible on the $n^{-1/2}$ scale, which would destroy 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 ℓ_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 $p \times p$ eigenvalues of $\widehat{\Sigma}$ to be bounded away from zero, which is impossible when $p > n$. 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 (5) 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 $\lambda \asymp \sigma\sqrt{(\log p)/n}$, 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 β. 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 ℓ_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

$$\frac{P}{2} \le \frac{3\lambda}{2}\frac{\sqrt{s}}{\phi_0}\sqrt{P},$$

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 λ and adding the prediction bound gives the stated inequality with a universal constant, for instance $C = 21$. The probabilistic rates follow by substituting $\lambda \asymp \sigma\sqrt{(\log p)/n}$. □

Theorem 11.2 is an estimation theorem, not an inferential theorem. Its prediction-error bound is consistent under the usual condition $s \log p/n \to 0$. Its ℓ_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 is fixed, $X^\top X/n \xrightarrow{P} \Sigma_0$, and

$$\frac{X^\top \varepsilon}{\sqrt{n}} \xrightarrow{d} W, \qquad W \sim N(0, \sigma^2 \Sigma_0).$$

If $\sqrt{n}\,\lambda_n \to \lambda_0$, then the limit of $\sqrt{n}(\widehat{\beta} - \beta)$ is the minimizer of a random convex function of the form

$$V(u) = \frac{1}{2} u^\top \Sigma_0 u - u^\top W + \lambda_0 \sum_{j=1}^{p} \left[u_j \operatorname{sign}(\beta_j) \mathbf{1}\{\beta_j \neq 0\} + |u_j| \mathbf{1}\{\beta_j = 0\} \right],$$

up to constants determined by the normalization of the Lasso objective [97]. This limit is not generally centered normal. If $\beta_j = 0$, the absolute value term creates a kink at the origin; if $\beta_j \neq 0$, the linear penalty term shifts the center. High-dimensional debiased methods should therefore be viewed as new estimators for inference, not as ordinary normal approximations to the raw Lasso.

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 jth row of M. The debiased Lasso estimator is

$$\widehat{b} = \widehat{\beta} + M \frac{X^\top (y - X\widehat{\beta})}{n}. \tag{6}$$

Substituting $y = X\beta + \varepsilon$ gives the exact decomposition

$$\widehat{b} - \beta = M \frac{X^\top \varepsilon}{n} + (I - M\widehat{\Sigma})(\widehat{\beta} - \beta). \tag{7}$$

For coordinate j,

$$\sqrt{n}(\widehat{b}_j - \beta_j) = \frac{m_j^\top X^\top \varepsilon}{\sqrt{n}} + \sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta), \tag{8}$$

where e_j is the jth standard basis vector.

The first term in (8) 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 central limit theorem 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 $p > n$.

Figure 11.2 turns the algebra into a geometric requirement: the debiasing direction must retain sensitivity to X_j while suppressing its nuisance-column correlations.

The remainder is controlled by Hölder's inequality:

$$\left| \sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta) \right| \leq \sqrt{n}\, \|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \|\widehat{\beta} - \beta\|_1. \tag{9}$$

If

$$\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

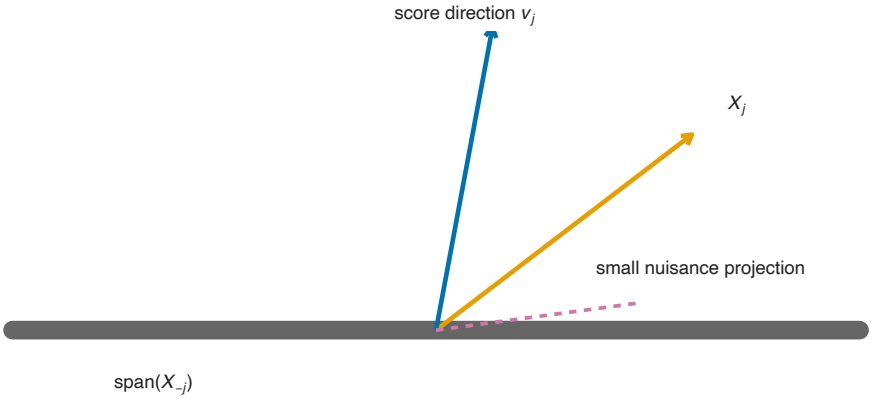

FIGURE 11.2. The debiasing direction should keep unit sensitivity to the target column X_j while being nearly orthogonal to nuisance columns X_{-j}. This geometry is the coordinate version of Neyman orthogonalization.

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

$$\frac{s \log p}{\sqrt{n}} \to 0. \tag{10}$$

This condition is stronger than the condition needed for prediction consistency, because inference must remove bias on the $n^{-1/2}$ scale.

Theorem 11.3: Coordinatewise debiased normality

Fix a coordinate j. 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 $\eta > 0$,

$$\frac{\sum_{i=1}^n (x_i^\top m_j)^2 E[\varepsilon_i^2 \mathbf{1}\{|(x_i^\top m_j)\varepsilon_i| > \eta\sigma\|Xm_j\|_2\} \mid X]}{\sigma^2\|Xm_j\|_2^2} \xrightarrow{P} 0.$$

Condition (ii) is the needed conditional Lindeberg condition. A uniform conditional $(2+\delta)$-moment bound together with

$$\frac{\max_{1 \le i \le n} |(Xm_j)_i|}{\|Xm_j\|_2} \to 0$$

is a convenient sufficient condition. Then

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j)}{\sigma\sqrt{m_j^\top \widehat{\Sigma} m_j}} \xrightarrow{d} N(0,1).$$

Proof of Theorem 11.3

Use the decomposition (8). 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

$$v_j^2 = \frac{1}{n}\sum_{i=1}^n (x_i^\top m_j)^2 \sigma_i^2.$$

If v_j is bounded away from zero and infinity and

$$\widehat{v}_j^2 = \frac{1}{n}\sum_{i=1}^n (x_i^\top m_j)^2 \widehat{\varepsilon}_i^2 \xrightarrow{P} v_j^2,$$

where the residuals come from a consistent or cross-fitted nuisance fit, 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. □

If $\widehat{\sigma} \xrightarrow{P} \sigma$, Slutsky's theorem replaces σ in the studentized statistic by $\widehat{\sigma}$, and a $1-\alpha$ confidence interval for β_j is

$$\widehat{b}_j \pm z_{1-\alpha/2}\,\widehat{\sigma}\sqrt{\frac{m_j^\top \widehat{\Sigma} m_j}{n}}. \tag{11}$$

The scaled Lasso [154] is one common route to estimating σ. Residual-based estimates can also work, but only under assumptions that make the residual degrees of freedom and model bias sufficiently small.

Read Figure 11.3 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 an autoregressive design of order one, abbreviated AR(1), means $\Sigma_{jk} = \rho^{|j-k|}$, so ρ directly controls how quickly feature correlations decay with coordinate distance.

FIGURE 11.3. 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 Monte Carlo repetitions per setting. Naive intervals centered at the Lasso estimate undercover because of shrinkage. Debiased score intervals are much closer to the nominal 95 percent level, with oracle support intervals included only as a benchmark. Vertical bars are pointwise 95 percent Wald Monte Carlo intervals for estimated coverage; marker shape as well as color identifies the method.

Figure 11.4 separates the error-law issue from the debiasing remainder. Under homoskedasticity the two variance estimators are similar. Under variance changing with leverage, the homoskedastic formula is no longer licensed, whereas the residual sandwich targets the variance in Corollary 11.4.

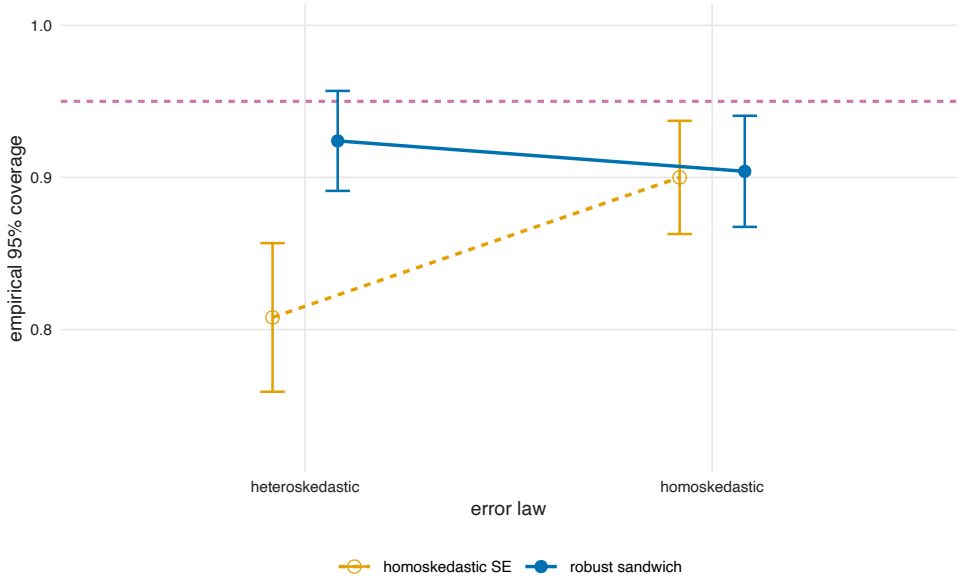

FIGURE 11.4. Homoskedastic and robust debiased-Lasso intervals under constant and covariate-dependent noise variance. The simulation uses $n = 120$, $p = 70$, an AR(1) design with $\rho = 0.4$, five nonzero coefficients $(0.55, 0.45, 0.40, 0.35, 0.30)$, and 250 repetitions per error law. Constant noise has standard deviation 1; covariate-dependent noise has conditional standard deviation $0.45 + 0.8|X_{i1}|$. Vertical bars are pointwise 95 percent Wald Monte Carlo intervals. Color is redundant with line type and marker shape; 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° 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° on X_{-j}° 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

$$\Gamma_{jj} = \frac{1}{\tau_j^2}, \qquad \Gamma_{j,-j} = -\frac{(\gamma_j^\star)^\top}{\tau_j^2}. \tag{12}$$

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

$$\widehat{\gamma}_j \in \arg\min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \|X_j - X_{-j}\gamma\|_2^2 + 2\lambda_j \|\gamma\|_1 \right\}. \tag{13}$$

(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 λ_j. Calibrating λ_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 (13) give

$$\frac{X_{-j}^\top \widehat{r}_j}{n} = \lambda_j \widehat{\kappa}_j, \qquad \widehat{\kappa}_j \in \partial \|\widehat{\gamma}_j\|_1,$$

so $\widehat{\gamma}_j^\top \widehat{\kappa}_j = \|\widehat{\gamma}_j\|_1$. 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

$$\widehat{\Gamma}_j^\top = \frac{\widehat{c}_j^\top}{\widehat{\tau}_j^2}. \tag{14}$$

Stacking these rows gives $\widehat{\Gamma}$, which can be used as M in (6). 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 (9); see Meinshausen and Bühlmann [121], van de Geer et al. [162], and Dezeure et al. [43].

Figure 11.5 shows how the population precision identity becomes an algorithm: residualizing one column against all others creates the direction that is rescaled into an approximate precision row.

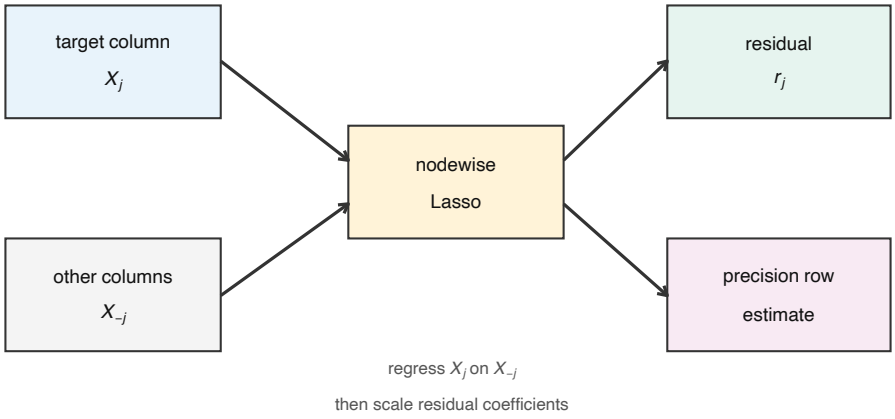

FIGURE 11.5. Nodewise Lasso estimates the direction used to orthogonalize the target coordinate. Regressing X_j on X_{-j} produces a residual direction and, after scaling, an estimated row of the precision matrix Γ.

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 β_{-j} is unknown. Replacing it by the Lasso gives

$$\widehat{\eta}_j = y - X_{-j}\widehat{\beta}_{-j} = X_j\beta_j + \varepsilon + X_{-j}(\beta_{-j} - \widehat{\beta}_{-j}).$$

Let $v_j \in \mathbb{R}^n$ be a projection direction satisfying $v_j^\top X_j = n$. Define

$$\widetilde{\beta}_j = \frac{v_j^\top \widehat{\eta}_j}{n}.$$

Then

$$\sqrt{n}(\widetilde{\beta}_j - \beta_j) = \frac{v_j^\top \varepsilon}{\sqrt{n}} + \frac{v_j^\top X_{-j}(\beta_{-j} - \widehat{\beta}_{-j})}{\sqrt{n}}. \tag{15}$$

The first term has conditional variance $\sigma^2\|v_j\|_2^2/n$. The second term is bounded by

$$\sqrt{n}\left\|\frac{v_j^\top X_{-j}}{n}\right\|_\infty \|\widehat{\beta}_{-j} - \beta_{-j}\|_1.$$

Thus v_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 v_j.

Javanmard and Montanari [90] formulated this tradeoff row by row. Their program lives in parameter space rather than sample space: parametrize the sample-space direction as $v_j = Xm_j$ for a vector $m_j \in \mathbb{R}^p$. Then

$$v_j^\top X/n = m_j^\top X^\top X/n = m_j^\top \widehat{\Sigma},$$

so requiring $v_j^\top X_j/n \approx 1$ and small $v_j^\top X_k/n$ 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

$$\widehat{m}_j \in \arg\min_{m \in \mathbb{R}^p} m^\top \widehat{\Sigma} m \quad \text{subject to} \quad \|m^\top \widehat{\Sigma} - e_j^\top\|_\infty \leq \eta, \tag{16}$$

with $\eta \asymp \sqrt{(\log p)/n}$. The program is trivially feasible when $\eta \geq 1$, because $m = 0$ then satisfies the constraint $\|e_j\|_\infty \leq \eta$. 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 η too small (e.g. $\eta = 0$ with $p > n$ and singular $\widehat{\Sigma}$) leaves an empty feasible set. The constraint controls the bias remainder, and the objective controls the asymptotic variance $\sigma^2 m_j^\top \widehat{\Sigma} m_j$, which equals $\sigma^2 \|v_j\|_2^2/n$ under $v_j = X m_j$. This makes the statistical tradeoff explicit: more exact orthogonalization reduces bias but may require a high-variance direction.

8. Many Coordinates and Multiple Testing

Route: Advanced

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 (9) 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. [39] 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 is conservative because it ignores the dependence among the debiased z-statistics, which can be substantial when the design has correlated columns. Zhang and Cheng [183] develop a bootstrap-assisted procedure that automatically incorporates this dependence. The construction has three steps. First, form the leading score vector $S_j = m_j^\top X^\top \varepsilon/\sqrt{n}$ for $j \in J$, so that the debiased statistic decomposes as $\sqrt{n}\,(\widehat{b}_j - \beta_j) = S_j + o_p(1)$ uniformly in j. Second, generate bootstrap weights ξ_i (multiplier bootstrap) and form

$$S_j^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i\, (m_j^\top X_i)\, \widehat{\varepsilon}_i,$$

where $\widehat{\varepsilon}_i$ are Lasso residuals. Third, use the bootstrap distribution of $\max_{j\in J}|S_j^*|$ to calibrate simultaneous critical values for the debiased statistics. The procedure is valid when the dimension of J is allowed to grow exponentially with n, provided the sparsity and design conditions of debiasing hold. The simultaneous $(1-\alpha)$ confidence intervals it produces are typically much narrower than the Bonferroni intervals when the design has substantial column correlation, because the bootstrap captures the correlation structure directly rather than bounding it through a worst-case dependence argument. 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.

9. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Plot approximate-inverse errors, leverage weights, robust versus homoskedastic standard errors, and tuning-path stability.
Not identified from these data	Sparsity of the true coefficient and precision rows, and correctness of a structural linear target, are not directly observable.
Sensitivity analysis	Vary Lasso/nodewise penalties, compare sandwich and homoskedastic intervals, and perturb high-leverage rows.
Conservative fallback	Use sample splitting, low-dimensional prespecified contrasts, or prediction-focused reporting.

The assumptions behind debiased Lasso are structural. Sparsity of β 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 ℓ_1 error and the approximate-inverse error must vanish. The condition is not a technical nuisance; it is exactly what keeps the regularization bias from contaminating the confidence interval.

10. Failure Modes and Diagnostics

Route: Core

Debiased Lasso can fail for several reasons.

- If β is too dense, the Lasso ℓ_1 error is too large and the bias remainder in (9) is not negligible.

- If the design is highly collinear, the approximate inverse can have large variance or fail to satisfy the required ℓ_∞ approximation.
- If the precision rows are dense, nodewise Lasso may not estimate Γ accurately even when β itself is sparse.
- If the noise variance is poorly estimated, the interval width in (11) 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 standard errors, the empirical correlations $m_j^\top \widehat{\Sigma} - e_j^\top$, and the plausibility of sparsity for both β and the relevant precision rows.

Tuning and diagnostic workflow. Use prediction-tuned Lasso penalties as a starting point, not as an automatic inferential certificate. Record a small prespecified multiplier grid around the prediction penalty for both the outcome Lasso and the nodewise regressions. For each target coordinate, report

$$\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty, \qquad \max_i \frac{|x_i^\top m_j|}{\|X m_j\|_2}, \qquad \widehat{v}_j,$$

along with the estimate over the tuning grid. A small approximate-inverse error with a huge standard error is not success; nor is a stable standard error when the debiased center moves materially across reasonable penalties.

Figure 11.6 should be read as a four-part diagnostic dashboard: inverse error, observation influence, variance sensitivity, and tuning-path drift must all be examined before the reported interval is trusted. The panels do not impose universal pass–fail cutoffs.

Example 11.5: End-to-end $p > n$ analysis

The reproducibility script generates $n = 120$ observations and $p = 180$ AR(1)-correlated features, with eight nonzero coefficients and covariate-dependent noise variance. It fits the outcome Lasso, constructs a nodewise approximate inverse, and debiases the first 30 predeclared coordinates. Each coordinate receives a residual-sandwich interval and a two-sided p-value; BH is then applied once at $q = 0.10$. The complete table of estimates, interval endpoints, adjusted p-values, approximate-inverse errors, and maximum influences is written to `generated/chapter11_p_gt_n_case.csv`. In the seeded realization, six coordinates are discovered, all six belong to the simulated nonzero support; the largest approximate-inverse error across the 30 targets is 0.320, and the largest normalized influence is 0.309. These diagnostics are large enough to warrant caution even though the synthetic truth lets us verify the selected list.

The workflow is deliberately target-first: the 30 coordinates are fixed before simulation, so BH concerns that family rather than the Lasso-selected support. Figure 11.6 is part of the reported analysis, not an ornamental afterthought. Large inverse error, leverage, or tuning drift would trigger a narrower target set, sample splitting, or a prediction-only conclusion even if an adjusted p-value happened to be small.

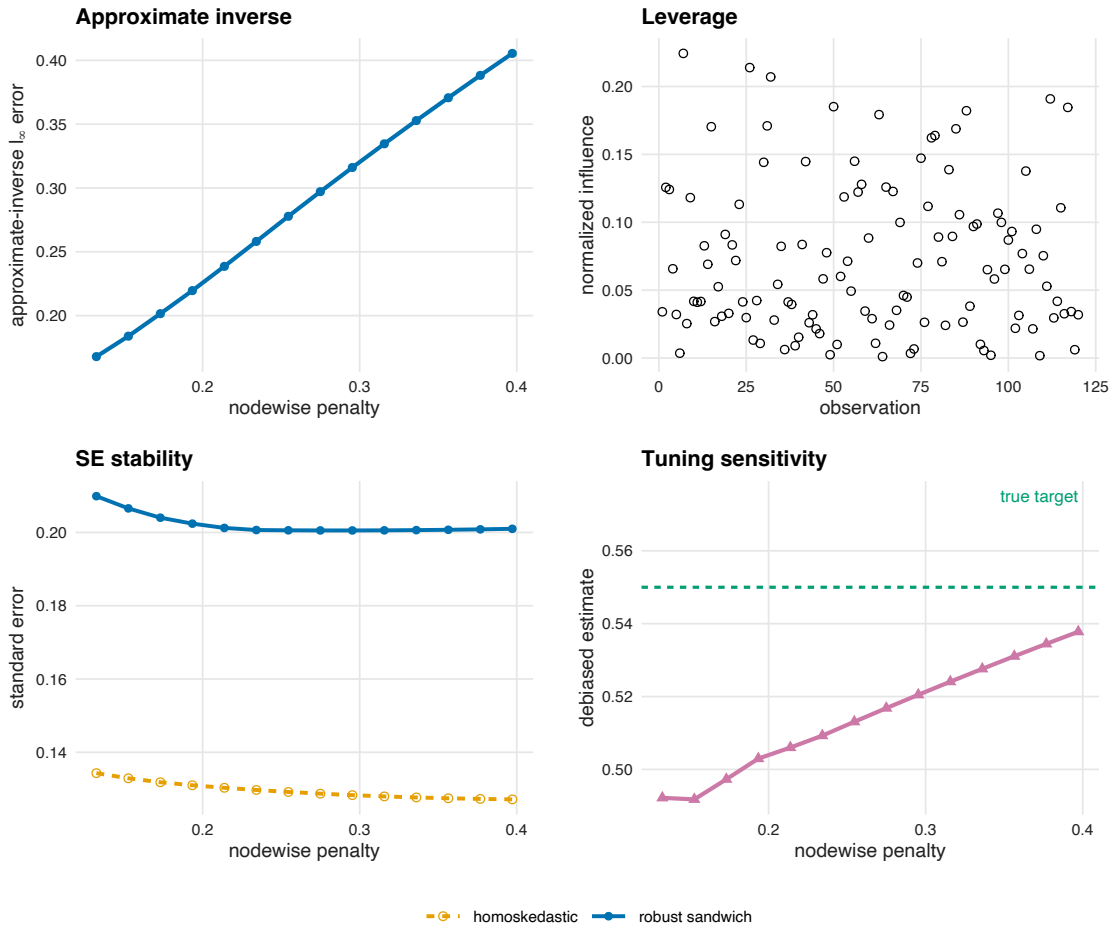

FIGURE 11.6. Four diagnostics from one synthetic fit with $n = 120$, $p = 180$, AR(1) correlation $\rho = 0.45$, eight nonzero coefficients, and covariate-dependent noise: approximate-inverse error, normalized observation influence, homoskedastic versus robust standard errors, and target-estimate sensitivity along the nodewise tuning path. The panels provide diagnostics, not universal pass–fail thresholds. Series are distinguished by line type and markers as well as color.

