**STAT 689: Large-scale and high-dimensional statistical inference** **Fall 2024**

Lecture 9

# 1 E-values

E-value is a recently proposed notion aiming to replace p-value in some statistical inference problems.

## 1.1 Definition

Consider a hypothesis $\mathcal{H}_0$ that includes a set of probability measures. Suppose we have observed the data $X$, where $X$ follows some distribution from $\mathcal{H}_0$ under the null.

**Definition (e-value).** A non-negative random $e$ is called an e-value for testing $\mathcal{H}_0$ if

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{E}_{P_0}[e(X)] \leq 1,$$

where $\mathbb{E}_{P_0}[e(X)]$ means the expectation of $e(X)$ for $X \sim P_0$.

**Definition (p-value).** A random variable p is called a p-value for testing $\mathcal{H}_0$ if

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{P}_{P_0}(p(X) \leq x) \leq x$$

for all $x$. In other words, $p(X)$ follows a super-uniform distribution for $X \sim P_0$ with any $P_0$ in $\mathcal{H}_0$.

We can construct a p-value through an e-value. Specifically, given an e-value $e$, we simply let $p = 1/e$.

**Exercise 9.1:** Show that $p = 1/e$ is a p-value according to the above definition.

## 1.2 Constructing e-values: Bayes factors

In Bayes hypothesis testing problem, we have

$$\mathcal{H}_0 = \{p_\theta | \theta \in \Theta_0\} \quad \text{versus} \quad \mathcal{H}_1 = \{p_\theta | \theta \in \Theta_1\}.$$

Let $G_i$ be a prior distribution on $\Theta_i$ for $i = 0, 1$. The Bayes factor collects the evidence in factor of $\mathcal{H}_1$ and is defined as

$$\text{BF} = \frac{p_{G_1}(X)}{p_{G_0}(X)},$$

where $p_{G_i}(X) = \int_{\Theta_i} p_\theta(X) dG_i(\theta)$ for $i = 0, 1$. We reject the null if this ratio is large enough. The Bayes factor is, in general, not an e-value. In some simpler cases, however, we can obtain e-values. Suppose we have a simple null hypothesis $\mathcal{H}_0 = \{p_0\}$ and $\mathcal{H}_1 = \{p_\theta | \theta \in \Theta_1\}$. The Bayes factor simplifies to

$$\text{BF} = e(X) = \frac{p_{G_1}(X)}{p_0(X)}.$$

Clearly, $e(X) \geq 0$ and

$$\mathbb{E}_{p_0}[e(X)] = \int e(x) p_0(x) dx = \int p_{G_1}(x) dx = \int \int_{\Theta_1} p_\theta(x) dG_1(\theta) dx = \int_{\Theta_1} \int p_\theta(x) dx dG_1(\theta) = 1.$$

Thus, Bayes factors can be used to obtain e-values.

## 1.3 Constructing e-values: reverse information projection

We now discuss another way to construct e-values. Recall that for two distributions $P, Q$ with the densities $p, q$, the Kullback–Leibler (KL) divergence is defined as

$$D(P\|Q) = \mathbb{E}_{X\sim P} \log \frac{p(X)}{q(X)}.$$

**Definition.** Given $G_1$, we define

$$G_0^* = \text{argmin}_{G_0} D(p_{G_1}\|p_{G_0}),$$

where the minimization is over all priors on $\Theta_0$. Then, $p_{G_0^*}$ is the reverse information projection of $p_{G_1}$ on the set $\{p_G : G \text{ is a prior on } \Theta_0\}$.

**Theorem (Grunwald et al., 2019, Theorem 1).** If $G_0^*$ exists, then $e(X) = p_{G_1}(X)/p_{G_0^*}(X)$ is an e-value, in the sense that

$$\mathbb{E}_0[p_{G_1}(X)/p_{G_0^*}(X)] \le 1.$$

Moreover, it achieves

$$\max \mathbb{E}_{X\sim p_{W_1}}[\log e(X)]$$

where the maximum is over all the e-values.

The quantity $\mathbb{E}_{X\sim p_{W_1}}[\log e(X)]$ can be viewed as an analog of power. But why should we look to maximize the logarithm of $e$ rather than the expectation of $e$ itself? One answer comes from an interpretation of e-values via betting; see the discussions in Section 1.4.

## 1.4 A betting perspective

We turn the hypothesis testing problem between two hypotheses, $H_0$ and $H_1$, into a betting game. At each turn, you may choose the amount to stake, and for each dollar bet, the payoff is equal to the e-value $e(x)$, dependent on the outcome $X$. Under the null, suppose that $\mathbb{E}_0[e(X)] = 1$. So you don't make money on average. Now, if the alternative is true, suppose that $\mathbb{E}_1[e(X)] > 1$. This situation corresponds to a game in which you think the alternative hypothesis is true, and under this regime, the payoff $e(X) - 1$ is positive. What is the long-term wealth you should expect to have at time $T$? In general, the wealth should be exponential in $T$. Taking the logarithm, we should expect the total wealth to be like $T \log e(x)$. In light of this, it is not unreasonable to maximize $\mathbb{E}_1[\log e(X)] = 1$. This strategy is known as Kelly Gambling or the Kelly Criterion.

## 1.5 Safe testing

**E-values based on $n$ observations.** In reality, we are given a set of independent samples $X_1, \ldots, X_n$. In this case, the e-value can be defined as

$$e(\mathbf{X}) = \prod_{i=1}^{n} e(X_i),$$

where $\mathbf{X} = (X_1, \ldots, X_n)$. Below for any $p_0 \in \mathcal{H}_0$, we write $\mathbb{P}_0 = \mathbb{P}_{p_0}$ and $\mathbb{E}_0 = \mathbb{E}_{p_0}$.