11. Limits of Adaptivity

Route: Research frontier

The previous sections describe a positive result: under enough sparsity and design regularity, one can obtain honest coordinatewise inference for pre-specified low-dimensional targets. There are also fundamental limits. High-dimensional confidence intervals cannot, in general, be simultaneously short, honest over a large sparse model, and fully adaptive to a much smaller unknown sparsity class. Cai and Guo [32] give minimax lower bounds and adaptivity results for confidence intervals in high-dimensional linear regression. Their results clarify why apparently automatic confidence intervals should be interpreted carefully: extra assumptions such as stronger sparsity, beta-min separation, known support structure, or sample splitting may be needed to obtain the desired length and coverage together.

Figure 11.7 contrasts the short interval available with known nuisance structure against the extra length needed for honesty over a larger unknown sparse class; the vertical gap is the price of adaptation, not an estimator-specific defect.

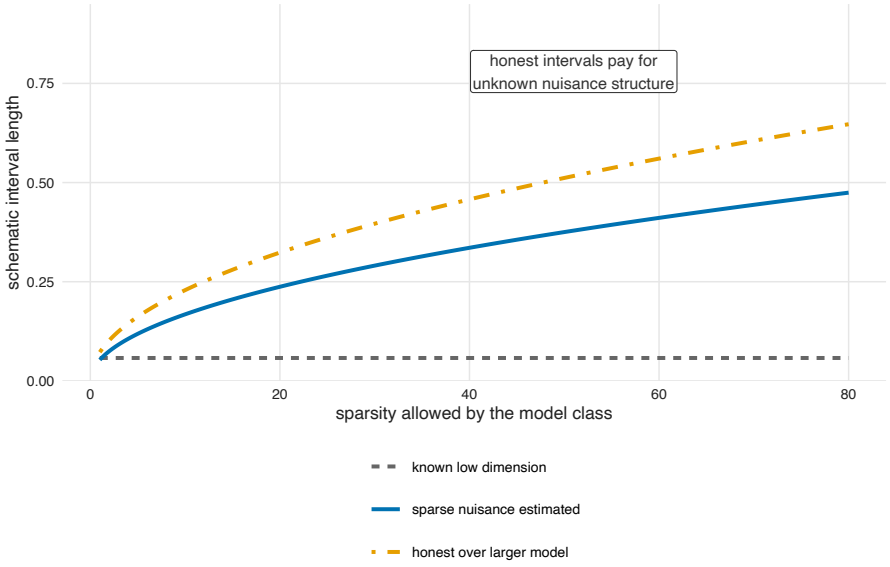

FIGURE 11.7. A schematic view of adaptivity limits. If the nuisance structure were known, interval length would behave like a low-dimensional standard error. Honest inference over larger sparse classes generally pays extra length for the unknown nuisance structure. The figure is illustrative, not a numerical theorem.

12. Bibliographic Notes

Route: Core

The Lasso was introduced by Tibshirani [158]. Fixed-dimensional asymptotics for Lasso-type estimators were developed by Knight and Fu [97]. The standard textbook reference for Lasso theory, compatibility/restricted-eigenvalue conditions, and high-dimensional regression foundations is Bühlmann and van de Geer [31]. Debiased and desparsified Lasso methods for high-dimensional confidence intervals and tests were developed in Zhang and Zhang [179], Javanmard and Montanari [90], and van de Geer et al. [162]. The nodewise Lasso construction is closely tied to Gaussian graphical model estimation by Meinshausen and Bühlmann [121]. Dezeure et al. [43] provide a broad account of high-dimensional inference methods and software, while Cai and Guo [32] 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. [39]. The bootstrap-assisted simultaneous-inference procedure introduced in Section 8 is from Zhang and Cheng [183], 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.

A practically important direction not pursued in detail here is the *streaming* version of debiased Lasso. In modern high-dimensional 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. [74] show that combining stochastic gradient descent updates

with one-pass debiasing achieves asymptotically optimal confidence-interval coverage with strictly $O(p)$ memory. An alternative anytime-valid approach uses confidence sequences (Chapter 8) directly on the debiased Lasso target, replacing fixed-sample intervals with intervals that remain valid under optional stopping.

13. Exercises

Route: Core

Basic.

Exercise 11.6: Soft-thresholding bias

Consider the orthonormal model with $X^\top X/n = I_p$. Show that the Lasso solution is

$$\widehat{\beta}_j = \operatorname{sign}(z_j)(|z_j| - \lambda)_+, \qquad z_j = X_j^\top y/n.$$

For a nonzero coordinate, compute the shrinkage bias in the regime $|\beta_j| \gg \lambda$. For a zero coordinate, show that the estimator has mean zero by symmetry but has a point mass at zero and a non-Gaussian sampling distribution. Explain why both features make ordinary Wald intervals centered at the Lasso estimate unreliable.

Exercise 11.7: Debiasing decomposition

Starting from

$$\widehat{b} = \widehat{\beta} + MX^\top(y - X\widehat{\beta})/n,$$

derive (7). Identify the leading stochastic term and the remainder term.

Exercise 11.8: Coordinate interval

Assume $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$ and M is fixed given X. Derive the conditional variance of $\sqrt{n}(\widehat{b}_j - \beta_j)$ after ignoring the bias remainder, and obtain the confidence interval (11).

Intermediate.

Exercise 11.9: Remainder control

Prove the bound

$$\left|\sqrt{n}\{e_j^\top - m_j^\top \widehat{\Sigma}\}(\widehat{\beta} - \beta)\right| \le \sqrt{n}\|m_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \|\widehat{\beta} - \beta\|_1.$$

Using the Lasso ℓ_1 rate and the approximate-inverse rate, show why $s \log p/\sqrt{n} \to 0$ is sufficient for this term to be $o_p(1)$.

Exercise 11.10: Compatibility condition

Work through the proof sketch of Theorem 11.2. Make the constants explicit under the event $\lambda \ge 2\|X^\top \varepsilon/n\|_\infty$, using the normalization in (5).

Exercise 11.11: Precision rows from regressions

Let $X^\circ \sim N(0, \Sigma)$ and $\Gamma = \Sigma^{-1}$. Use the block inverse formula to prove (12). Interpret the zeros of $\Gamma_{j,-j}$ in terms of conditional linear regression.

Exercise 11.12: Projection direction

For a fixed coordinate j, derive (15). Show how the two requirements $\|v_j\|_2^2/n = O_p(1)$ and $\|v_j^\top X_{-j}/n\|_\infty = o_p\{1/(s\sqrt{\log p})\}$ lead to an asymptotically normal projection estimator.

Computational.

Exercise 11.13: Small debiasing simulation

Simulate $n = 150$, $p = 80$ Gaussian linear models with AR(1) feature correlation. Compare empirical coverage for three intervals for β_1: an interval centered at the Lasso estimate, a debiased interval using a nodewise residual direction, and an oracle interval using the true support. Report coverage and average interval length as the sparsity and correlation change. As a basic sanity check, the naive Lasso-centered intervals should undercover once shrinkage bias is non-negligible, while the debiased intervals should move substantially closer to nominal coverage; compare the setup in `fig09_coverage_sim()` in `scripts/make_figures.R`.

Exercise 11.14: Nodewise Lasso diagnostics

In a simulated design, compute nodewise Lasso rows for several target coordinates. For each coordinate, report

$$\|\widehat{\Gamma}_j^\top \widehat{\Sigma} - e_j^\top\|_\infty \quad \text{and} \quad \widehat{\Gamma}_j^\top \widehat{\Sigma} \widehat{\Gamma}_j.$$

Explain how these quantities correspond to estimated bias and variance.

Exercise 11.15: JM row program

For a small high-dimensional design with $p > n$ (so that $\widehat{\Sigma}$ is singular), solve the Javanmard–Montanari program (16) for several values of η. Compare the resulting variance $m^\top \widehat{\Sigma} m$ and approximation error $\|m^\top \widehat{\Sigma} - e_j^\top\|_\infty$. What happens as η is made smaller? Why does the analogous program with $p < n$ and $\eta = 0$ miss the point of the exercise?

Advanced.

Exercise 11.16: Uniform inference

Theorem 11.3 is stated for a fixed coordinate. Formulate additional conditions that would be needed to justify simultaneous inference over a set J whose size grows with n. Which parts of the proof need to become uniform over $j \in J$?

Exercise 11.17: Post-selection versus coordinate targets

Construct a simple example in which the Lasso selects variable j only when its noisy estimate is unusually large. Explain why a debiased interval for the pre-specified coefficient β_j is not the same as a selective interval conditional on the event that j was selected.

Exercise 11.18: Adaptivity limits

For $\Theta(s) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \le s\}$, with $s_0 < s_1$, compare a coordinate interval honest uniformly over $\Theta(s_1)$ with an oracle interval designed for $\Theta(s_0)$. Using the debiasing remainder $s \log p/\sqrt{n}$, explain why the former generally cannot attain the latter's length; relate this to the lower-bound message of Cai and Guo [32].

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.

The issue is not that exploration is illegitimate. Exploration is often how scientific questions are found. The issue is that an interval with marginal coverage for every fixed parameter need not cover at the advertised rate after we restrict attention to the intervals that looked interesting. 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 [148].

This chapter covers two broad responses. The first controls an average error rate over selected intervals, the false coverage rate. 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 θ_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.

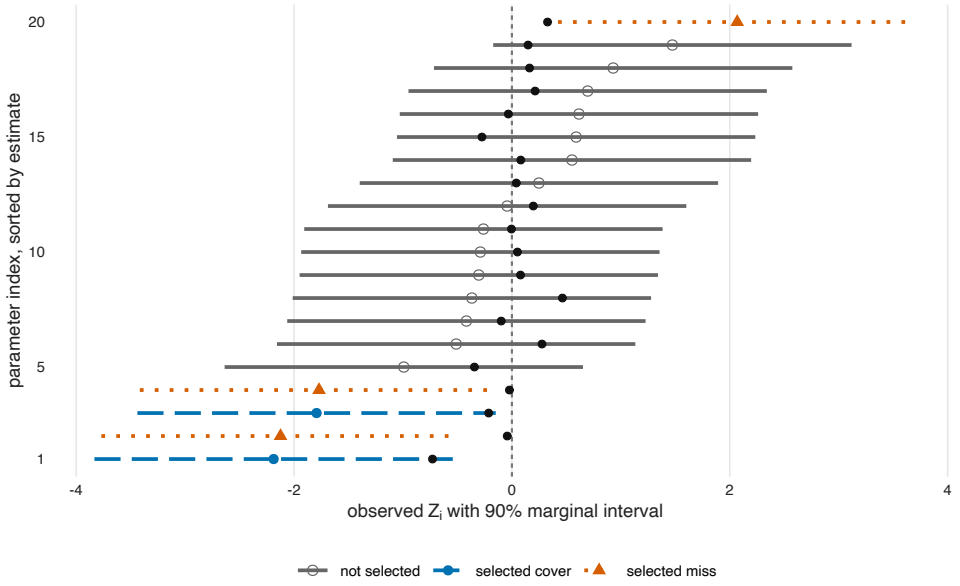

FIGURE 12.1. Marginal intervals after selection. Each horizontal segment is a 90 percent marginal interval; black dots mark the true θ_i's. Selecting intervals that exclude zero concentrates attention on noisy extremes, so selected intervals can miss far more often than their marginal confidence level suggests. Line type and endpoint shape reinforce the color coding for unselected intervals, selected covers, and selected misses.

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 false coverage rate control. The targets remain θ_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

$$\beta_M = (X_M^\top X_M)^{-1} X_M^\top \mu, \tag{17}$$

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 μ 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:

$$\mathbb{P}_\theta(\theta_i \in C_i \mid i \in \widehat{\mathcal{S}}) \geq 1 - \alpha. \tag{18}$$

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 [21], 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 false coverage rate is

$$\text{FCR} = \mathbb{E}\left[\frac{V_{\text{CI}}}{R_{\text{CI}} \vee 1}\right]. \tag{19}$$

This is the confidence-interval analogue of the false discovery rate. It does not promise that every selected interval has conditional coverage (18); 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.

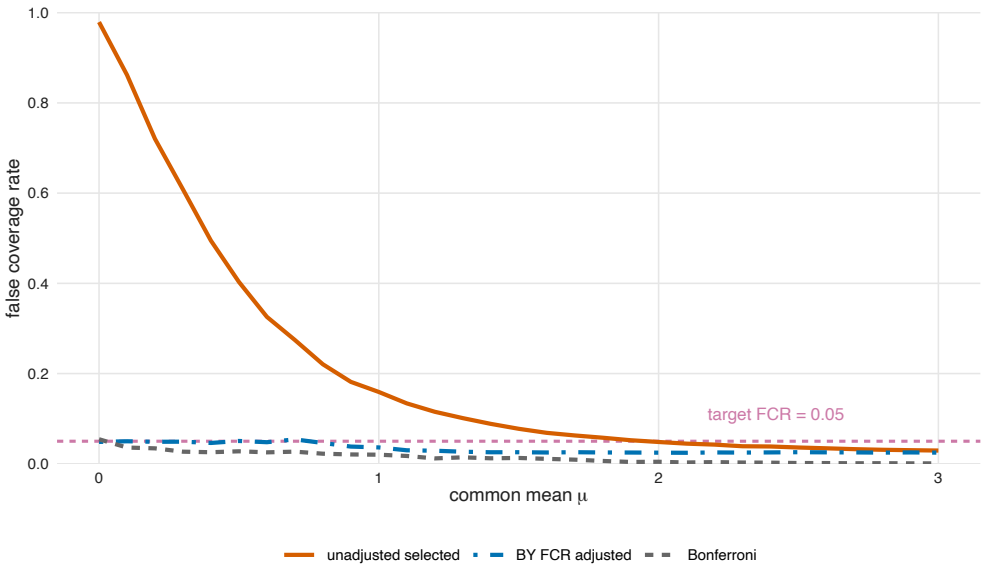

FIGURE 12.2. FCR in the Gaussian common-mean example with $m = 80$, $\alpha = 0.05$, and 2500 Monte Carlo repetitions at each μ. Intervals selected for excluding zero can have very large FCR when the true mean is near zero. The Benjamini–Yekutieli selected-interval adjustment controls FCR while being 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

$$R^{(i)} = \min\bigl(\{|\widehat{\mathcal{S}}(T^{(i:t)})| : i \in \widehat{\mathcal{S}}(T^{(i:t)}),\ t \in \mathbb{R}\} \cup \{\infty\}\bigr). \tag{20}$$

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 (20) is nonempty on the event $i \in \widehat{\mathcal{S}}(T)$, since the realized $t = T_i$ is feasible there; off this event the value is ∞. This makes $R^{(i)}$ a function of T_{-i} alone for all T_{-i}, while the reported interval below uses it only when i is selected. For many monotone selection rules, $R^{(i)}$ equals the realized number R_{CI}, but the definition above is the one used in the finite-sample proof.

Suppose $C_i(\gamma)$ denotes a marginal interval with noncoverage probability at most γ:

$$\mathbb{P}_\theta\{\theta_i \notin C_i(\gamma)\} \le \gamma.$$

For each $i \in \widehat{\mathcal{S}}$, report

$$C_i\left(\frac{\alpha R^{(i)}}{m}\right). \tag{21}$$

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 γ. Then the selected intervals in (21) satisfy

$$\mathrm{FCR} \le \alpha.$$

Proof of Theorem 12.1

Let

$$\Delta_i = \frac{\mathbf{1}\{i \in \widehat{\mathcal{S}},\ \theta_i \notin C_i(\alpha R^{(i)}/m)\}}{R_{\mathrm{CI}} \vee 1}.$$

Then $\mathrm{FCR} = \sum_{i=1}^m \mathbb{E}[\Delta_i]$. 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

$$\mathbb{E}[\Delta_i \mid T_{-i}] \le \sum_{k=1}^{m} \frac{\mathbf{1}\{R^{(i)} = k\}}{k} \frac{\alpha k}{m} = \frac{\alpha}{m}.$$

Taking expectations and summing over i gives $\mathrm{FCR} \le \alpha$. □

As with the Benjamini–Hochberg procedure, the independence assumption can be weakened. Benjamini and Yekutieli [21] establish the same FCR bound when the joint law of $(T_1, \dots, T_m)$ and the coverage events are positive regression dependent in the PRDS sense reviewed in Chapter 6; under arbitrary dependence, replacing α by α/L_m with $L_m = \sum_{k=1}^m 1/k$ restores the control.

The FCR criterion is useful, but it is not a cure for every post-selection effect. It controls the frequency of noncoverage for the stated targets. It does not necessarily correct the direction of selection bias, the slope of empirical-Bayes shrinkage, or the fact that a regression target may have changed after model selection. In large screening problems, intervals that control FCR can still look visually odd because they widen symmetrically around selected estimates rather than directly modeling regression to the mean. This is one motivation for empirical-Bayes refinements and for target-specific selective methods.

6. POSI: Protection Against Arbitrary Selection

Route: Advanced

False coverage rate concerns a fixed family of targets. POSI, short for post-selection inference, 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. [23] protect against this by constructing simultaneous intervals over all selected-model targets.

Let $\mathcal{M} = \{M \subseteq \{1, \dots, p\} : M \neq \emptyset,\ X_M^\top X_M \text{ invertible}\}$ 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

$$\widehat{\beta}_{j\bullet \widehat{M}} \pm K_{1-\alpha}(X)\, \sigma \|\eta_{j\bullet \widehat{M}}\|_2 \tag{22}$$

cover all selected-model coefficients simultaneously, no matter how $\widehat{M}$ was chosen. In practice σ is unknown and is replaced by an estimate $\widehat{\sigma}$ independent of y (or based on a t-type pivot), which leads to a slightly different constant $K_{1-\alpha}(X, n-p)$; the simultaneous-coverage logic is the same.

Theorem 12.2: POSI simultaneous protection

Under the fixed-design Gaussian linear model, if $K_{1-\alpha}(X)$ is chosen as above with known σ, then for every data-dependent model selection rule $\widehat{M} = \widehat{M}(y)$,

$$\mathbb{P}\left\{\beta_{j\bullet \widehat{M}} \in C_{j\bullet \widehat{M}} \text{ for all } j \in \widehat{M}\right\} \geq 1 - \alpha,$$

where $C_{j\bullet \widehat{M}}$ is the interval in (22).

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 $b = 1, \dots, B$, 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 compares the price of different protections through their interval multipliers: fixed-model intervals pay for no selection search, POSI pays for every allowed search, and FCR or conditional methods occupy the middle by protecting a specified selection target.

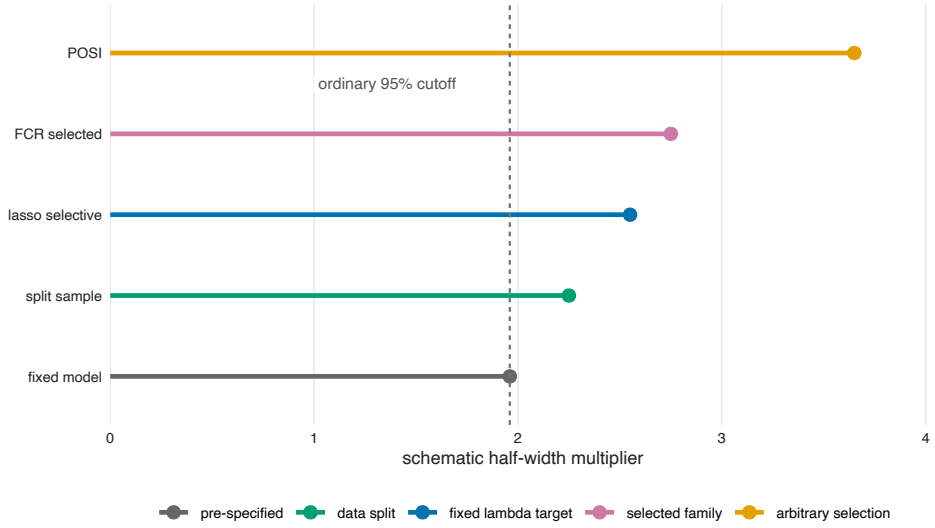

FIGURE 12.3. A schematic comparison of interval width multipliers. Ordinary fixed-model intervals are narrow because the target is pre-specified. POSI protects against arbitrary selection and can be much wider. FCR and selective intervals protect more specific selection problems and targets.

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. [103] assumes that the Lasso tuning parameter λ is fixed before looking at y. Cross-validation is itself a selection procedure; treating a cross-validated λ as fixed changes the conditioning event and is not covered by the basic theorem.

Consider

$$y \sim N(\mu, \sigma^2 I_n),$$

with fixed design X. For fixed $\lambda > 0$, the Lasso estimator is

$$\widehat{\beta}^{\lambda} \in \arg\min_{b} \left\{ \frac{1}{2}\|y - Xb\|_2^2 + \lambda \|b\|_1 \right\}.$$

Let

$$\widehat{M} = \{j : \widehat{\beta}_j^{\lambda} \neq 0\}, \qquad \widehat{s} = \operatorname{sign}(\widehat{\beta}_{\widehat{M}}^{\lambda}).$$

The Karush–Kuhn–Tucker (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 $\operatorname{diag}(s)\widehat{\beta}_M^{\lambda} > 0$ 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

$$\{\widehat{M} = M, \widehat{s} = s\} = \{Ay \leq b\}, \tag{23}$$

where A and b depend on X, λ, M, and s, but not on the unknown mean μ.

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.

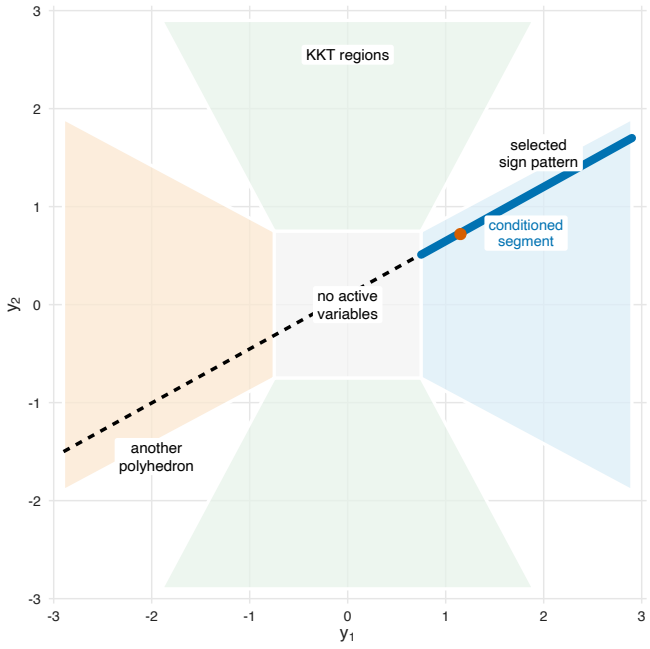

FIGURE 12.4. A schematic lasso selection partition in the response space. For a fixed active set and sign vector, the KKT conditions define a polyhedron $\{Ay \leq b\}$. Conditioning further on the nuisance projection restricts inference to the highlighted solid line segment inside the selected polyhedron; the dashed continuation lies outside that selection event.

Fix a selected model M, a selected coordinate $j \in M$, and the contrast

$$\eta^\top = e_j^\top (X_M^\top X_M)^{-1} X_M^\top,$$

where, throughout this section, $e_j \in \mathbb{R}^{|M|}$ denotes the local basis vector picking out the position of j within M, not the global basis vector in $\mathbb{R}^p$. The selected-model 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-λ 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-λ 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 observations

$$W_i \sim N(\mu_i, \sigma^2 I_d), \qquad i = 1, \dots, n,$$

are clustered by an algorithm $\mathcal{C}$. 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.

For fixed disjoint sets C_1, C_2, define

$$v_i(C_1, C_2) = \frac{\mathbf{1}\{i \in C_1\}}{|C_1|} - \frac{\mathbf{1}\{i \in C_2\}}{|C_2|}.$$

Then

$$\bar{W}_{C_1} - \bar{W}_{C_2} = W^\top v(C_1, C_2).$$

Let

$$P_v^\perp = I_n - \frac{v v^\top}{\|v\|_2^2}.$$

Under the null, the norm of the mean difference, its direction, and the orthogonal nuisance projection can be separated using Gaussian independence. The selective clustering p-value conditions on enough information to make the remaining one-dimensional law tractable:

$$p_{\text{sel}} = \mathbb{P}_{H_0}\{D(W) \ge D(w) \mid E_{\text{sel}}(W, w)\},$$

where $D(W) = \|\bar{W}_{C_1} - \bar{W}_{C_2}\|_2$, and $E_{\text{sel}}(W, w)$ is the event that $C_1, C_2 \in \mathcal{C}(W)$, $P_v^\perp W = P_v^\perp w$, and

$$\text{dir}(\bar{W}_{C_1} - \bar{W}_{C_2}) = \text{dir}(\bar{w}_{C_1} - \bar{w}_{C_2}).$$

After conditioning on the nuisance projection and the observed direction, the data vary along a scalar separation parameter δ_{sep}. The event that C_1, C_2 appear in the clustering becomes

$$\delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2),$$

a data-dependent truncation set. Thus the selective p-value can be written as

$$\frac{\mathbb{P}\{\delta_{\text{sep}} \geq \delta_{\text{sep,obs}},\ \delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)\}}{\mathbb{P}\{\delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)\}}, \tag{24}$$

where, under the global null $\mu_1 = \cdots = \mu_n$ (the pairwise null $\bar{\mu}_{C_1} = \bar{\mu}_{C_2}$ alone leaves residual mean structure in the nuisance projection),

$$\delta_{\text{sep}} \sim \sigma \sqrt{|C_1|^{-1} + |C_2|^{-1}}\, \chi_d,$$

and χ_d denotes the chi distribution with d degrees of freedom, that is, the law of the Euclidean norm of a standard $N(0, I_d)$ vector. This is the clustering analogue of the lasso truncated-normal pivot: condition on the selection event and nuisance information until the remaining statistic has a truncated, computable reference distribution.

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 partition.

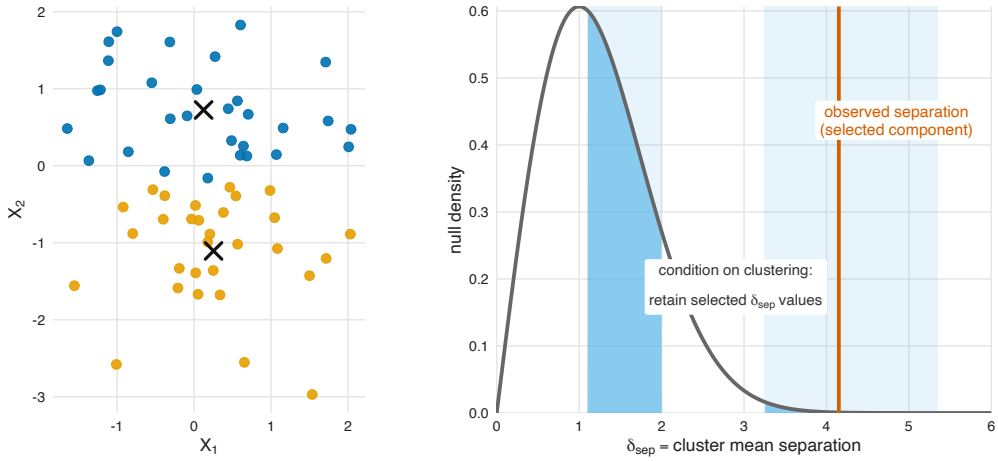

FIGURE 12.5. Selective inference for clustering. Even under a null model, clustering creates separated groups. Conditional inference varies the cluster separation δ_{sep} along a line determined by the observed nuisance information and evaluates the null law truncated to values that reproduce the selected clustering event. In the schematic right panel, that selected support has two disconnected components, and the observed separation lies in the right component.

The practical computation may require Monte Carlo sampling over δ_{sep} and rerunning the clustering algorithm to determine whether $\delta_{\text{sep}} \in \mathcal{S}(w, C_1, C_2)$. The same principle applies to hierarchical clustering [62] and k-means [38], although the selection events differ by algorithm.

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. The construction is elegant, finite-sample valid, and immune to the polyhedral-conditioning 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. [124] 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 θ: 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 $\epsilon \in (0,1)$. Then there is a distributional split with $(X^{(1)}, X^{(2)})$ 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

For the abstract split, 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. If the conditional law of (Y_1, Y_2) given $Y_1 + Y_2 = x$ is available and parameter-free (or has known parameters), then sampling from that conditional law after observing $X = x$ reconstructs the same joint distribution as (Y_1, Y_2) conditional on its sum. Averaging over $X \sim P_\theta$ therefore gives unconditional marginals $P_{\epsilon\theta}$ and $P_{(1-\epsilon)\theta}$, and the components are independent because the reconstructed joint law is the original product law.
The cleanest concrete case is Poisson, where the conditional law is the classical binomial split. Let $X \sim \text{Poisson}(\lambda)$ and, conditional on $X = k$, draw $X^{(1)} \sim \text{Binomial}(k, \epsilon)$ and set

$X^{(2)} = X - X^{(1)}$. The joint mass function factorizes:

$$\mathbb{P}(X^{(1)} = j, X^{(2)} = k) = e^{-\lambda}\frac{\lambda^{j+k}}{(j+k)!}\binom{j+k}{j}\epsilon^j(1-\epsilon)^k = \underbrace{e^{-\epsilon\lambda}\frac{(\epsilon\lambda)^j}{j!}}_{X^{(1)}\sim\text{Pois}(\epsilon\lambda)} \cdot \underbrace{e^{-(1-\epsilon)\lambda}\frac{((1-\epsilon)\lambda)^k}{k!}}_{X^{(2)}\sim\text{Pois}((1-\epsilon)\lambda)}.$$

Independence and the two marginal distributions fall out of this single algebraic identity. Convolution closure of the Poisson family gives $X^{(1)} + X^{(2)} \sim \text{Poisson}(\lambda)$, so the sum has the same distribution as X.

The same recipe works whenever the conditional split law given the sum is available in a form that can be sampled without estimating the target from the same data. For Gaussian data, the analogous independent views are usually treated as data fission via external Gaussian noise when the variance is known; negative binomial thinning uses a beta-binomial split; the abstract split is the conditional law of the sum $Y_1 + Y_2 = X$ for hypothetical independent draws $Y_1 \sim P_{\epsilon\theta}$, $Y_2 \sim P_{(1-\epsilon)\theta}$. The detailed verification and the parameter conditions for common families are in Neufeld et al. [124]. □

The operational consequence is that we can perform any data-driven step (clustering, PCA, change-point detection, feature selection) on $X^{(1)}$ and then conduct *exact* inference on $X^{(2)}$, because $X^{(2)}$ is independent of the choices made from $X^{(1)}$. 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, modeled as Poisson with mean μ_{ij}. 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 $X^{(1)}$ — for instance, by Louvain on the cosine-similarity graph of normalized $X^{(1)}$ — to obtain a partition $\widehat{\mathcal{C}} = (\widehat{C}_1, \ldots, \widehat{C}_g)$. 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 $X^{(1)}$, the conditional distribution of any test statistic computed from $X^{(2)}$ given $\widehat{\mathcal{C}}$ is the unconditional distribution under the null, and the resulting two-sample Poisson test has nominal type I error.

Remark 12.6: Why sample splitting fails here

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. [44] 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. [109] 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 mandatory "debiasing" step that uses the conditional distribution to compute the test statistic on the inference part.

Example 12.7: Gaussian fission

Let $X \sim N(\mu, \tau^2)$ with τ^2 known. Generate an external $Z \sim N(0, \sigma^2)$ 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 μ 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. The general data fission framework generalizes the Gaussian construction by replacing the τ^2/σ^2 coefficient with the appropriate analogue from the known conditional distribution of $g(X)$ given $f(X)$; for example, when X is Poisson, the role of σ^2 is played by the expected-information ratio between the two components.

Remark 12.8

Data thinning and data fission both retain the full sample size for inference, unlike sample splitting. The cost is a noise inflation: each component carries only a fraction of the original signal. In practice the inflation is modest for moderate ϵ, and the gain in target stability (latent clusters and structures identified on $X^{(1)}$ apply unchanged to $X^{(2)}$) is decisive in unsupervised settings.

The relationship to the polyhedral and clustering selective inference of the earlier sections is illuminating. Polyhedral selective inference computes the conditional distribution of a pivot given the selection event, which can be combinatorially intricate. Data thinning bypasses the combinatorics by constructing an independent "fresh" dataset for inference, at the cost

of distributional assumptions on the data-generating process. Each approach has its place: polyhedral SI is distribution-free in the design matrix and gives exact pivots for fixed-λ Lasso under Gaussian noise; data thinning is parametric in the data distribution but applies to arbitrary selection rules.

10. Selective Inference Beyond Fixed-λ Models

Route: Research frontier

Polyhedral selective inference for the Lasso, as developed in Section 7, conditions on a particular fixed λ and a particular selected active set and sign pattern. In practice, λ is rarely fixed: it is chosen by cross-validation, 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 [117] 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 λ, active set, and signs consequently gives a quadratically constrained selection region, not the single polyhedron that arises for a fixed-λ 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 is to add small randomization to the selection rule itself. Fithian et al. [58] show that adding independent noise to the selection step produces a smoother conditional distribution and tighter selective intervals. When the randomization is small relative to the information content, the selective pivot approaches the unconditional pivot in width, and "data carving" — using both the discovery half and a small fraction of the inference half — recovers most of the lost power.

Selective inference for non-linear models. Taylor and Tibshirani [157] extend selective inference to ℓ_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. [144] develop selective inference for change-point detection by recurrent neural networks: 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. [125] apply selective inference to saliency maps of graph neural networks: 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, by contrast, sidesteps conditioning

entirely: it constructs an independent dataset on which any selection rule can be applied without post-selection adjustment.

11. Assumptions in Plain Language

Route: Core

Assumption diagnostic	
Checkable	Replay the complete selection map, verify linear/quadratic constraints, and inspect truncation widths and Monte Carlo error.
Not identified from these data	Unrecorded analyst choices and the scientific relevance of a selected-model target cannot be recovered after selection.
Sensitivity analysis	Compare conditional, POSI, FCR, randomized, and split-sample intervals.
Conservative fallback	Use sample splitting or simultaneous POSI-style protection.

Selective inference is target-specific. A valid interval must name the target and the selection rule. FCR controls an average over selected intervals, not conditional coverage for every selected interval. POSI protects against arbitrary model selection by paying a simultaneous-inference price. Lasso polyhedral inference is sharper because it conditions on a particular, fixed-λ selection rule. Clustering selective inference is valid for the clusters generated by the specified clustering algorithm and conditioning scheme.

The most common mistake is to combine a data-adaptive workflow with a reference law derived for a fixed workflow. If the selection rule is changed, the selective distribution changes as well.

12. Failure Modes and Diagnostics

Route: Core

- FCR intervals may cover the stated fixed targets on average while failing to represent regression-to-the-mean effects in a scientifically useful way.
- POSI intervals may be too wide to answer the practical question if the selection process was actually much more restricted than arbitrary search.
- Lasso selective intervals require the selection event to be explicit. Cross-validation, screening before the Lasso, or tuning by visual inspection adds selection layers that must be included or justified separately.
- Conditioning on very rare selection events can produce unstable truncated-law calculations and very wide intervals.
- Clustering selective inference depends on the exact clustering algorithm, distance, number of clusters, and any preprocessing used before clustering.

Useful diagnostics include reporting the selection rule in enough detail to be replicated, identifying whether the target is fixed or selected-model, checking the number of selected intervals R_{CI}, and reporting when conditioning events are rare enough to make selective intervals unstable.

13. Bibliographic Notes

Route: Core

Sorić's warning about selected confidence intervals is an early statement of the problem [148]. False coverage rate and selected confidence interval adjustments were introduced by Benjamini and Yekutieli [21]. POSI was developed by Berk et al. [23]. The general selective-inference framework that organizes polyhedral and other conditional pivots, including the optimality viewpoint, is due to Fithian et al. [58]. Polyhedral selective inference for the Lasso is due to Lee et al. [103]; related sequential regression procedures are developed in Tibshirani et al. [159]. Selective inference after clustering is developed for hierarchical clustering in Gao et al. [62] and for k-means in Chen and Witten [38].

Data thinning for convolution-closed distributions is due to Neufeld et al. [124]; the sufficient-statistic generalization that covers uniform, beta, and Wishart families is in Dharamshi et al. [44]. Data fission, which decomposes a single observation into independent components or components with known conditional distribution, is from Leiner et al. [109]. Selective inference for cross-validation-selected models was developed by Loftus and Taylor [117]; the related randomization and data-carving techniques are in Fithian et al. [58]. Post-selection inference for ℓ_1-penalized likelihood models is from Taylor and Tibshirani [157]. Recent extensions to RNN change-point detection and graph neural network saliency are from Shiraishi et al. [144] and Nishino et al. [125].

14. Exercises

Route: Core

Basic.

Exercise 12.9: Selected marginal intervals

Let $Z_i \sim N(\theta_i, 1)$ independently and let

$$C_i = Z_i \pm z_{1-\alpha/2}.$$

Select $\widehat{\mathcal{S}} = \{i : 0 \notin C_i\}$. Show that if $\theta_i = 0$, then a selected interval for i never covers θ_i.

Exercise 12.10: FCR definition

For selected intervals C_i, $i \in \widehat{\mathcal{S}}$, define R_{CI}, V_{CI}, and FCR. Letting $\mathrm{FWER}_{\mathrm{CI}} = \mathbb{P}(V_{\mathrm{CI}} \geq 1)$ be the probability that at least one selected interval fails to cover, show that $\mathrm{FCR} \leq \mathrm{FWER}_{\mathrm{CI}}$.

Exercise 12.11: No-selection case

Suppose all m intervals are reported and each has marginal noncoverage at most α. Show directly from (19) that $\mathrm{FCR} \leq \alpha$, without assuming independence.

Intermediate.

Exercise 12.12: The $R^{(i)}$ construction

Let $\widehat{\mathcal{S}} = \{i : |T_i| > c\}$. For a selected i, compute $R^{(i)}$ in (20). Then repeat the calculation when $\widehat{\mathcal{S}}$ consists of the indices of the k largest $|T_i|$'s.

Exercise 12.13: FCR proof

Fill in the conditioning step in Theorem 12.1. Where is independence of the T_i's used? Construct a short explanation of why the same proof does not automatically apply under arbitrary dependence.

Exercise 12.14: 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 12.15: KKT polyhedron

Consider the Lasso with fixed λ, two predictors, and selected active set $M = \{1\}$ with positive sign. Write the KKT conditions and express them as linear inequalities in y, treating X as fixed.

Computational.

Exercise 12.16: Selected-interval simulation

Reproduce the common-mean simulation in Figure 12.2. Compare the FCR of unadjusted selected intervals, Bonferroni intervals, and the Benjamini–Yekutieli selected-interval adjustment as m changes. A successful reproduction should show the unadjusted selected intervals exceeding the target FCR, with Bonferroni conservative and the Benjamini–Yekutieli adjustment closer to the target; see `fig10_fcr_intervals()` in `scripts/make_figures.R`.

Exercise 12.17: POSI constants by simulation

For a small fixed design with $p = 5$, enumerate all nonempty models M and simulate the POSI statistic

$$\max_{M} \max_{j \in M} |Z_{j\bullet M}|.$$

Estimate its 95 percent quantile and compare it with the Bonferroni cutoff and the ordinary 1.96 cutoff.

Exercise 12.18: Truncated-normal pivot

Simulate $y \sim N(\mu, I_n)$ and a simple polyhedral event $Ay \le b$. For a chosen contrast η, compute $V^-(z)$ and $V^+(z)$ after conditioning on $P_{\eta^\perp} y = z$. Check by simulation that the truncated-normal CDF pivot is approximately uniform conditional on the event.

Advanced.

Exercise 12.19: Clustering truncation set

Let $W = (W_1, W_2, W_3) \sim N(\mu, \sigma^2 I_3)$ 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 $w = (-1.0, -0.4, 0.8)$, the selected clusters are $C_1 = \{1, 2\}$ and $C_2 = \{3\}$. Let

$$v_i = \mathbf{1}\{i \in C_1\}/|C_1| - \mathbf{1}\{i \in C_2\}/|C_2|.$$

Condition on $P_v^\perp W = P_v^\perp w$, and parametrize the remaining line by $W(\delta_{\text{sep}}) = P_v^\perp w + \delta_{\text{sep}} v/\|v\|_2^2$. Derive the interval of δ_{sep} values for which the threshold split still returns $C_1 = \{1, 2\}$ and $C_2 = \{3\}$.

Compare the naive Gaussian p-value for $v^\top W$ with the selective p-value obtained from the normal law truncated to this interval.

Exercise 12.20: Choosing the criterion

Give a data-analysis scenario where FCR control is the appropriate goal, one where POSI is more defensible, and one where a fully conditional selective interval is necessary. For each case, state the target parameter precisely.

Exercise 12.21: Poisson thinning

Let $X \sim \text{Poisson}(\lambda)$ and, conditional on X, draw $X^{(1)} \sim \text{Binomial}(X, \epsilon)$; set $X^{(2)} = X - X^{(1)}$. Show that $X^{(1)}$ and $X^{(2)}$ are independent and marginally Poisson with rates $\epsilon\lambda$ and $(1-\epsilon)\lambda$.
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 λ_j's. 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 complete workflow and verify that the resulting exact conditional two-sided p-value has rejection probability at most the nominal level. Explain why independence of $X^{(1)}$ and $X^{(2)}$ and prespecification of j_0 are relevant, and why a non-randomized exact test may be conservative because of discreteness. Optionally randomize at the boundary to target exact size.

Exercise 12.22: CV-selected lambda as a compound selection event

For Lasso regression with cross-validation choosing λ from a finite grid $\Lambda = \{\lambda_1, \dots, \lambda_K\}$, explain why the event "CV selects λ_k" is not a single fixed-λ Lasso polyhedron. After conditioning on the active sets and signs along the foldwise Lasso paths, describe how the event is represented as a union of many selection regions. How does the number of pieces grow with K and the possible foldwise active sets? Discuss why importance sampling or randomization is typically needed to compute the corresponding selective pivot.

CHAPTER 13

Applications and Research Frontiers

Route: Research frontier

Rapidly evolving and preprint-based material in this chapter is current as of August 2026.

The methods in this book were developed for statistical inference, not for a single scientific area. They become most useful when an application has three features at the same time: many questions, strong dependence, and a sharp distinction between a calibrated claim and an attractive story. Genomics is the classical example. A single experiment may measure expression levels for thousands of genes or test millions of variants. Large language models (LLMs) give a new example. A model may be evaluated on many benchmarks, audited for watermarks or contamination, and asked to make factual claims whose reliability must be calibrated.

The two domains look different on the surface. The statistical grammar is the same. 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. This chapter is therefore a capstone rather than a collection of unrelated case studies. 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 13.1 shows a stylized genomic discovery pipeline. The important point is not the biology in the boxes; it is the placement of inference after preprocessing and before scientific follow-up. Quality control 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.

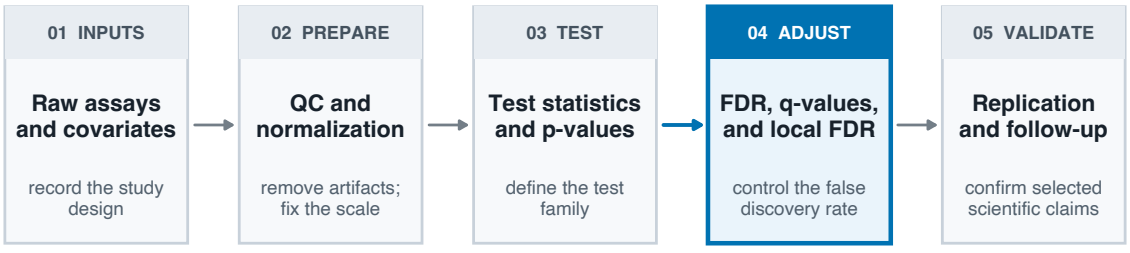

FIGURE 13.1. A genomic discovery pipeline. Multiplicity control follows quality control and test-family definition; selected findings then require replication or biological validation.

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 genome-wide association study (GWAS), H_{0i} might state that a variant has no marginal association with a phenotype after adjustment for pre-specified covariates.

The first temptation is to sort the p-values and report the smallest ones. That is not an inferential procedure. If $m = 20{,}000$, then even perfectly null data will produce many small-looking p-values. The question is not whether the best gene looks interesting; it is how many false discoveries are expected among the genes declared interesting. This is why the Benjamini–Hochberg procedure became central in genomics [19, 49].

For a target FDR level q, BH rejects

$$H_{0(1)}, \dots, H_{0(\widehat{k})}, \qquad \widehat{k} = \max\left\{k : p_{(k)} \le \frac{qk}{m}\right\},$$

where $p_{(1)} \le \cdots \le p_{(m)}$. The selected set $\mathcal{R}$ is a discovery list, not a list of guaranteed truths:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \le q$$

under the assumptions of the procedure. This average statement is often more scientifically useful than familywise error control, because the goal is to produce a stable set of candidates for follow-up rather than to certify that no false positive exists anywhere.

Storey's q-value and Efron's local FDR refine this view [52, 53, 149]. A Storey q-value Q_i can be read as the smallest FDR level at which feature i would enter the discovery set. The local FDR, defined under a two-groups empirical-Bayes model in which each feature is null with prior probability π_0 and nonnull with probability $1 - \pi_0$, is

$$\mathrm{lfdr}(z) = \mathbb{P}(H_{0i} \text{ is true} \mid Z_i = z)$$

and can be read as a posterior null probability under that mixture model. BH, q-values, and local FDR answer related but distinct questions. BH controls an error rate over a selected set; Q_i records where a feature enters such a set; $\mathrm{lfdr}(z)$ summarizes feature-level evidence under a model for the mixture of null and nonnull effects.

3. Dependence, Linkage Disequilibrium (LD), and Structured Discoveries

Route: Core

Genomic p-values are rarely independent. Nearby variants are correlated by linkage disequilibrium (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 Benjamini–Yekutieli 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 conditional randomization tests (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 linkage disequilibrium with a causal variant X_k: X_j carries no independent biological signal once X_k is conditioned on, yet a marginal test of X_j versus y will reject because of the correlation. The conditional null $Y \perp X_j \mid X_{-j}$ is the one a model-X knockoff procedure targets, and X_j is null under that null even though it is highly significant marginally. Model-X knockoffs construct artificial variables $\widetilde{X}_j$ that act as negative controls 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 [8, 36].

Figure 13.2 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 negative control, and the marked threshold turns those contrasts into an FDR-controlled discovery set.

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.

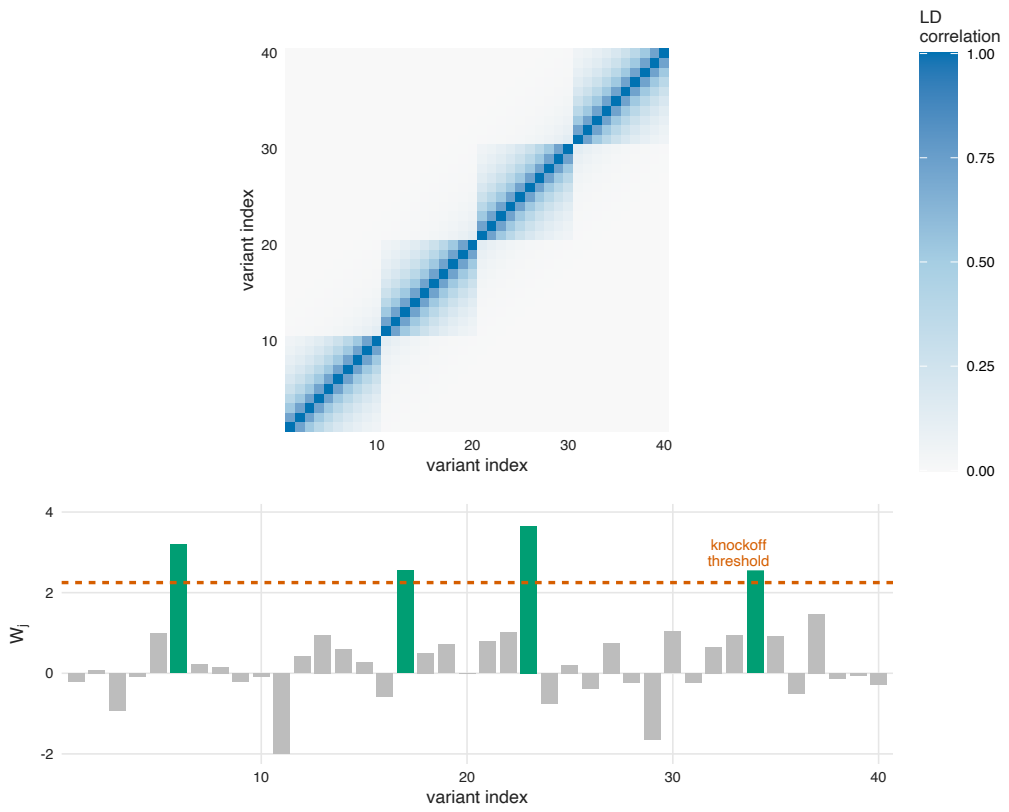

FIGURE 13.2. A schematic GWAS dependence problem. The heatmap displays an LD-like correlation structure among variants, with the color bar reporting LD correlation. The lower panel shows knockoff feature statistics W_j, where large positive values indicate evidence for conditional importance and the threshold is chosen to control FDR.

4. Benchmark Multiplicity for Language Models

Route: Core

Large language models are usually compared across many tasks, datasets, metrics, prompts, and decoding settings. Let Δ_{ab} denote the performance difference between a model and a baseline on task b, or between two models a and a' on task b. Each comparison may have a standard error from paired examples, bootstrap resampling, or a binomial model for success/failure outcomes. If many comparisons are inspected, then the largest improvements will be noisy in the same way that the smallest genomic p-values are noisy.

What is random? The inferential unit must be named before a standard error is computed.

TABLE 13.1. Sources of randomness in benchmark evaluation.

Random unit	Natural estimand	Resampling or variance unit
None: fixed finite census	Average performance on these exact items and runs	No population sampling error; report the finite-census result.
Sampled prompts/items	Expected performance on future items represented by the benchmark	Resample or cluster by item.
Decoding seeds	Expected stochastic-decoding performance on fixed items	Repeat and resample seeds within item; retain item pairing.
Annotators or judges	Mean score under a specified annotator/judge population	Cluster repeated labels and model judge/item effects.
Independent training runs	Expected performance of the training algorithm	The training run is the outer replicate; item-level repeats do not replace it.

Only a population-of-future-items target automatically licenses item resampling as population inference. "Average performance on this fixed benchmark" is a valid but different finite-census claim.

Example 13.1: Worked benchmark evaluation: 30 predeclared comparisons

The reproducible synthetic case in Figure 13.3 compares five candidate models with a common baseline on six tasks. Each task contains paired items, and each item is evaluated under three nested decoding seeds. For model a, task b, item i, and seed r, let d_{abir} be the candidate-minus-baseline outcome and first average over seeds, $\bar{d}_{abi} = 3^{-1} \sum_r d_{abir}$. The task effect is the mean of $\bar{d}_{abi}$, and its standard error clusters at the item level. The 30 one-sided nulls $H_{ab} : E(d_{abir}) \leq 0$ are declared before the simulation and BH is applied once at $q = 0.10$. Circles mark the resulting discoveries, and every cell prints its paired effect, so the plot remains readable without color. In the seeded realization BH discovers 23 of the 30 predeclared cells; the full effect, standard-error, raw-p-value, and adjusted-p-value table is written to `generated/chapter13_benchmark_case.csv`.
Two interpretations are reported. If the items are a random sample, the intervals and tests concern future items from the represented task population. If the benchmark is a fixed finite census, the same point estimates describe only those items; item clustering is then a stability analysis, not automatic population inference. The random seed and full data-generating specification are recorded in the reproducibility bundle.

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.

Dependence is substantial. Benchmark examples overlap in topic and difficulty; models share pretraining data and architectures; metrics can be correlated within a task. This does not make inference impossible, but it discourages naive independence calculations. Paired designs, bootstrap procedures that resample examples rather than cells, hierarchical testing plans, and FDR-controlling procedures are often more defensible than a table of unadjusted p-values. Hierarchical FDR procedures from Chapter 7 are a natural fit when benchmarks are nested inside capability domains. AI governance documents such as National Institute of Standards and Technology [123] and benchmark-design work such as Liang et al. [114] provide the policy framing for how these multiplicity-adjusted reports should be presented to decision makers.

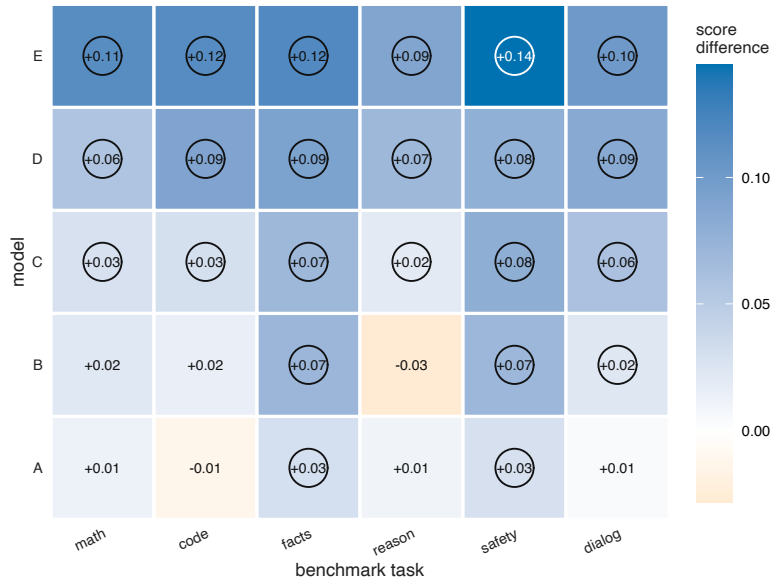

FIGURE 13.3. Benchmark multiplicity for model evaluation. Each cell is a model-by-task comparison, and circled cells are discoveries after a BH-style adjustment. Reporting only the best cells without multiplicity control overstates the strength of evidence. Labels and discovery outlines switch to white on the darkest cells to preserve contrast.

5. LLM-as-a-Judge and Latent Entanglement

Route: Research frontier

A particularly subtle multiplicity problem arises when language models are themselves used to evaluate the outputs of other language models. The LLM-as-a-Judge paradigm has become standard for evaluating tasks where ground-truth answers are expensive or impossible to collect: ask a strong LLM to score the output of a candidate model, treat the score as a measurement, and average across many examples.

Dependence of a useful judge on the candidate output is expected; independence from the output is not the relevant validity criterion. The statistical questions are calibration to a latent target, systematic bias, shared item effects, and dependence among judge errors. One simple measurement model is

$$J_{irk} = L_{ir} + b_k + u_{ik} + \varepsilon_{irk}, \tag{25}$$

Here L_{ir} is latent answer quality for item i and response r, b_k is judge k's average bias, and u_{ik} is an item-by-judge effect. A family-favoritism extension adds $c_{k,f(r)}$, where $f(r)$ records the candidate model family. Shared pretraining data, tokenizers, distillation pipelines, and alignment can make both u_{ik} and the remaining errors correlated across judges.

Mavrogiannis et al. [120] formalize this as a problem of *behavioral entanglement*: ensembles of judges that look superficially independent share correlated error modes that produce *confabulation consensus* — apparent agreement that reflects shared bias rather than independent validation. Statistically, an analysis that omits these terms can understate uncertainty and, more importantly, estimate the wrong quality contrast because of systematic judge bias. Sun et al. [153] propose multi-agent reasoning-tree auditing as an alternative to flat majority voting, demonstrating that

conventional LLM-as-Judge ensembles can be Pareto-dominated by procedures that account for reasoning-trace correlations.

For our purposes the import is methodological: an LLM-as-Judge evaluation is not a flat multiple-testing family of independent measurements. It is at best a clustered family in which the cluster structure is partially observable through pretraining-data and architecture metadata. The hierarchical and group-aware procedures of Chapter 7 therefore suggest an analysis plan only after valid cluster-level p-values or e-values and their dependence assumptions have been specified: control error at the cluster (model-family) level, and treat within-cluster variance as nuisance. Hierarchical FDR cannot repair a biased measuring instrument; calibration, design-based comparisons, or a defensible measurement model must come first. Cumulative Information Gain and similar dependence-corrected metrics offer one operationalization [120]. Selective inference (Chapter 12) gives a second planning principle: when an analyst picks the "best" judging ensemble from multiple candidates, the post-selection inference target should be explicitly the selected ensemble's performance, not the population performance.

6. Watermarked Generation

Route: Research frontier

Watermarking asks whether a piece of text was generated using a known hidden randomization scheme. The typical setting has a trusted model provider, a secret key sequence, and a public text that may have been edited. A detector uses the key to test whether the text carries the statistical signature of the watermark. Modern watermarking methods differ in their engineering details, but the inferential core is a hypothesis test [96, 139].

Let $\mathcal{A}$ be the vocabulary and $K = |\mathcal{A}|$. Write $Y_{<t}$ for the history of tokens generated before step t (truncated to the context window when one is in force). At time t, the autoregressive model assigns probabilities

$$p_t(k) = \mathbb{P}(Y_t = k \mid Y_{<t} = y_{<t}), \qquad k \in \mathcal{A}.$$

A keyed decoder receives p_t and a key ξ_t drawn from a specified key distribution that is independent of $Y_{<t}$, and outputs

$$Y_t = \Gamma(\xi_t, p_t).$$

In practice the ξ_t across positions are produced by a pseudorandom function (PRF) of a master secret and the position, so each token uses fresh keyed randomness but the entire sequence is reproducible by anyone holding the master secret. The watermark should not degrade the distribution of model outputs in a way that is visible to ordinary users. This motivates the following definition.

Definition 13.2: Distortion-free keyed generation

A keyed decoder Γ is distortion-free for a key distribution under which ξ_t is independent of $Y_{<t}$ if

$$\mathbb{P}\{\Gamma(\xi_t, p_t) = k \mid Y_{<t} = y_{<t}\} = p_t(k), \qquad k \in \mathcal{A}.$$

Thus the secret key changes the coupling between the text and the key, but not the marginal distribution of the next token.

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:

$$Y = \arg\min_{1 \le k \le K} E_k.$$

Proposition 13.3: 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 13.3

The minimum of independent exponential variables with rates $p_1, \dots, p_K$ has total rate $\sum_j p_j = 1$. For a fixed k,

$$\mathbb{P}(Y = k) = \int_0^\infty \mathbb{P}(E_j > t \text{ for all } j \neq k)\, p_k e^{-p_k t}\, dt.$$

Independence gives

$$\mathbb{P}(E_j > t \text{ for all } j \neq k) = \exp\left(-t \sum_{j \neq k} p_j\right).$$

Therefore

$$\mathbb{P}(Y = k) = \int_0^\infty p_k e^{-t \sum_j p_j}\, dt = p_k.$$

Tokens with $p_k = 0$ can be excluded or assigned $E_k = \infty$. □

7. Watermark Detection and Scanning

Route: Research frontier

Let $\widetilde{y}_{1:\ell}$ be a public text segment and let ξ denote the secret key sequence. A detector computes a watermark statistic

$$T_{\mathrm{wm}}(\xi, \widetilde{y}_{1:\ell}) := \phi(\xi, \widetilde{y}_{1:\ell}),$$

where larger values mean stronger agreement between the text and the key. The null hypothesis is that the text was not generated using this watermark key. The alternative is that the text was generated, or partly generated, by the keyed decoder.

The cleanest p-value is obtained by key resampling. Draw independent fake keys $\xi^{(1)}, \dots, \xi^{(B)}$ from the same key distribution and compute their scores against the same text. If large scores are evidence for the watermark, use

$$p_{\mathrm{wm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T_{\mathrm{wm}}(\xi^{(b)}, \widetilde{y}_{1:\ell}) \ge T_{\mathrm{wm}}(\xi, \widetilde{y}_{1:\ell})\}}{B + 1}. \tag{26}$$

The direction of the inequality is part of the method. If the statistic were defined so that smaller values were more suspicious, the inequality would have to be reversed.

Proposition 13.4: Key-resampling p-value

Suppose $\xi, \xi^{(1)}, \dots, \xi^{(B)}$ 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 (26) is super-uniform:

$$\mathbb{P}(p_{\text{wm}} \le \alpha) \le \alpha.$$

Proof of Proposition 13.4

Write the observed-key score as S_0 and the resampled-key scores as $S_1, \dots, S_B$. Conditional on the text, these $B+1$ scores are exchangeable under the null. Attach independent continuous tie breakers $U_0, \dots, U_B$, and rank the pairs (S_b, U_b) from largest to smallest, using U_b only when scores tie. The randomized rank R_0 of (S_0, U_0) is then exactly uniform on $\{1, \dots, B+1\}$, so

$$p_{\text{rand}} = \frac{R_0}{B+1}$$

is discrete-uniform on $\{1/(B+1), \dots, 1\}$. The reported p-value uses "$\ge$" and therefore counts *every* score tied with S_0 as at least as extreme. It is no smaller than the randomized-rank p-value: $p_{\text{wm}} \ge p_{\text{rand}}$. Consequently, conditionally on the text,

$$\mathbb{P}(p_{\text{wm}} \le \alpha \mid \widetilde{y}_{1:\ell}) \le \mathbb{P}(p_{\text{rand}} \le \alpha \mid \widetilde{y}_{1:\ell}) = \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \le \alpha.$$

Integrating over the text proves super-uniformity. The added 1 in (26) is the observed key's own contribution to this rank count. □

Remark 13.5: Statistical random keys versus a deployed PRF

The proposition is information-theoretic: the observed and reference keys are genuinely exchangeable draws from a declared distribution. A deployed system usually derives key material from a secure PRF. In that setting the analogous statement is computational, not literal exchangeability: an efficient adversary should be unable to distinguish the PRF outputs from iid random keys, under the PRF's security assumptions. A statistical proof should state the ideal random-key experiment; an implementation claim should separately state the cryptographic reduction and threat model.

Editing creates an additional search problem. A long document may contain a short watermarked passage, or a watermarked passage may have been moved. A substring detector might scan over key locations a, text locations c, and a fixed block length L:

$$T_{\text{wm,max}} = \max_{a,c} \mathcal{M}(\xi_{a:a+L-1}, \widetilde{y}_{c:c+L-1}),$$

where $\mathcal{M}$ is a local match score. This is a multiple-testing problem in disguise. Calibrating a single window and then taking the largest window score would be anti-conservative. One should either use a Bonferroni correction over the scan or, more powerfully, resample the entire maximum statistic:

$$T^{(b)}_{\text{wm,max}} = \max_{a,c} \mathcal{M}(\xi^{(b)}_{a:a+L-1}, \widetilde{y}_{c:c+L-1}), \qquad b = 1, \dots, B,$$

using resampled keys $\xi^{(b)}$ indexed by b. The reference law must include the same search that was used to find the reported match.

Example 13.6: Worked watermark audit: resample the full scan

The reproducible synthetic audit first canonicalizes the displayed text, fixes the candidate window lengths and scan domain, and evaluates the observed key with the maximum-scan statistic $T_{\mathrm{wm}}^{\max}$. It then draws $B = 999$ iid reference keys from the declared ideal key distribution and repeats the entire maximization for every key. The valid p-value is

$$p_{\max} = \frac{1 + \sum_{b=1}^{999} \mathbf{1}\{T_{\mathrm{wm},b}^{\max} \geq T_{\mathrm{wm}}^{\max}\}}{1000}.$$

The companion calculation calibrates the reported maximum against a reference distribution for one fixed, predeclared window. That shortcut is invalid: the observed statistic received many chances to be large while each reference statistic received only one. The simulation records both rejection rates and shows the resulting anti-conservatism.

This is an ideal random-key experiment. If reference keys are generated by a PRF in deployment, the audit inherits the computational caveat in Proposition 13.4; canonicalization removes representation selection but does not replace key exchangeability or the identical-scan requirement.

Figure 13.4 visualizes the worked full-scan audit. Its upper panel gives every reference key the same maximization opportunity as the observed key, which is the valid comparison. The lower panel instead compares the observed maximum with fixed-window reference scores; the resulting shift in the reference distribution shows why omitting the scan is anti-conservative.

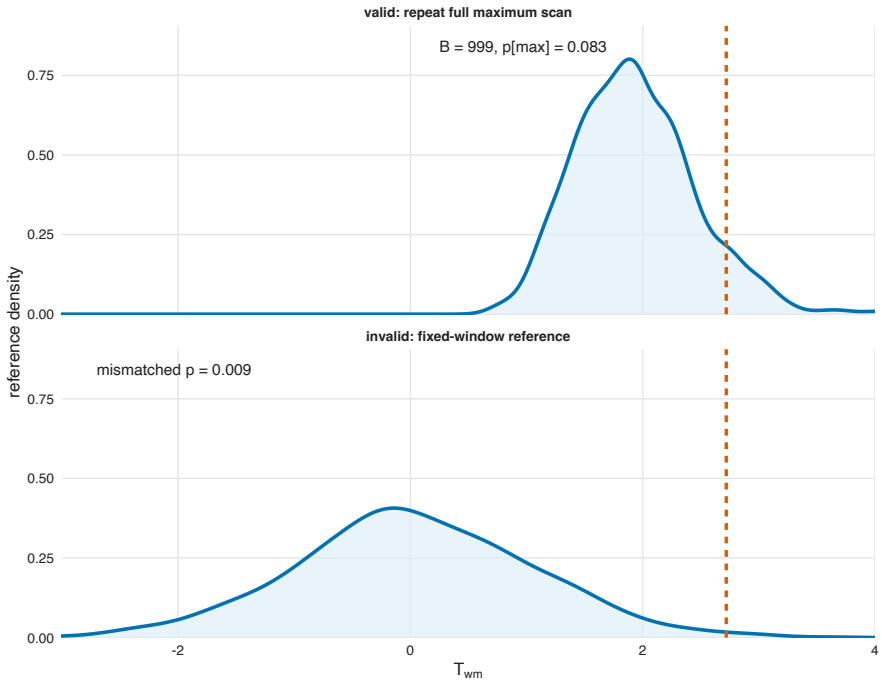

FIGURE 13.4. A watermark maximum-scan audit with $B = 999$ reference keys. The upper panel repeats the full scan for every key; the lower panel shows the invalid fixed-window reference for the same observed maximum. Direct labels and a dashed observed-score line keep the comparison legible in grayscale.

Tokenization multiplicity. A subtle complication of any token-level watermark or auditing scheme is *tokenization multiplicity*. With a fixed deterministic tokenizer, the encoding of a displayed string is usually unique. The statistical issue arises when the testing protocol accepts token sequences, byte strings, provider-side encodings, or other representations that are not canonically tied to the displayed text. In that broader protocol, an identical semantic string can be represented by multiple distinct token sequences. For example, "Multilingual" might decode as (`Multi`, `ling`, `ual`), (`Mu`, `lti`, `lingual`), or (`M`, `ulti`, `lin`, `gual`), each carrying a different probability under the model.

Artola Velasco et al. [6] show that this multiplicity is not merely a theoretical curiosity: an LLM service provider can exploit alternative tokenizations to inflate token counts (and therefore pay-per-token charges) by arbitrary amounts *without altering the displayed text*. For statistical auditing, the same multiplicity can attack watermark detectors if the protocol lets a service provider choose, after the fact, a representation that minimizes the detection statistic while preserving the displayed output.

The statistical remedy is *canonical generation*. Define a deterministic mapping from semantic strings to a unique canonical token sequence, and compute the watermark detection statistic only on the canonical sequence. Canonicalization prevents the analyst or provider from choosing among representations after seeing their watermark scores; it does not itself create exchangeability. Resampling validity still requires exchangeable keys under the null and the identical scan or maximization for the observed and resampled statistics. The representation constraint can be enforced cryptographically by hashing the canonical token sequence and including the hash in the watermark verification protocol.

Verifiable oblivious watermark detection. A second cryptographic refinement targets the user-privacy implications of watermark detection. In the basic detection protocol, the user submits the generated text to the provider for verification; the provider learns the text content and can use it for training or auditing. When the underlying text is sensitive — legal documents, medical communications, or private correspondence — this raises a privacy concern.

Fairoze et al. [55] propose *verifiable oblivious watermarking* (VOW): the user can verify watermark presence with the provider's cooperation, but the provider learns nothing about the text being verified. The construction combines a watermark embedding scheme with a verifiable oblivious pseudorandom function. The detection statistic is computed inside a secure two-party computation between the user and the provider, with the user contributing the text and the provider contributing the watermark key; the output of the computation is a p-value, but neither party learns the other's input.

For our statistical purposes, the key point is conditional: if the VOW implementation exactly evaluates the same detection statistic with the same key distribution and canonicalization rule as the public detector, then the resulting p-value has the same null distribution as the public detection p-value. Under that protocol-level equivalence, verifiable oblivious detection inherits the validity statements developed in this chapter without modification.

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. [126] uses a simple idea. Let

$$X = (X_1, \dots, X_n)$$

be a benchmark dataset in its canonical order, and let $\mathcal{P}$ be the language model being audited. Write $T(X)$ for a score that should be large if the model assigns unusually high likelihood to the canonical sequence. For the test to have power, $T(X)$ must actually depend on the order: scoring rules that evaluate each example independently and then sum or average produce a score that is invariant under permutations, giving zero power. A sequence-level model evaluation that conditions on previously seen examples, or a score that explicitly uses adjacent-example structure, is needed. A canonical choice with this property is

$$T(X) = \log \mathcal{P}(X)$$

when $\mathcal{P}$ scores the whole sequence with cross-example conditioning. The null hypothesis is not "the model is generally clean." A sharper null is that $\mathcal{P}$ is independent of the benchmark dataset and that the benchmark examples are exchangeable under permutations relevant to the audit. Under this null, the canonical ordering is not special.

Proposition 13.7: Permutation p-value for canonical-order contamination

Let $\pi_1, \dots, \pi_B$ be independent 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 13.7

Under the null, the canonical ordering and the randomly permuted orderings are exchangeable. Conditional on the unordered collection of examples, the scores

$$T(X), T(X_{\pi_1}), \dots, T(X_{\pi_B})$$

therefore have an exchangeable joint distribution. The rank argument used in Proposition 13.4 applies verbatim. □

Figure 13.5 makes the permutation calibration concrete. The canonical-order score is compared with the full reference distribution of scores obtained by permuting the same examples, and the upper-tail fraction is the p-value. A visible separation supports the stated contamination claim only when the examples are exchangeable under the audit's permutations and the score is genuinely sensitive to their order.

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 [139]. 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

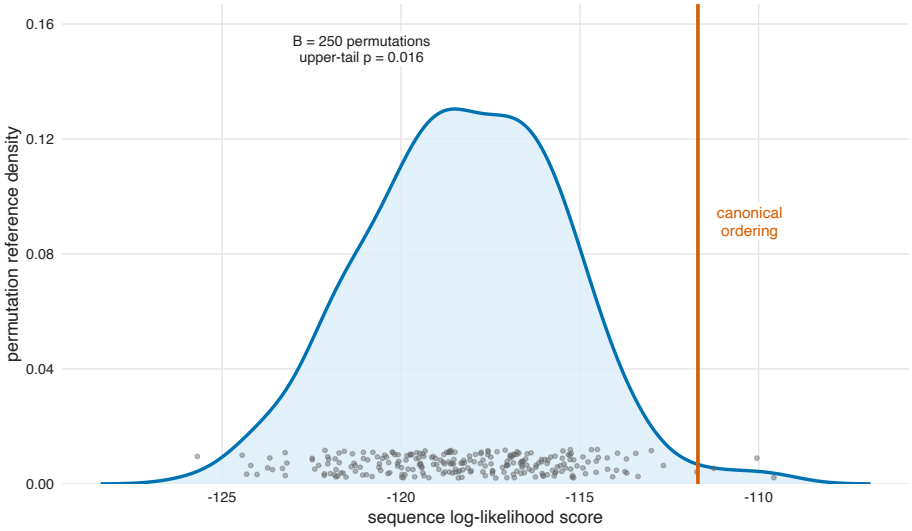

FIGURE 13.5. A canonical-versus-permuted contamination audit. If the canonical ordering receives much larger likelihood than random orderings, the upper-tail permutation p-value based on $B = 250$ random permutations is small. The test supports a specific contamination claim only under the exchangeability and scoring assumptions.

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 resulting p-values form a family. A single small p-value among thousands of audits is not the same claim as a pre-specified audit. FWER, FDR, or e-value methods from earlier chapters should be chosen according to the intended conclusion: one certified violation, a list of suspicious benchmarks, or an ongoing audit that may continue as new models are released.

9. Reference-Relative Factuality

Route: Research frontier

Watermarking and contamination auditing are defensive evaluations of a model. Reference-relative factuality uses the same exchangeability logic to modify model outputs so that they satisfy a coverage guarantee. The construction below follows the statistical structure of Mohri and Hashimoto [122].

> **Remark 13.8: What the guarantee does not mean**
>
> The result below concerns marginal, reference-relative entailment over deployment inputs under exchangeability. It is not a guarantee of metaphysical truth, and it does not certify correctness when the reference or the entailment verifier is wrong.

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 entails the answer a. 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. A back-off family $\{F_t : t \in \mathcal{T}\}$ indexed by a totally ordered threshold set $\mathcal{T} \subseteq \mathbb{R}$ (so the infimum below is well-defined) 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.

For a calibration pair $(X_i, Y_i^\star)$, define the strict-safe score

$$A_{\text{safe}}(X_i, Y_i^\star) = \inf\{t \in \mathcal{T} : Y_i^\star \in E(F_s(X_i)) \text{ for every } s \ge t\}. \tag{27}$$

The phrase "strict-safe" is useful. It is not enough that one threshold happens to be safe; every more conservative threshold should also be safe. This avoids pathologies when the back-off path is not perfectly monotone in its surface form. The infimum is over a possibly empty set: if no level of back-off entails the reference (for example, when the prompt is unanswerable in the chosen family), the score is set to ∞, and A_{safe} takes values in $\mathcal{T} \cup \{\infty\}$. In that case the conformal quantile may also equal ∞, and the conformalized output abstains from a factual claim. For the theorem below, assume the upper-safe set in (27) is closed in $\mathcal{T}$ for every $(x, y^\star)$; this is automatic when $\mathcal{T}$ is a finite grid.

Figure 13.6 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 trades specificity for the stated marginal reference-entailment guarantee.

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 $\alpha \ge 1/(n+1)$, set

$$k = \lceil (n+1)(1-\alpha) \rceil$$

and let $\widehat{t}_\alpha$ be the kth order statistic of $A_{\text{safe},1}, \ldots, A_{\text{safe},n}$. The conformalized output is

$$\widetilde{\mathcal{P}}(x) = F_{\widehat{t}_\alpha}(x).$$

Theorem 13.9: 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. Assume also that the back-off family is sound in the sense that a sufficiently large threshold abstains or makes a claim entailed by every reference in the target class, and that each upper-safe set defining (27) is closed so its infimum is safe. Then

$$\mathbb{P}\left\{Y_{n+1}^\star \in E(\widetilde{\mathcal{P}}(X_{n+1}))\right\} \ge 1 - \alpha.$$

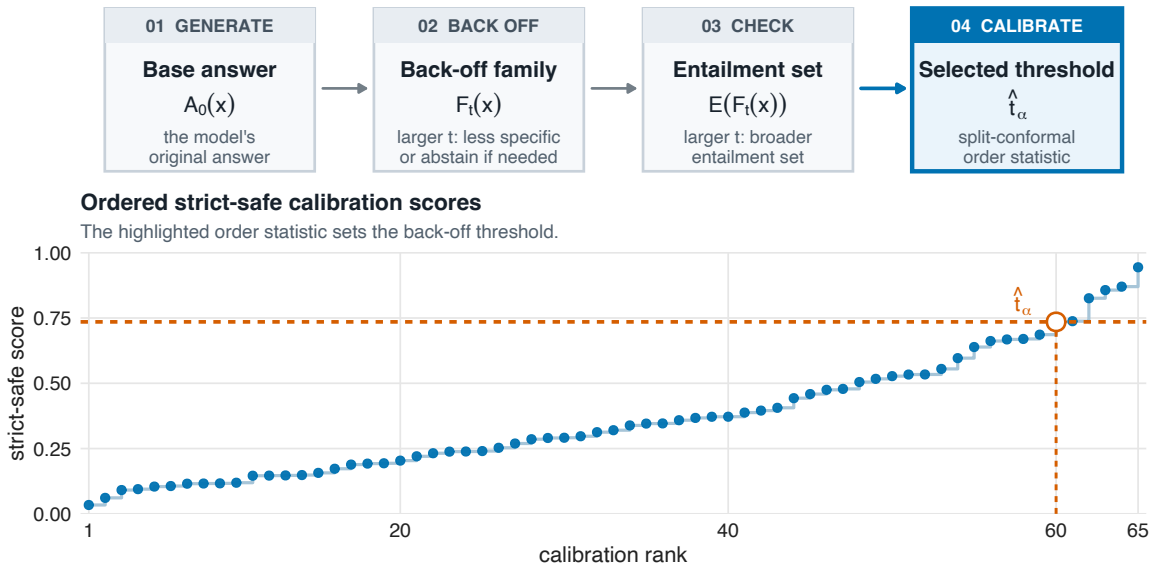

FIGURE 13.6. Reference-relative factuality. Increasing t trades specificity for a larger entailment set; the kth ordered strict-safe score supplies the calibrated threshold $\widehat{t}_\alpha$.

Proof of Theorem 13.9

The scores

$$A_{\text{safe},1}, \ldots, A_{\text{safe},n}, A_{\text{safe},n+1} = A_{\text{safe}}(X_{n+1}, Y^\star_{n+1})$$

are exchangeable because they are deterministic functions of exchangeable pairs. By the split conformal rank argument from Chapter 10,

$$\mathbb{P}(A_{\text{safe},n+1} \le \widehat{t}_\alpha) \ge 1 - \alpha.$$

On the event $A_{\text{safe},n+1} \le \widehat{t}_\alpha$, the upper-safe set property and the closedness assumption imply that $\widehat{t}_\alpha$ itself is a safe threshold, so

$$Y^\star_{n+1} \in E(F_{\widehat{t}_\alpha}(X_{n+1})).$$

This is exactly the desired reference-relative entailment event. □

The guarantee is marginal over future prompts drawn like the calibration prompts. It is not a pointwise guarantee for every possible prompt. It also inherits the definition of entailment used to score the calibration examples. If entailment is judged by humans, the target is human-labeled factuality. If entailment is judged by an automated verifier, the guarantee is only as meaningful as that verifier and its calibration labels.

10. What the Assumptions Are Doing

Route: Core

Assumption diagnostic	
Checkable	Audit the inferential unit, family, measurement model, resampling symmetry, full scan, and reference/verifier error.
Not identified from these data	Latent quality, undisclosed training contamination, PRF security, and future exchangeability are not identified by the reported p-values.
Sensitivity analysis	Vary clustering units, judges, feature/key models, scan domains, and reference/entailment rules.
Conservative fallback	Narrow the claim, use conservative multiplicity control, and report an audit result rather than a causal or truth claim.

The applications in this chapter are useful precisely because the assumptions are visible.

In genomics, the main threats are measurement artifacts, population structure, dependence among features, and target confusion. A marginal association test does not become a conditional importance statement just because the p-value is small. A selected gene list controlled at FDR q does not imply that every gene has posterior null probability at most q. A debiased Lasso interval for a pre-specified coefficient is not the same object as an interval after choosing a model.

In LLM watermarking, the p-value is only valid for the statistic and search procedure that were calibrated. If the analyst scans many substrings, edits the statistic after seeing the text, or reports the best key among many keys, the reference law must include that search. In contamination auditing, exchangeability is the engine of the test. If the benchmark ordering is not exchangeable under the null, the permutation p-value is not justified without modification. In reference-relative factuality, the coverage statement is marginal and distributional; it does not remove the need to define the deployment population and the entailment target.

11. Bibliographic Notes

Route: Core

For genomic multiple testing, see Dudoit and van der Laan [49] for a broad treatment of multiple testing in genomics and Efron [53] for empirical Bayes and local FDR. Knockoffs were introduced for fixed designs by Barber and Candès [8] and extended to the model-X setting by Candès et al. [36]; hidden Markov model knockoffs for genome-wide association studies are developed by Sesia et al. [141]. Standard data resources for human genetic association studies include the NHGRI-EBI GWAS Catalog [147] and the GTEx tissue-expression atlas [70], which 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. [48]. The Holistic Evaluation of Language Models framework of Liang et al. [114] provides the operational template for benchmarking AI systems across many metrics; the NIST AI Risk Management Framework [123] gives the policy framing for how these evaluations should be reported. Green-token watermarking for LLM outputs is due to Kirchenbauer et al. [96]; distortion-free watermarks based on exponential minimum sampling are developed by Kuditipudi et al. [99], and benchmark-level watermarking is studied by Sander et al. [139]. Tokenization multiplicity and its consequences for pricing and watermark auditing are documented by Artola Velasco et al. [6]. Verifiable oblivious watermark detection, which preserves user privacy through a secure two-party computation, is from Fairoze et al. [55]. The

exchangeability-based contamination test is developed by Oren et al. [126]. The reference-relative factuality framework is due to Mohri and Hashimoto [122].

LLM-as-a-Judge latent entanglement is analyzed by Mavrogiannis et al. [120], who propose a cumulative-information-gain metric to quantify behavioral correlation between judges. Multi-agent reasoning-tree auditing as a corrective procedure is from Sun et al. [153]. Replicability across multiple studies, which is increasingly relevant for AI-evaluation reproducibility, is treated in Bogomolov and Heller [28] and Heller and Bogomolov [76].

12. Exercises

Route: Core

Basic.

Exercise 13.10: A genomic discovery list

An expression study tests $m = 18{,}000$ genes and reports $R = 240$ discoveries using BH at $q = 0.05$. State the FDR guarantee in words. Does the guarantee imply that every reported gene has probability at most 0.05 of being null? Explain the difference between FDR and local FDR in this example.

Exercise 13.11: Watermark null and alternative

Write a null and an alternative hypothesis for a watermark detector whose statistic $T_{\mathrm{wm}}(\xi, \widetilde{y})$ is larger when the text agrees more strongly with the secret key. Then write the key-resampling p-value and state which inequality direction should be used.

Exercise 13.12: Exponential minimum sampling

Prove Proposition 13.3 directly for $K = 2$ by integrating over the value of E_1. Then explain how the same calculation extends to general K.

Exercise 13.13: 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. What is the natural family if the team wants to report all improved cells? What is the natural family if it pre-specified one task as primary and treats all others as exploratory?

Intermediate.

Exercise 13.14: Marginal versus conditional genomics

Let X_j be a genetic variant correlated with X_k, and suppose X_k is causal for a phenotype Y while X_j is not. Explain why a marginal test of X_j may reject. State the conditional null targeted by model-X knockoffs.

Exercise 13.15: Scanning correction

A watermark detector scans 500 possible substrings and computes a single-window p-value for each substring. Describe a Bonferroni correction. Then describe a permutation or key-resampling procedure that calibrates the maximum statistic directly. Which approach is likely to be less conservative, and why?

Exercise 13.16: Contamination p-value orientation

In a contamination audit, $T(X)$ is the log-likelihood of the canonical benchmark ordering. Large values are suspicious. Write the permutation p-value using B random permutations. How would the formula change if the statistic were a loss for which smaller values were suspicious?

Exercise 13.17: Reference-relative factuality score

Suppose $F_0(x)$ is the original answer, $F_1(x)$ removes unsupported numbers, $F_2(x)$ removes unsupported named entities, and $F_3(x)$ abstains from factual claims. For a calibration reference $y^\star$, define the strict-safe score in this four-level family. Why is it important to require all more conservative thresholds to be safe?

Advanced.

Exercise 13.18: Multiple contamination audits

An organization audits one model against 1,000 benchmark datasets and obtains one permutation p-value per dataset. Propose an analysis plan if the goal is to identify a list of suspicious datasets while controlling FDR. What dependence concerns arise, and how might e-values or hierarchical testing help if audits are repeated over time?

Exercise 13.19: 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, and the target value of α. State clearly what the resulting guarantee does and does not say.

Exercise 13.20: Knockoff interpretation

In a genotype study with strong LD, a knockoff procedure selects one variant from a correlated block. Explain why this should not automatically be interpreted as identifying the unique causal variant. What additional analyses or study designs would strengthen the biological interpretation?

Exercise 13.21: From a p-value to a claim

For each of the following, state the inferential target, the statistic or score, the reference distribution or calibration law, and the claim justified by the reported output: a gene-level BH analysis, a watermark key-resampling test, a canonical-order contamination test, and a reference-relative factuality calibration procedure. For the first three items, identify the null hypothesis and the claim justified by rejection. For factuality calibration, identify the prediction-set or back-off output and its marginal coverage claim; do not recast it as a p-value rejection.

APPENDIX A

Advanced Safe and Sequential Inference

Route: Advanced

The core route in Chapter 8 introduces e-values, e-BH, and the principal guarantees. This appendix collects the derivations that are most useful for a second course or research reading: numeraire optimality, reverse information projection, process-level arguments, asymptotic constructions, online FDR, and confidence-sequence inversion.

1. Numeraires and reverse information projection

Route: Advanced

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 in one line: by Jensen's inequality applied to the concave logarithm,

$$\mathbb{E}_Q\left[\log \frac{E}{E^*}\right] \le \log \mathbb{E}_Q\left[\frac{E}{E^*}\right] \le \log 1 = 0.$$

Three structural facts, whose fully general proofs are beyond this chapter, organize the whole topic [100, 130]. 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. 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 Kullback-Leibler divergence, it solves the minimization

$$D(Q \,\|\, Q_{G^*}) = \min_G D(Q \,\|\, Q_G),$$

where G ranges over priors on Θ_0 and $D(Q \,\|\, P) = \mathbb{E}_Q[\log\{dQ/dP\}]$ is the Kullback-Leibler 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_θ. 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 q and p with respect to a common measure ν. 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 Kullback-Leibler 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 $p^* = 0$:

$$E^*(X) = \frac{q(X)}{p^*(X)}$$

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:

$$\mathbb{E}_P[E^*] = \int_{\{p^*>0\}} p\,\frac{q}{p^*}\,d\nu = \mathbb{E}_Q[r] \leq 1.$$

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

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] = \int_{\{q>0\}} Ep^*\,d\nu \leq \mathbb{E}_{P^*}[E] \leq 1.$$

The integral is restricted to $\{q > 0\}$; 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 q/p_{θ^*} 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.

For practice, what matters is a verification criterion that turns a guess into a certificate of optimality, and that involves only the null distributions, never abstract mixtures.

Proposition A.3: Guess and verify

Suppose densities exist, let G be a prior on Θ_0 with mixture density p_G, and suppose $p_G(X) > 0$ and $q(X) > 0$ almost surely under Q and under every null distribution. If the candidate

$$E^* = \frac{q(X)}{p_G(X)} \qquad \text{satisfies} \qquad \mathbb{E}_{P_\theta}[E^*] \leq 1 \quad \text{for every } \theta \in \Theta_0,$$

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

$$\mathbb{E}_Q\left[\frac{E}{E^*}\right] = \int E(x)\,\frac{p_G(x)}{q(x)}\,q(x)\,dx = \int E(x)\,p_G(x)\,dx = \int_{\Theta_0} \mathbb{E}_{P_\theta}[E]\,dG(\theta) \leq 1,$$

where the interchange of the two integrals uses Tonelli's theorem for nonnegative integrands, and the final inequality holds because $\mathbb{E}_{P_\theta}[E] \leq 1$ for every $\theta \in \Theta_0$. □

The recipe is: guess the null mixture closest to the alternative, often a single boundary point, and check the expectation inequality over the whole null parameter set. Passing the check

certifies not only that q/p_G is valid but that no e-value whatsoever grows faster under Q. The Gaussian half-line example below runs this verification directly.

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 Kullback-Leibler 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 $\mu = 0$, upgrades this conclusion: the display certifies 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 [130].

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 $(M_t)_{t \ge 0}$ is a *supermartingale* under P if $\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] \le M_{t-1}$ for every $t \ge 1$; 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

$$\begin{aligned}\mathbb{E}_P[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}_P[M_{t-1}E_t \mid \mathcal{F}_{t-1}] \\ &= M_{t-1}\mathbb{E}_P[E_t \mid \mathcal{F}_{t-1}] \le M_{t-1}.\end{aligned}$$

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 are the canonical way to build one. In the literature, conditional e-values are also called *sequential e-values*; the two names describe the same object.

The general notion of a sequential evidence process is broader than a test supermartingale, and the extra breadth is used by real constructions.

Definition A.7: E-process

An adapted sequence of nonnegative random variables $(M_t)_{t\ge 0}$ with $M_0 \le 1$ is an *e-process* for $\mathcal{H}_0$ if

$$\mathbb{E}_P[M_\tau] \le 1 \qquad \text{for every stopping time } \tau \text{ and every } P \in \mathcal{H}_0,$$

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, $\mathbb{E}_P[M_{\tau\wedge K}] \le 1$, and letting $K \to \infty$ with Fatou's lemma gives $\mathbb{E}_P[M_\tau] \le 1$. Notice that no conditions on the stopping time are needed; the nonnegativity of the process is doing the work that boundedness or uniform integrability does in the optional stopping theorems of a standard probability course.

The converse fails: an e-process need not be a supermartingale under any null distribution. The equivalent characterization, whose proof is beyond this chapter, 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 [82, 130, 132]. Two examples show that the inclusion is strict. The sequential version of universal inference,

$$M_t = \frac{\prod_{j\le t} \widehat{q}_{j-1}(X_j)}{\max_{\theta\in\Theta_0} \prod_{j\le t} p_\theta(X_j)},$$

in which the numerator uses predictable alternative fits as in the plug-in method and the null maximum likelihood in the denominator is recomputed from all data at every step, is an e-process

for the composite null but is not a supermartingale, because the denominator can jump upward and is not a running product. And for the exchangeability null of Example 8.8 in sequential form, it can be shown that every test supermartingale is essentially constant, so no interesting supermartingale evidence exists, while nontrivial e-processes do [133]. The betting language extends to the general definition: the analyst plays a separate fair game against each null distribution P and reports the worst-case wealth over the games.

Theorem A.8: Ville's inequality

If $(M_t)_{t\geq 0}$ is a nonnegative supermartingale under a probability measure $\mathbb{P}$, with $M_0 = 1$, then for every $\alpha \in (0, 1)$,

$$\mathbb{P}\left(\sup_{t\geq 0} M_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

Proof of Theorem A.8

Let

$$\tau = \inf\{t : M_t \geq 1/\alpha\},$$

with $\inf \varnothing = \infty$. Fix a deterministic horizon K. The stopped time $\tau \wedge K$ is bounded. For a bounded stopping time $\rho \leq K$, the stopped process $M_{t\wedge\rho}$, $0 \leq t \leq K$, is still a supermartingale; iterating the one-step inequalities gives $\mathbb{E}[M_\rho] \leq \mathbb{E}[M_0]$. Applying this with $\rho = \tau \wedge K$ gives

$$\mathbb{E}[M_{\tau\wedge K}] \leq \mathbb{E}[M_0] = 1.$$

On the event $\{\tau \leq K\}$, the stopped value is at least $1/\alpha$. On the complementary event it is nonnegative. Hence

$$1 \geq \mathbb{E}[M_{\tau\wedge K}] \geq \frac{1}{\alpha}\mathbb{P}(\tau \leq K).$$

Thus $\mathbb{P}(\tau \leq K) \leq \alpha$. The events $\{\tau \leq K\}$ increase to $\{\tau < \infty\}$ as $K \to \infty$, so continuity from below bounds the probability of the crossing event $\{\exists\, t : M_t \geq 1/\alpha\}$ by α. Finally, on the event $\{\sup_t M_t \geq 1/\alpha\}$, the process exceeds $1/\alpha'$ at some finite time for every $\alpha' > \alpha$; applying the crossing bound at level α' and letting α' decrease to α 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 τ 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. The e-process framework is also complete for sequential testing: nothing outside it can be reached.

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-α 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, ϕ_τ 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 τ 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 say that e-processes are exactly the right abstraction for monitored testing: every e-process yields a level-α sequential test by thresholding at $1/\alpha$, and every level-α sequential test arises this way.

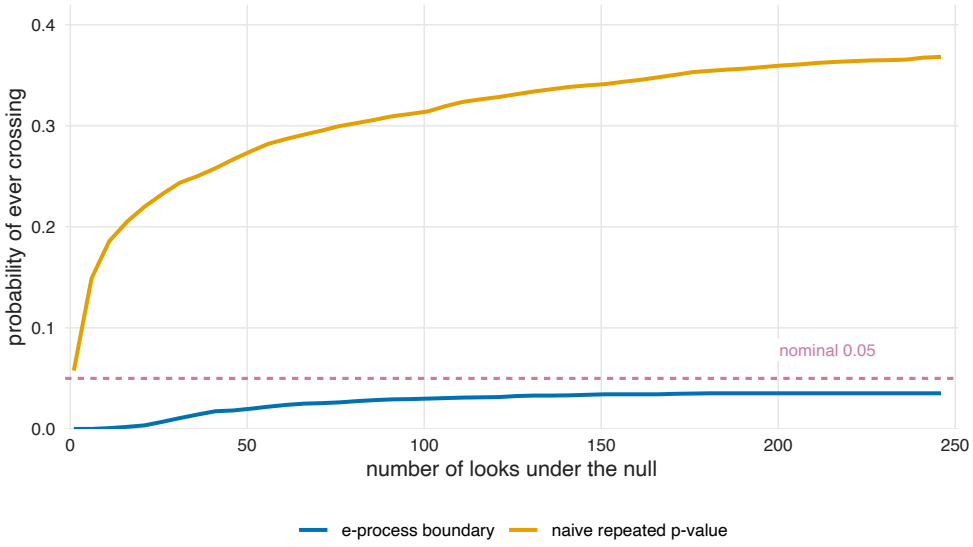

FIGURE A.1. A null simulation with 4000 Monte Carlo runs and 250 looks. Repeatedly checking a fixed-sample one-sided Gaussian p-value inflates the probability of ever crossing 0.05. The likelihood-ratio e-process boundary stays controlled because Ville's inequality applies to the whole monitored path.

Optional stopping versus optional continuation. Two distinct freedoms are protected here, and it pays to name them separately. *Optional stopping* concerns one monitored study: the analyst may stop at any stopping time τ and report M_τ. *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. E-values simply multiply; the running product acts as a continuously updated meta-analysis.

The next example shows what validity under an unknown stopping rule buys.

Example A.10: The design does not need to be known

A collaborator reports eight coin tosses,

$$1,\ 1,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,$$

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

$$\mathbb{P}_0(\text{stop by toss 8}) = \frac{1}{32} + \frac{3}{64} = \frac{5}{64} \approx 0.078.$$

Identical data, and the verdict at level 0.05 flips with the unverifiable design choice. Now compute the likelihood-ratio e-process against a design alternative with success probability $q > 1/2$:

$$M_8 = \prod_{j=1}^{8} \frac{q^{X_j}(1-q)^{1-X_j}}{1/2} = 2^8\, q^7(1-q),$$

which equals, for instance, 12.2 at $q = 0.9$. Because (M_t) is a test supermartingale under the null, this number is a valid e-value at any stopping time: the analyst does not need to know whether the collaborator used a fixed sample size, stopped on a streak, or stopped when the evidence looked promising. No method, however, survives hidden data: if unsuccessful runs of the whole experiment are discarded and only the best run is reported, p-values and e-values are broken alike.

One practical hazard of plain products deserves a fix. A single factor $E_s = 0$ sends the product M_t to zero permanently; all accumulated evidence, and all future evidence, is destroyed. The remedy is to hedge each factor toward one:

$$M_t = \prod_{s=1}^{t} \{1 - \lambda_s + \lambda_s E_s\}, \qquad \lambda_s \in [0,1] \text{ predictable.}$$

Each hedged factor is again a conditional e-value, since

$$\mathbb{E}_P[1 - \lambda_s + \lambda_s E_s \mid \mathcal{F}_{s-1}] = 1 - \lambda_s + \lambda_s\, \mathbb{E}_P[E_s \mid \mathcal{F}_{s-1}] \le 1,$$

so Proposition A.6 applies unchanged, and no single disastrous factor can zero out the account when $\lambda_s < 1$. The affine form is canonical. If a function g has the property that $g(E)$ is an e-value whenever E is, then g is dominated pointwise by $e \mapsto 1 - \lambda + \lambda e$ for some $\lambda \in [0, 1]$: affine shrinkage toward one is the only admissible way to post-process a single e-value [130, 170]. For instance, $\sqrt{E}$ is tempting as a way to soften zeros, and indeed $\sqrt{e} \le (1 + e)/2$ pointwise, so the hedge with $\lambda = 1/2$ dominates it. Data-driven choices of λ_s, tuned on the past to approximately maximize the growth rate, are developed by Waudby-Smith and Ramdas [175].

3. Universality of e-BH

Route: Advanced

Universality of e-BH. 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 α 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 α 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 $E_{(k)} \ge m/(\alpha k)$ 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 explains why compound e-values are the common structure underneath FDR control: any procedure with a valid FDR proof implicitly certifies condition (A) for its induced evidence vector. The threshold template of the next section is the constructive version of this statement, deriving the induced e-values in closed form for threshold-based rules. The universality viewpoint also gives a general recipe for aggregating several FDR procedures, of entirely different constructions, into one rejection list: convert each procedure ℓ 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 two sections below. 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 [85, 135]. A refinement worth knowing: if the original procedure is admissible among FDR procedures, meaning that no other level-α 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. [85] and Chapter 9 of Ramdas and Wang [130].

4. Asymptotic E-Values

Route: Advanced

Much of this book runs on test statistics that are only asymptotically normal: the debiased lasso statistics of Chapter 11 and the applied pipelines of Chapter 13 are leading examples. For such statistics a useful finite-sample e-value may be out of reach. For the unrestricted nonparametric null that the observations have mean zero and finite variance, Ramdas and Wang [130] report that no nonconstant nonasymptotic e-value is currently known. This is a statement about current knowledge, not an impossibility theorem. A bridge between this chapter and large-sample practice therefore needs a relaxed notion, together with a statement of what survives of the downstream guarantees.

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 (ε, δ)-*approximate e-value* for $\mathcal{H}_0$ if

$$\mathbb{E}_P[\min\{E, t\}] \le 1 + \varepsilon(P) + \delta(P)t \qquad \text{for every } t \ge 0 \text{ and every } P \in \mathcal{H}_0.$$

When ε and δ 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 slack parameters play different roles. For a fixed P, the parameter $\varepsilon(P)$ inflates the evidence budget, while $\delta(P)$ permits a rare wild event: if there is an event A_P with $\mathbb{P}_P(A_P) \le \delta(P)$ such that $\mathbb{E}_P[E\,\mathbf{1}_{A_P^c}] \le 1 + \varepsilon(P)$, 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 (ε, δ)-approximate. 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/α.

The central limit theorem 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 [130], 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 [130].

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 $X_1, \ldots, X_n$. Take the self-normalized statistic

$$Z^{(n)} = \frac{\sqrt{n}\,\bar{X}_n}{S_n}, \qquad S_n^2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2,$$

which converges in distribution to $N(0,1)$ under every null distribution, by the central limit theorem and Slutsky's lemma. Then

$$E^{(n)} = \exp\left\{\lambda \frac{\sqrt{n}\,\bar{X}_n}{S_n} - \frac{\lambda^2}{2}\right\}$$

satisfies the pointwise guarantee of Proposition A.14 under every null distribution, even though, as noted above, no nonconstant nonasymptotic e-value is currently known for this unrestricted null. Thus it is a pointwise asymptotic e-value. If the null is restricted to a class over which an appropriate self-normalized central limit theorem 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 point of the truncated definition is that the e-BH guarantee degrades gracefully rather than breaking.

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 (ε, δ)-approximate condition for fixed constants $\varepsilon \geq 0$ and $\delta \in [0,1]$. Then e-BH at level α 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 δ slack is multiplied by the number of true nulls: a rare wild event per hypothesis is not rare once m hypotheses can each have one. Consequently, if all true-null coordinates obey a common $(\varepsilon_n, \delta_n)$ bound, e-BH has asymptotic FDR control at level α provided $\varepsilon_n \to 0$ and $m\,\delta_n \to 0$. When m grows with n, the wild-event probability must shrink faster than $1/m$. This is the honest statement licensing central-limit-theorem e-values inside the e-BH pipelines of later chapters: the guarantee is asymptotic exactly because the e-values are.

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. This is the right abstraction for 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 how hard to test 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 τ, 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.

The common design template is simple. Before seeing the current evidence, choose a predictable testing level α_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 α_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. 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 α_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 α_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 α_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 $b_t \ge 0$ 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 α_t and the rewards.

LORD, short for significance Levels based On Recent Discovery, 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 $\sum_{j \ge 1} \gamma_j = 1$. Think of γ_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.

Proof outline for LORD-type rules. For true nulls,

$$\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] \le \alpha_t.$$

Therefore, for any deterministic horizon K,

$$\mathbb{E}[V_K] \le \mathbb{E}\left[\sum_{t\le K: t\in\mathcal{I}_0} \alpha_t\right] \le \mathbb{E}\left[\sum_{t\le K} \alpha_t\right].$$

The LORD design makes the cumulative spending $\sum_{t\le K}\alpha_t$ controlled by initial wealth plus rewards from R_K discoveries. This is why discoveries matter in the denominator of FDR: if there are few discoveries, the rule has little reward and cannot keep spending large levels; if there are many discoveries, the denominator $R_K \vee 1$ is larger. A full LORD theorem turns this accounting into a bound for $\mathbb{E}[V_K/(R_K \vee 1)]$, not just for $\mathbb{E}[V_K]$. Standard LORD FDR theorems also impose assumptions on how true-null p-values relate to the past, commonly independence or conditional super-uniformity strong enough for the rejection history being used. The proof has two moving parts: the p-value assumption controls false rejections at predictable levels, and the wealth equation prevents those predictable levels from growing too fast without discoveries.

SAFFRON, short for Serial estimate of the Alpha Fraction that is Futilely Rationed On true Null hypotheses, 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 λ_t with $\alpha_t \le \lambda_t \le 1$, and define

$$A_t = \mathbf{1}\{p_t \le \lambda_t\}.$$

A candidate is not necessarily rejected; it is a hypothesis whose p-value is small enough to be worth counting as a serious opportunity for rejection. 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.

For implementation, track four objects at every time: the predictable test level α_t, the predictable candidate threshold λ_t, the candidate indicator A_t, and the rejection indicator $I_t = \mathbf{1}\{p_t \le \alpha_t\}$. 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 null candidates.

E-value online rules: a clean e-LOND theorem. E-values give a particularly transparent online theorem. The p-value rules above must carefully ration α_t. For e-values, the test $E_t \ge 1/\alpha_t$ is controlled directly by conditional expectation. We do not need independence between the current evidence and the past; we need only conditional e-value validity. The resulting rule is called e-LOND, the e-value analogue of LOND, short for Levels based On Number of Discoveries, a simpler predecessor of LORD in which the testing level is proportional to the number of discoveries made so far.

The design is deliberately simple. Choose a deterministic spending schedule γ_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 γ_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 τ,

$$\mathbb{E}\left[\frac{V_\tau}{R_\tau \vee 1}\right] \le \alpha.$$

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 τ 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 α_t, R_{t-1}, and γ_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 θ 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 ϑ, 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_∞ read as the whole parameter space. Adaptedness makes this first-exclusion time a stopping time, and coverage at this τ 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 θ_0, suppose $(M_t^{\theta_0})_{t\geq 0}$ is a nonnegative supermartingale under the null model $H_0 : \theta = \theta_0$, with $M_0^{\theta_0} = 1$. 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 θ is excluded at some time if and only if

$$M_t^\theta \geq 1/\alpha$$

for at least one t. Since M_t^θ is a nonnegative supermartingale under the true distribution, Ville's inequality gives

$$\mathbb{P}_\theta\left(\sup_{t\geq 1} M_t^\theta \geq 1/\alpha\right) \leq \alpha.$$

Thus the probability of ever excluding the true θ is at most α. □

Example A.20: A simple bounded-mean confidence sequence

Let $X_1, X_2, \ldots \in [0,1]$ be independent with common mean θ. 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. For bounded means, 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 λ_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 θ_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.

The affine form of the factors $1 + \lambda_s(X_s - \theta_0)$ is not a stylistic choice. Each observation $X_s \in [0,1]$ yields two elementary conditional e-values for the null $\theta = \theta_0$, namely X_s/θ_0 and $(1 - X_s)/(1 - \theta_0)$, each with conditional expectation one, and, as noted in the discussion of hedged products, affine shrinkage toward one is the only admissible way to post-process an e-value. Shrinking the first e-value produces 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 recover exactly the displayed range of bets, so the betting construction uses the entire admissible family, not a convenient corner of it.

Confidence sequences also seed a multiple-testing story that this book takes up later. A family of e-values $E(\theta_0)$ indexed by the candidate value defines an *e-confidence interval* $\{\theta_0 : E(\theta_0) < 1/\alpha\}$, and a confidence sequence stopped at any stopping time is exactly of this form by the inversion theorem above. When intervals are reported only for parameters selected after looking at the data, the relevant error rate is the *false coverage rate*, the expected fraction of reported intervals that miss their targets: $\mathrm{FCR} = \mathbb{E}[|\{i \in S : \theta_i \notin C_i\}|/(|S| \vee 1)]$ for the selected set S. Reporting each e-confidence interval at level $\alpha_i = \delta|S|/m$, out of m candidates, controls the false coverage rate at level δ under arbitrary selection rules and arbitrary dependence, by an argument with the same three-line structure as Theorem 8.15 [177]. We return to selective confidence statements systematically in Chapter 12.

APPENDIX B

Advanced Conformal Theory

Route: Advanced

Chapter 10 contains the core full- and split-conformal algorithms and their principal coverage statements. This appendix collects proof mechanisms and extensions that can be deferred on a first reading.

1. What Conditional Coverage Means

Route: Advanced

The word "conditional" is used in several different ways in conformal prediction, and mixing them up is a common source of mistakes.

First, split conformal has a clean statement 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 is no longer guaranteed to be exactly $1 - \alpha$; 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. The answer is a classical Beta law. 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 $\mathbb{P}(T \leq t) \leq \mathbb{P}(T' \leq t)$ 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 [166]

Assume $Z_1, \ldots, Z_{n+1}$ 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 $\widehat{q} = R_{(k)}$, and

$$k = \lceil (1 - \alpha)(n_2 + 1) \rceil \leq n_2.$$

Let

$$G(t) = \mathbb{P}\{S(X, Y; \widehat{f}_{n_1}) \leq t \mid D_1\}$$

be the conditional distribution function of the score of a fresh draw, so that

$$T = G(\widehat{q}) = \mathbb{P}\{Y_{n+1} \in C_n(X_{n+1}) \mid D_1, D_2\}$$

is the realized coverage of the fitted prediction set. Then, conditional on D_1:

(a) T stochastically dominates a Beta$(k,\, n_2 + 1 - k)$ random variable;
(b) if G is continuous, then $T \mid D_1 \sim \text{Beta}(k,\, n_2 + 1 - k)$ exactly;
(c) for every $\Delta \geq 0$, $\mathbb{P}(T \leq 1 - \alpha - \Delta \mid D_1) \leq \exp(-2n_2\Delta^2)$.

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 is an order statistic in disguise. The function G is nondecreasing, so applying G preserves the ordering of the scores, and $\widehat{q} = R_{(k)}$ gives

$$T = G(R_{(k)}) = U_{(k)},$$

the kth smallest of $U_1, \ldots, U_{n_2}$.

Step 1: the transformed scores are super-uniform. For every $\tau \in [0, 1]$,

$$\mathbb{P}(U_1 \leq \tau \mid D_1) \leq \tau.$$

The inequality is trivial when $\tau = 1$, so take $\tau < 1$ and let $t_\tau = \sup\{t : G(t) \leq \tau\}$; 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 $G(t) \leq \tau$ for every $t < t_\tau$. If $G(t_\tau) \leq \tau$, then $\{G(R_1) \leq \tau\} \subseteq \{R_1 \leq t_\tau\}$, so $\mathbb{P}(U_1 \leq \tau \mid D_1) \leq G(t_\tau) \leq \tau$. If $G(t_\tau) > \tau$, then $\{G(R_1) \leq \tau\} \subseteq \{R_1 < t_\tau\}$, whose probability is $\lim_{t \uparrow t_\tau} G(t) \leq \tau$. 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 $\mathbb{P}(U_1 \leq \tau \mid D_1) \geq \mathbb{P}(R_1 \leq t \mid D_1) = \tau$, and combining with super-uniformity gives $\mathbb{P}(U_1 \leq \tau \mid D_1) = \tau$.

Step 2: a binomial representation. For $\tau \in [0, 1]$, the event $\{U_{(k)} \leq \tau\}$ states that at least k of the n_2 variables U_i fall at or below τ. The number of such variables is a binomial count, $\text{Bin}(n_2, p_\tau)$ with $p_\tau = \mathbb{P}(U_1 \leq \tau \mid D_1)$, so

$$\mathbb{P}(U_{(k)} \leq \tau \mid D_1) = \mathbb{P}\{\text{Bin}(n_2, p_\tau) \geq k\}.$$

A binomial count is stochastically increasing in its success probability: realizing the ith Bernoulli variable as $\mathbf{1}\{V_i \leq p\}$ for i.i.d. uniform $V_1, \ldots, V_{n_2}$ couples the counts at two success probabilities so that the count with the larger p is pointwise larger. Since $p_\tau \leq \tau$ by Step 1,

$$\mathbb{P}(U_{(k)} \leq \tau \mid D_1) \leq \mathbb{P}\{\text{Bin}(n_2, \tau) \geq k\}.$$

Step 3: uniform order statistics are Beta. Let $U_1^*, \ldots, U_{n_2}^*$ be i.i.d. uniform on $(0, 1)$. The same counting identity gives, for $t \in [0, 1]$,

$$\mathbb{P}\{U_{(k)}^* \leq t\} = \mathbb{P}\{\text{Bin}(n_2, t) \geq k\} = \sum_{j=k}^{n_2} \binom{n_2}{j} t^j (1-t)^{n_2 - j}.$$

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:

$$\frac{d}{dt}\,\mathbb{P}\{U_{(k)}^* \leq t\} = n_2 \binom{n_2 - 1}{k - 1} t^{k-1}(1-t)^{n_2-k} = \frac{n_2!}{(k-1)!\,(n_2-k)!}\, t^{k-1}(1-t)^{n_2-k},$$

which is the Beta$(k, n_2 + 1 - k)$ density. Hence $U^*_{(k)} \sim \text{Beta}(k, n_2 + 1 - k)$, 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 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 $\tau = 1 - \alpha - \Delta$,

$$\mathbb{P}(T \le 1 - \alpha - \Delta \mid D_1) \le \mathbb{P}\{\text{Bin}(n_2, 1 - \alpha - \Delta) \ge k\}.$$

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 $\mathbb{P}\{\text{Bin}(n_2, p) \ge n_2 p + s\} \le \exp(-2s^2/n_2)$ for every $s \ge 0$; the same inequality reappears in the Learn-Then-Test calibration of Section 7. Taking $s = n_2\Delta$ gives $\exp(-2n_2\Delta^2)$, which is part (c). □

The theorem says that the realized coverage of a fitted split conformal set is not the constant $1 - \alpha$; it is a random variable whose distribution is pinned down by the calibration size alone. The 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 at any finite n_2 it falls below $1 - \alpha$ with appreciable probability on unlucky calibration draws. Larger calibration sets do not change the marginal guarantee, which holds at every finite n_2; they tighten the distribution of the realized coverage around it [4, 166].

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 α' 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 δ as soon as

$$\alpha - \alpha' \ge \sqrt{\frac{\log(1/\delta)}{2n_2}}.$$

In words: a high-probability guarantee over the calibration draw costs only a reduction of the nominal level of order $\sqrt{\log(1/\delta)/n_2}$, which vanishes as the calibration set grows [166]. If the displayed feasibility condition fails, this tail bound does not supply a nontrivial positive working level α'; 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 α, even though marginal coverage continues to hold [25]; this is

one of the few places in the theory where the independence assumption genuinely buys more than exchangeability.

Test-conditional coverage is impossible. Third, and strongest, is coverage conditional on the exact covariate value:

$$\mathbb{P}\{Y_{n+1} \in C_n(x) \mid X_{n+1} = x\} \geq 1 - \alpha \qquad \text{for every } x.$$

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 [12, 105]

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 ϵ.

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 ith term changes the ith 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 exactly tight, and it can be read as saying that, at each fixed covariate value of a nonatomic feature distribution, any distribution-free test-conditional procedure is no more informative than this trivial construction. The conclusion holds for every distribution with a nonatomic covariate marginal, not merely for some worst-case distribution: there is no class of nice distributions on which such a procedure gives short intervals while paying the price only elsewhere.

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 transfer free. This template is a general tool 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. The impossibility theorem closes one door and points to several others. Distribution-free conditional coverage survives in weakened forms in which the conditioning event has positive probability. 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 [171]. 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 chapter, which gives marginal coverage under the target distribution [12, 105].

2. The conformal PRDS argument

Route: Optional proof

Proof of Theorem 10.6

Consider first p_1 and assume $Z_{n+1} \sim P$. Let

$$S_{(1)} \le \cdots \le S_{(n+1)}$$

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

$Z_{n+2}, \dots, Z_{n+m}$ changes nothing, because those points are independent of $(Z_1, \dots, Z_{n+1})$. 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 $\{1, \dots, n+1\}$. If $p_1 = k/(n+1)$, then $S(Z_{n+1}) = S_{(k)}$, and the reference scores used to compute $p_2, \dots, p_m$ 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

$$N_j(k) = \#\{\ell : S_{(\ell)} \le r\} - \mathbf{1}\{S_{(k)} \le r\}.$$

If k is increased, the indicator $\mathbf{1}\{S_{(k)} \le r\}$ 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 $(S_{(1)}, \dots, S_{(n+1)})$ 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 $(p_1, \dots, p_m)$ 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. □

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.

There is a conformal fix when the selection rule treats the calibration and test points symmetrically. To state the clean and most common version, 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 $\widehat{f}(X_i) \ge c$. 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_Ith 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

$$C_I(x) = \{y : S(x, y) \le \widehat{q}_I\}.$$

Theorem B.4: Selection-conditional conformal coverage [93]

Under exchangeability of the calibration and test observations conditional on D_1, and for a symmetric selection rule I as above,

$$\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I\} \ge 1 - \alpha,$$

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 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 σ 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 σ 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 $I = \{i : \widehat{f}(X_i) \ge c\}$ is symmetric, and so is the rule that selects the points with the k largest predicted responses.

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. 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 σ 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 σ 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 σ 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 $N \leq r-k$. 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 $\{S_{n+1} \leq \widehat{q}_{I_0}\}$.

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 $\{S_{n+1} \leq S_{(k)}\}$ says that the test score is among the k smallest of these $r+1$ exchangeable scores, so Proposition 10.2 gives

$$\mathbb{P}\{S_{n+1} \leq S_{(k)} \mid I = I_0\} \geq \frac{k}{r+1} \geq 1-\alpha,$$

equivalently $\mathbb{P}\{p \leq \alpha \mid I = I_0\} \leq \alpha$.

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,

$$\mathbb{P}\{p \leq \alpha,\ n+1 \in I\} = \sum_{I_0 \ni n+1} \mathbb{P}\{I = I_0\}\, \mathbb{P}\{p \leq \alpha \mid I = I_0\} \leq \alpha\, \mathbb{P}\{n+1 \in I\},$$

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 face value. The bound holds for every α simultaneously, and no continuity of the scores is required; ties only make the p-value larger, hence more conservative. With the corollary in hand, the theorem is a restatement.

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

$$\{Y_{n+1} \in C_I(X_{n+1})\} = \{S(X_{n+1}, Y_{n+1}) \leq \widehat{q}_I\}.$$

On each selection event $\{I = I_0\}$ with $n+1 \in I_0$, the deterministic equivalences in the proof of Corollary B.6 identify $\{S_{n+1} \leq \widehat{q}_{I_0}\}$ 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

$$\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I,\, D_1\} = \mathbb{P}\{p > \alpha \mid n+1 \in I,\, D_1\} \geq 1-\alpha.$$

This is the conditional-on-D_1 form of the theorem. For the unconditional form, note the joint bound $\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}),\, n+1 \in I \mid D_1\} \geq (1-\alpha)\, \mathbb{P}\{n+1 \in I \mid D_1\}$, 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 $\mathbb{P}\{Y_{n+1} \in C_I(X_{n+1}) \mid n+1 \in I\} \geq 1-\alpha$. □

The lemma-and-corollary route follows the textbook treatment of selective coverage in Angelopoulos et al. [4]; the result and its extensions to more general selection rules are developed by Jin and Ren [93].

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 [7, 92, 93].

Group-conditional coverage as a special case. Let $g : \mathcal{X} \to \{1, \ldots, G\}$ 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 [171]. 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, which is the theme of Chapter 12.

4. The jackknife+ tournament argument

Route: Optional proof

The tournament argument. The proof rests on a single combinatorial object. 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

$$A_{ij} + A_{ji} \leq 1 \qquad \text{for all } i, j \in \{1, \ldots, N\},$$

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 [4].

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 heavy losers; 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 is a double count. Each heavy loser needs 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, roughly $|J|^2/2$ losses in total. A large set of heavy losers cannot find enough games to lose.

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 player $n+1$ to lose heavily. Suppose first that $Y_{n+1} > q_{\mathrm{up}}$. This forces $q_{\mathrm{up}} < \infty$, so q_{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_{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: exchangeability makes the losing row uniform. 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 σ 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 σ. Fix $i \in \{1, \ldots, n+1\}$ and choose σ 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 $n+1$ is a heavy loser. 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,

$$\begin{aligned}\mathbb{P}\{Y_{n+1} \notin [q_{\text{low}}, q_{\text{up}}]\} &\le \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\Big\{\sum_{j=1}^{n+1} A_{ij} \ge (1-\alpha)(n+1)\Big\} \\ &= \mathbb{E}\Big[\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\Big\{\sum_{j=1}^{n+1} A_{ij} \ge (1-\alpha)(n+1)\Big\}\Big].\end{aligned}$$

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 2α is the price of the tournament bound, not of a union bound over the two tails. The tournament route follows Angelopoulos et al. [4]; the original proof of Barber et al. [11] bounds the same event by a closely related combinatorial argument. Stronger statements are possible under additional stability assumptions, but the robust message stands: jackknife+ restores a finite-sample guarantee, at level $1 - 2\alpha$, that the ordinary jackknife does not have in general.

Why the factor of two. It is natural to ask whether the constant 2α 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. A single score list induces a transitive comparison structure, that is, a total order: 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 precisely when its score is in the top α fraction of the list, so at most $\alpha(n+1)$ players can be heavy losers. 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 $2m + 1 = 2\alpha(n+1) - 1$ heavy losers, nearly twice the transitive count [4]. 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 [4, 11]. 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 [138, 169], and this doubling is exactly where the $1 - 2\alpha$ level arises in the p-value route to

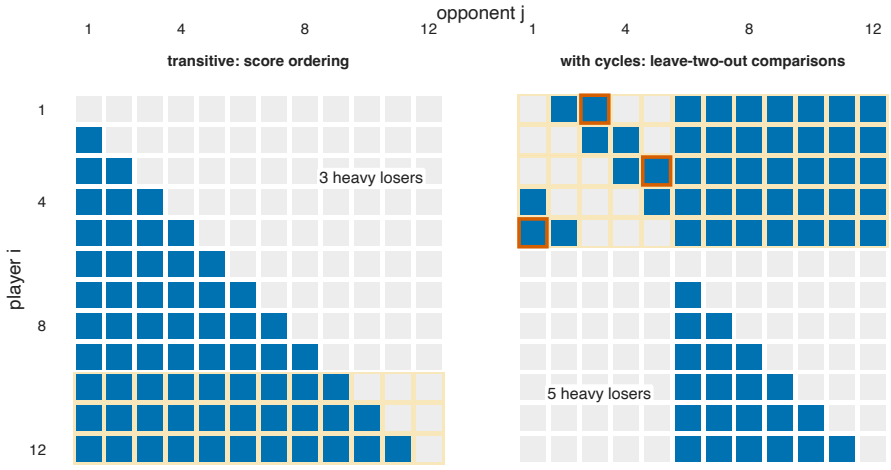

FIGURE B.1. Transitive versus cyclic tournaments. Left: a single score list, as in full conformal prediction, induces a total order. Players are displayed in score order, best conformity first, so losses fill a staircase, and a player can lose at least $(1-\alpha)(n+1)$ games only by having one of the $\alpha(n+1)$ worst scores; these are the bottom rows, so at most $\alpha(n+1)$ players are heavy losers. 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.

the cross-conformal guarantee [4]. By contrast, the average of e-values is always a valid e-value (Chapter 8), which is one motivation for the conformal e-prediction methods discussed later in this chapter.

5. Online Conformal P-Values and Testing Exchangeability

Route: Research frontier

The chapter has so far calibrated once, on a batch. Many deployments instead see a stream. Data arrive one at a time, and each new point must be predicted from all earlier ones: a bank scores transactions as they occur, a hospital monitors each incoming patient, a deployed model is checked on each new query. 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 this chapter 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 $\mathbb{P}\{Y_t \in C_t(X_t)\} \geq 1 - \alpha$. This is not new. The new question is the joint behavior over time. Every observation is reused as training data at every later step, so the coverage errors at different times share data and look as though they should be dependent in a

complicated way. The surprise of this section is that, 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 rest of the chapter. 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 σ 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 tth 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 [165, 172].

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. The proof establishes 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 kth largest score, so

$$p_t \mid \mathcal{G}_t \ \sim \ \text{uniform on } \{1/t, 2/t, \dots, 1\}.$$

Claim 2: for every $s > t$, the p-value p_s is a function of $\mathcal{G}_t$. Let σ be any permutation of $\{1, \dots, T\}$ that permutes $\{1, \dots, t\}$ and fixes every index larger than t, and relabel the stream by σ. 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 $Z_{t+1}, \dots, Z_T$, which is the claim.

Combining the claims. By Claim 2, the vector $(p_{t+1}, \dots, p_T)$ is a function of $\mathcal{G}_t$. Claim 1 and the tower property of conditional expectation therefore give

$$p_t \mid (p_{t+1}, \dots, p_T) \ \sim \ \text{uniform on } \{1/t, 2/t, \dots, 1\},$$

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 $a_t \in \{1/t, \dots, 1\}$, the chain rule of conditional probability gives

$$\mathbb{P}(p_1 = a_1, \dots, p_T = a_T) = \prod_{t=1}^{T} \mathbb{P}(p_t = a_t \mid p_{t+1} = a_{t+1}, \dots, p_T = a_T) = \prod_{t=1}^{T} \frac{1}{t}.$$

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. □

Three remarks keep the theorem honest. 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 $Z_{t+1}, \dots, Z_T$ are fixed, all later p-values are determined, while the position of the tth 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. Independence is bought by refitting on all data at every step, not by conformal prediction as such. 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 [4, 172].

Corollary B.9: Independent errors and long-run coverage

Suppose $Z_1, Z_2, \dots$ 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 α. Then $\mathrm{err}_1, \mathrm{err}_2, \dots$ 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 $\{1/t, \dots, 1\}$, the number of grid points at most α 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 α. Hence the miscoverage frequency converges almost surely to α, which is the displayed statement. □

Three dependence regimes. It is worth pausing to organize how conformal p-values interact with the multiple-testing machinery of this book, because the dependence structure is decided by 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, and the Benjamini–Hochberg procedure [19] applies exactly as in the independent case. (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 Benjamini–Yekutieli correction [20], or on conformal e-values combined with the e-BH procedure of Chapter 8 [14, 104, 118]. 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 $M_0, M_1, \dots$ 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 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 α, so the process can be monitored at every time without a multiplicity penalty.

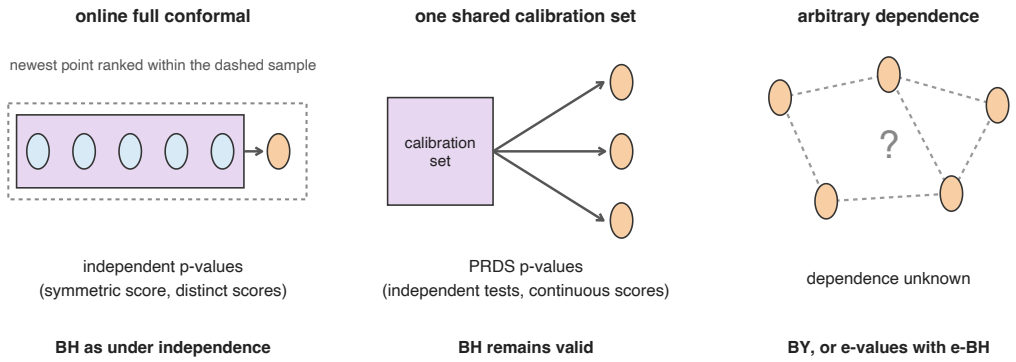

FIGURE B.2. Three dependence regimes for conformal p-values, and the matching multiple-testing tool. Online full conformal retraining with a symmetric score and distinct scores gives exactly independent p-values, so BH applies as under independence. Fresh independent smoothing gives the analogous conclusion 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 Benjamini–Yekutieli 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 on which every guarantee of this chapter rests, and in deployment it is the assumption most likely to 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.

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 α 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 Chapter 8) gives $\mathbb{P}(\sup_T M_T \ge 1/\alpha) \le \alpha$. For any stopping rule τ, 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 λ_t, which may be tuned from the past p-values, controls how aggressively the test bets on small p-values [172].

Three punchlines summarize why this construction is remarkable. First, the p-values are computed from the same growing data stream, with each observation reused at every later step, yet Theorem B.8 means that no multiplicity correction for data reuse is needed: each factor $f_t(p_t)$ is a valid conditional e-value given the past. Second, the division of labor is clean: conformal prediction supplies the e-process, and Chapter 8 supplies the betting interpretation, the optional-stopping guarantee, and the reading of large M_T as accumulated evidence against exchangeability. Third, this is a monitoring tool in the strict sense: it detects when exchangeability fails, which is exactly the assumption whose failure the next section addresses. A word of caution completes the picture: failing to reject is not evidence that the stream is exchangeable, and the test has power only against departures that make the chosen score produce unusually small conformal p-values.

Finally, when the data are not merely drifting but adversarially nonstationary, no guarantee based on exchangeability survives, and a different idea takes over: 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 [66]. Its guarantee is deterministic and long-run: the average miscoverage over time approaches α for every sequence, random or

not. This is a different kind of statement from the distributional coverage of this chapter, and we do not develop it here [4].

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.

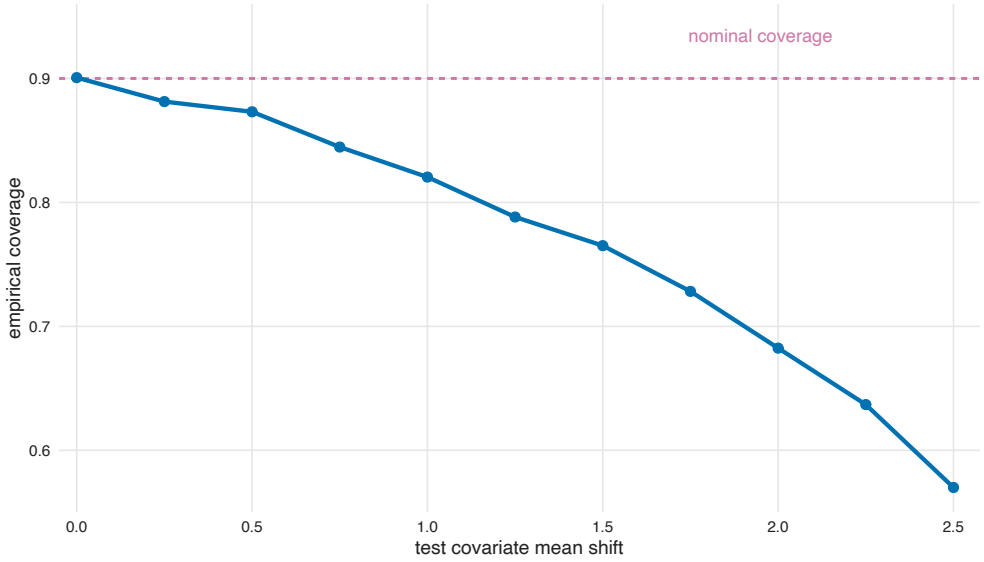

FIGURE B.3. A split conformal interval calibrated for nominal 90% coverage under one covariate distribution can undercover after test covariate shift. The issue is not model misspecification alone; the exchangeable-rank argument no longer applies.

Weighted and distribution-shift conformal methods modify the calibration weights or the target coverage statement when the shift structure is known or estimable. Those methods require extra assumptions. Without such structure, finite-sample conditional or shifted-distribution coverage cannot be obtained for free.

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 [160] 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 assume throughout that it is finite at every point. 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. The weighted distribution is

$$\widehat{\mathcal{F}}_y^w = \sum_{i=1}^{n} \frac{w(X_i, Y_i)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)} \delta_{S_i} + \frac{w(X_{n+1}, y)}{\sum_{\ell=1}^{n} w(X_\ell, Y_\ell) + w(X_{n+1}, y)} \delta_{+\infty}.$$

The finite-score part of this distribution has CDF

$$\widehat{F}_n^w(s) = \frac{\sum_{i=1}^n w(X_i, Y_i)\,\mathbf{1}\{S_i \le s\}}{\sum_{i=1}^n w(X_i, Y_i) + w(X_{n+1}, y)}.$$

For a finite collection of atoms $z_1, \dots, z_m \in \mathbb{R} \cup \{+\infty\}$ carrying nonnegative weights $w_1, \dots, w_m$ 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 τ, with the convention that the infimum of the empty set is $+\infty$. The weighted conformal prediction set is

$$\widehat{C}_n^w(x) = \{y : S(x, y) \le \text{Quantile}(\widehat{\mathcal{F}}_y^w;\, 1 - \alpha)\}.$$

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 [75], 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 $z_1, \dots, z_m$ and nonnegative weights $w_1, \dots, w_m$ 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 jth 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 jth 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_{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. 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 $S_{n+1} \leq \text{Quantile}(\widehat{\mathcal{F}}_{Y_{n+1}}^w; 1-\alpha)$ contains the event

$$E = \Big\{S_{n+1} \leq \text{Quantile}\Big(\sum_{i=1}^{n+1} p_i \delta_{S_i};\, 1-\alpha\Big)\Big\}.$$

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

$$\mathbb{P}(E \mid \{Z_1, \dots, Z_{n+1}\}) = \sum_{j=1}^{n+1} p_j\, \mathbf{1}\Big\{S_j \leq \text{Quantile}\Big(\sum_{i=1}^{n+1} p_i \delta_{S_i};\, 1-\alpha\Big)\Big\} \geq 1 - \alpha,$$

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 special cases lead the applications. Covariate shift, where the test and calibration laws share the conditional distribution of Y given X so that the likelihood ratio $w(x, y) = w(x)$ depends only on the covariate, is the setting of Tibshirani et al. [160]; label shift, where the two laws share the conditional distribution of X given Y and $w(x, y) = w(y)$ depends only on the response, is handled by the same theorem, the unknown test label being supplied by the candidate value y inside $\widehat{\mathcal{F}}_y^w$. 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 [13]. 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 [13]

Let $w_1, \dots, w_{n+1}$ be fixed nonnegative constants with $\sum_{i=1}^{n+1} w_i = 1$ and $w_{n+1} \geq w_i$ for every $i \leq n$. 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

$$C^w(x) = \Big\{ y : S(x,y) \leq \text{Quantile}\Big(\sum_{i=1}^{n} w_i \delta_{S_i} + w_{n+1} \delta_{+\infty};\, 1-\alpha \Big) \Big\}.$$

Then

$$\mathbb{P}\{Y_{n+1} \in C^w(X_{n+1})\} \geq 1 - \alpha - \sum_{i=1}^{n} w_i \, d_{\mathrm{TV}}(Z,\, Z^{(i)}),$$

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. [13], 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 $(S_1, \dots, S_{n+1})$, 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; comparing the two quantile events uses a mixture generalization of the quantile comparison, stated as an exercise at the end of this chapter, and it is here that 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 certifies 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 guarantee has the same validity-minus-drift structure as the dependence-robust false discovery rate bounds of Chapter 6: an exact guarantee under an ideal assumption degrades continuously, with an explicit penalty, as the assumption fails, and choosing weights that decay with the age of an observation trades statistical efficiency against robustness to drift. The penalty involves the unknown joint law, so the bound quantifies robustness rather than certifying an achieved coverage level from data. Localized variants that reweight calibration points by proximity to the test covariate deliver approximate test-conditional coverage by the same weighting mechanism [71, 81].

7. Conformal Risk Control and Learn-Then-Test

Route: Research frontier

The conformal guarantee controls miscoverage, the expectation of a 0/1 loss. Many modern deployments care about quantities that are not miscoverage: 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. [3] introduce *conformal risk control* (CRC), which generalizes coverage control to the expectation of any bounded, monotone loss function.

Theorem B.15: Conformal risk control

Let λ 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 $0 \leq L_\lambda(X, Y) \leq B$. 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

$$\frac{1}{n+1}\left(\sum_{i=1}^{n} L_\lambda(X_i, Y_i) + B\right) \leq \alpha.$$

Then the test-point loss satisfies

$$\mathbb{E}[L_{\widehat{\lambda}}(X_{n+1}, Y_{n+1})] \leq \alpha$$

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 $\alpha \geq B/(n+1)$ 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 $i \in \{1, \dots, n+1\}$, let $\widehat{\lambda}^{(-i)}$ be the threshold obtained by applying the same rule to all observations except Z_i:

$$\frac{1}{n+1}\left(\sum_{\ell \neq i} L_\lambda(Z_\ell) + B\right) \le \alpha.$$

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

$$\sum_{i=1}^{n+1} L_{\widehat{\lambda}^{(-i)}}(Z_i) \le (n+1)\alpha.$$

Let i_* be an index for which $\widehat{\lambda}^{(-i_*)}$ is smallest. Since losses are non-increasing in λ,

$$L_{\widehat{\lambda}^{(-i)}}(Z_i) \le L_{\widehat{\lambda}^{(-i_*)}}(Z_i) \qquad \text{for every } i.$$

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 the calibration sample certify a bound that transfers to the test point. When $L_\lambda(X, Y) = \mathbf{1}\{Y \notin C_\lambda(X)\}$ and C_λ is the prediction set indexed by quantile level λ, 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 with no change in the proof; only the requirement that the loss be monotone in λ and uniformly bounded.

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 *Learn-Then-Test* (LTT) framework of Angelopoulos et al. [2] reduces calibration of an arbitrary risk-controlling parameter to a multiple-testing problem.

Fix a finite grid $\Lambda = \{\lambda_1, \dots, \lambda_K\}$, 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 λ, 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_λ. 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

$$p_\lambda^{\text{Hoeff}} = \exp\left(-\frac{2n}{B^2}(\alpha - \bar{L}_\lambda)_+^2\right), \qquad \bar{L}_\lambda = \frac{1}{n}\sum_{i=1}^{n} L_\lambda(X_i, Y_i),$$

which is super-uniform under H_λ (a direct application of Hoeffding to $\bar{L}_\lambda$ bounded around its mean). Sharper Bernstein- or e-value-based p-values can be used when the loss has small variance or when the experimenter prefers anytime-valid certificates.

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_λ is a valid super-uniform p-value for H_λ under the calibration sample. Apply Bonferroni at level δ to the family $\{p_\lambda : \lambda \in \Lambda\}$, so $\widehat{\Lambda}_{\text{Bonf}} = \{\lambda : p_\lambda \le \delta/K\}$. If this set is nonempty, then for any selection rule $\widehat{\lambda} \in \widehat{\Lambda}_{\text{Bonf}}$,

$$\mathbb{P}\Big(R(\widehat{\lambda}) \le \alpha\Big) \ge 1 - \delta.$$

If $\widehat{\Lambda}_{\text{Bonf}}$ is empty, the procedure abstains.

Proof of Theorem B.16

The event "$\widehat{\Lambda}_{\text{Bonf}}$ contains a null λ" 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_λ under H_λ. Hence with probability at least $1 - \delta$, every λ 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 $\widehat{\Lambda}_{\text{Bonf}}$. □

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 λ's that fail to satisfy $R(\lambda) \le \alpha$ is at most δ. For bounded losses calibrated with Hoeffding p-values, the concentration scale at familywise level δ is $B\sqrt{\log(K/\delta)/n}$, matching the usual uniform-concentration rate.

The interplay between LTT and the rest of this book is direct. 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. [15] are a special case in which the test is constructed from a known concentration inequality.

APPENDIX C

Method-Selection Matrix

The matrix is a routing device, not a substitute for the theorem statements. Start with the target and output columns; then check whether the required model or dependence condition describes the intended data-generating process.

Method	Target	Output	Guarantee	Status	Dependence/model requirement	Main failure mode	Conservative fallback
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 q-values	Flat family with estimable null fraction	Adaptive rejection set, q-values	Procedure-specific FDR	Model-based	Independence or stated tail-estimation conditions	Alternatives contaminate the estimation tail	Ordinary BH or BY
Local FDR	Posterior null status	Posterior probabilities/ranking	Model-based posterior interpretation	Model-based	Correct two-groups model and density estimation	Misspecified mixture/null density	Frequentist BH with valid p-values
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	Valid layer p-values and stated dependence	Interpreted as recursive selected-family control	Layerwise Bonferroni
TreeBH	Tree-structured selected families	Rejected branches/leaves	Recursive selected-family FDR	Finite	Prespecified tree and valid family p-values	Read as unconditional levelwise FDR	Hierarchical Holm/closed testing
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 under optional continuation	E-value/rejection	Type I error $\leq \alpha$ for $E \geq 1/\alpha$	Finite	Null expectation ≤ 1	Post-hoc construction breaks expectation budget	Fixed-horizon p-test
E-process	Sequential null	Anytime evidence; confidence sequence	Time-uniform Type I error	Finite	Conditional e-values and predictable bets	Hindsight tuning or filtration errors	Prespecified fixed looks with alpha spending

Method	Target	Output	Guarantee	Status	Dependence/model requirement	Main failure mode	Conservative fallback
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	Sensitivity across feature models
Fixed-X knockoffs	Coefficient nulls in fixed Gaussian linear model	Selected features	FDR at target q	Finite	Fixed X, Gaussian linear errors, Gram construction	Conditioning on y or target confusion	Holm/debiased prespecified contrasts
Model-X knockoffs	Conditional feature nulls	Selected features	FDR at target q	Finite	Exchangeable knockoffs from correct feature law	Approximate or nonexchangeable knockoffs	CRT or BY with sensitivity analysis
Split conformal	Next response	Prediction set	Marginal coverage $\geq 1-\alpha$	Finite	Exchangeability and fixed score pipeline	Read as conditional coverage; tuning after calibration	Wider recalibrated set
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	Remainder, leverage, variance mismatch	Sample splitting or lower-dimensional target
FCR adjustment	Fixed parameters, selected reporting set	Selected intervals	False coverage rate $\leq \alpha$	Finite	Valid marginal intervals and selection rule conditions	Read as per-selected-interval conditional coverage	Simultaneous intervals
POSI	All selected-model projection targets	Intervals after arbitrary selection	Simultaneous coverage	Finite, model-based	Gaussian linear mean and simultaneous constant	Wide intervals or structural-target confusion	Sample splitting
Polyhedral SI	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	Data carving or splitting
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 deployment pairs, sound reference relation	Interpreted as truth or pointwise coverage	Abstention/narrower claim

APPENDIX D

Glossary

Familywise error rate (FWER): The probability of at least one false rejection in a specified family: FWER $= \mathbb{P}(V \geq 1)$.

False discovery rate (FDR): The expected false discovery proportion: FDR $= E\{V/(R \vee 1)\}$.

Positive FDR (pFDR): The expected false discovery proportion conditional on making at least one rejection: $E(V/R \mid R > 0)$.

Marginal FDR (mFDR): The ratio of expected false to expected total discoveries, commonly $E(V)/E(R)$, with the exact denominator convention stated by the method.

False coverage rate (FCR): The expected proportion of noncovering intervals among the intervals selected for reporting.

Selected-family FDR: An average FDR across data-selected families, usually conditional or weighted according to a hierarchical selection path. It is not automatically an unconditional levelwise FDR.

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.

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. PRDS plus super-uniform nulls licenses ordinary BH.

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 null expectation at most one. The rule $E \geq 1/\alpha$ is a valid level-α e-test, not a numerical p-value conversion.

E-process: A nonnegative process dominated under the null by a test supermartingale, usually normalized at one. Ville's inequality gives time-uniform evidence.

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 μ and selected design X_M, the coefficient $(X_M^\top X_M)^{-1} X_M^\top \mu$. It is not automatically a population or structural regression coefficient.

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, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the 12th International Conference on Learning Representations*. ICLR, 2024.

[4] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. Cambridge University Press, 2026. Pre-publication version.

[5] 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.

[6] Ander Artola Velasco, Ivi Chatzi, Renato Vieira, Stratis Tsirtsis, and Manuel Gomez Rodriguez. Tokenization multiplicity leads to arbitrary price variation in LLM-as-a-service. arXiv preprint arXiv:2506.06446, 2025. Preprint.

[7] Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Selective conformal inference with false coverage-statement rate control. *Biometrika*, page asae010, 2024.

[8] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

[26] 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.

[27] Gilles Blanchard and Etienne Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10:2837–2871, 2009.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] T. Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_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.

[35] 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.

[36] 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.

[37] 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.

[38] Yiqun T. Chen and Daniela M. Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

[39] 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.

[40] Giovanni Cherubin. Majority vote ensembles of conformal predictors. *Machine Learning*, 108(3):475–488, 2019.

[41] Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplement Issue*, 1:205–237, 1984.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] Sandrine Dudoit and Mark J. van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, 2008.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Jaiden Fairoze, Jiwon Kim, Lichao Yang, Sanjam Garg, Mingyuan Ma, and Dawn Song. Verifiable and oblivious watermark detection for large language models. arXiv preprint arXiv:2509.27666, 2025. Preprint.

[56] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 4 edition, 1932.

[57] 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.

[58] William Fithian, Dennis L. Sun, and Jonathan Taylor. Optimal inference after model selection, 2017.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote. arXiv preprint arXiv:2401.09379, 2024.

[64] Etienne Gauthier, Francis Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction. arXiv preprint arXiv:2503.13050, 2025. Preprint.

[65] 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.

[66] Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[67] 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.

[68] 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.

[69] 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.

[70] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

[71] Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

[72] 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.

[73] 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.

[74] Ruijian Han, Lan Luo, Yiming Lin, and Jian Huang. Adaptive debiased Lasso in high-dimensional GLMs with streaming data. arXiv preprint arXiv:2405.18284, 2024. Preprint.

[75] Matthew T Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

[76] Ruth Heller and Marina Bogomolov. Replicability across multiple studies. *Statistical Science*, 38(4):602–620, 2023. doi: 10.1214/23-STS890.

[77] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988. doi: 10.1093/biomet/75.4.800.

[78] Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2): 901–928, 2023.

[79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] Nikolaos Ignatiadis, Ruodu Wang, and Aaditya Ramdas. Asymptotic and compound e-values: Multiple testing and empirical bayes. arXiv preprint arXiv:2409.19812, 2024.

[86] Yu. I. Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[87] 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.

[88] 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.

[89] 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.

[90] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[91] 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.

[92] Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000. doi: 10.1214/aos/1015957397.

[98] Nick W. Koning. Post-hoc α hypothesis testing and the post-hoc p-value. arXiv preprint arXiv:2312.08040, 2024.

[99] 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.

[100] 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.

[101] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

[102] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2 edition, 2000.

[103] 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.

[104] Junu Lee and Zhimei Ren. Boosting e-BH via conditional calibration. *arXiv preprint arXiv:2404.17562*, 2024.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] Guanxun Li and Xianyang Zhang. A note on e-values and multiple testing. *Biometrika*, 112(1):asae050, 2025. doi: 10.1093/biomet/asae050.

[113] Jonathan Q. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

[120] Christoforos Mavrogiannis, Angelos Mavrogiannis, and Constantine Dovrolis. How independent are large language models? A statistical framework for auditing behavioral entanglement and reweighting verifier ensembles. arXiv preprint arXiv:2410.07650, 2024. Preprint.

[121] 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.

[122] 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.

[123] National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, 2023.

[124] 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.

[125] 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.

[126] Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. Proving test set contamination in black box language models, 2023.

[127] 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.

[128] Mehrdad Pournaderi and Yu Xiang. Differentially private model-X knockoffs via johnson–lindenstrauss transform. arXiv preprint arXiv:2508.04800, 2025. Preprint.

[129] 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.

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[137] 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.

[138] Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

[139] Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking, 2025.

[140] 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.

[141] 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.

[142] 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.

[143] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64(3):479–498, 2002.

[150] 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.

[151] 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.

[152] 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.

[153] Hanqi Sun, Xiaoyu Wu, Jiansheng Maddu, Yuping Yu, Wenliang Wang, and Sarath Ramnath. Auditing multi-agent LLM reasoning trees outperforms majority vote and LLM-as-judge. arXiv preprint arXiv:2410.09341, 2024. Preprint.

[154] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. doi: 10.1093/biomet/ass043.

[155] 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.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[164] 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.

[165] 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.

[166] 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.

[167] Vladimir Vovk. Transductive conformal predictors. In *Proceedings of the International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360, 2013.

[168] Vladimir Vovk and Ivan Petej. Venn–Abers predictors. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.

[169] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4): 791–808, 2020.

[170] 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.

[171] 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.

[172] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[173] 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.

[174] 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.

[175] 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.

[176] 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.

[177] 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.

[178] Daniel Yekutieli. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008. doi: 10.1198/016214507000001373.

[179] 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.

[180] Xianyang Zhang. Testing High Dimensional Mean Under Sparsity, 2015. URL https://arxiv.org/abs/1509.08444. arXiv:1509.08444v2.

[181] Xianyang Zhang. Mirror and Knockoff+ thresholds under dependence, 2026. URL https://arxiv.org/abs/2607.17084. arXiv:2607.17084v2.

[182] 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.

[183] 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.

[184] Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values, 2024. URL https://arxiv.org/abs/2407.15733.

[185] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011. doi: 10.1214/11-STS356.

[186] 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.

[187] 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