**Safe testing.** The safe testing procedure rejects the null when $e(\mathbf{X}) \ge 1/\alpha$. Under the null, the safe testing procedure controls the Type I error at level $\alpha$ as

$$\mathbb{P}_0(e(\mathbf{X}) \ge 1/\alpha) \le \alpha \mathbb{E}_0[e(\mathbf{X})] \le \alpha,$$

where the first inequality is due to Markov's inequality. In comparison, the Neyman-Pearson (NP) test rejects the null if $e(\mathbf{X}) \geq c_\alpha$ for $c_\alpha$ such that

$$\mathbb{P}_0(e(\mathbf{X}) \geq c_\alpha) = \alpha.$$

As a result, the safe testing procedure necessarily leads to power loss as compared to the NP test.

**Exercise 9.2:** Let $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_i \sim N(\mu, 1)$ independently. Consider the problem of testing

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu = \mu_1.$$

Compute $e(\mathbf{X})$. Find $c_\alpha$ in the NP test and compare it with the threshold $1/\alpha$ in safe testing when $\alpha = 0.05$.

## 1.6 Advantages of using e-values

Although safe testing incurs power loss, e-values offer some unique advantages as compared to p-values.

- They allow us to perform sequential inference and gradual appraisal of information and evidence.
- E-values concern expectations, which are robust to data dependence, whereas tail bounds (associated with the use of p-values) are not.
- They are easy to combine to aggregate results.

## 1.7 Safety under optional continuation

The main purpose of e-values is to address the issue of optional continuation, which involves deciding whether to collect new data and conduct further testing based on previous test outcomes. For instance, if research group A tests a new medication and gets a "promising but inconclusive" result, another group, B, might decide to conduct their own test with new data. Subsequently, group C might observe the results of group B and decide to collect data for further testing. To perform hypothesis testing in this scenario, it is necessary to combine results from multiple tests in a statistically valid manner. Using p-value-based methods is inadequate because the experiments are not independent - each subsequent group decides to collect data and conduct testing only after seeing the results of previous groups. Consequently, combining the data and recalculating the p-value as if all the data were fixed in advance leads to misleading results and can be considered p-hacking. E-values enable the use of safe tests, which are valid in the optional continuation setting. This allows researchers to monitor results and stop whenever they want while still ensuring statistically valid results and preserving Type I error guarantees.

Suppose we have data $(X_1, Z_1), (X_2, Z_2), \ldots$ coming in batches of size $n_1$, $n_2$, and so on. We can view $Z_i$ as side information, such as how much money we have to continue data collection. Define $N_t = \sum_{i=1}^{t} n_i$ as the amount of data collected after the $t$-th batch. A safe testing procedure works as follows. We first compute an e-value $e_1 = e(X_1, \ldots, X_{n_1})$. If the outcome is within a certain range (e.g., promising but not conclusive) and $(Z_1, \ldots, Z_{n_1})$ have certain values (the budget is enough to collect more data). Then we move to collect more data $(X_{n_1+1}, Z_{n_1+1}), \ldots, (X_{N_2}, Z_{N_2})$ and calculate the corresponding e-value $e_2 = e(X_{n_1+1}, \ldots, X_{N_2})$. Otherwise, we stop. Let $T$ be the number of data batches collected when we do stop. We report the final result as

$$E := \prod_{i=1}^{T} e_i.$$

Let $\mathcal{F}_t$ be a filtration generated by $(X_1, Z_1), (X_2, Z_2), \ldots, (X_{N_t}, Z_{N_t})$. Define $e_t$ as a conditional e-value if $e_t$ is non-negative, measurable with respect to $\mathcal{F}_t$ and satisfies that

$$\mathbb{E}_0[e_t | \mathcal{F}_{t-1}] \leq 1.$$

**Theorem.** With $e_1, e_2, \ldots$ defined above, the process

$$V_t = \prod_{i=1}^{t} e_i$$

is a non-negative supermartingale under the null.

*Proof.* Note that

$$\mathbb{E}_0[V_t|\mathcal{F}_{t-1}] = V_{t-1}\mathbb{E}_0[e_t|\mathcal{F}_{t-1}] \leq V_{t-1}.$$

Now, suppose $T$ is a stopping time. By the optional stopping theorem,

$$\mathbb{E}_0[V_T] \leq 1.$$

As a consequence, $V_T$ is an e-value, and thus, we can use it for testing.

**Ville's Inequality.** Under any $p_0 \in \mathcal{H}_0$, we have

$$\mathbb{P}_0\left(\sup_t V_t \geq 1/\alpha\right) \leq \alpha.$$

*Proof.* Define $\tau = \inf\{t : V_t \geq 1/\alpha\}$. Then, $\tau$ is a stopping time as $\{\tau \leq t\} \in \mathcal{F}_t$. By the optional stopping theorem,

$$\begin{aligned}
1 = \mathbb{E}_0[V_0] \geq& \mathbb{E}_0[V_\tau] \\
=& \mathbb{E}_0[V_\tau|\tau = \infty]\mathbb{P}_0(\tau = \infty) + \mathbb{E}_0[V_\tau|\tau < \infty]\mathbb{P}_0(\tau < \infty) \\
\geq& \mathbb{E}_0[V_\tau|\tau < \infty]\mathbb{P}_0(\tau < \infty).
\end{aligned}$$

As $V_\tau \geq 1/\alpha$, we have $\mathbb{E}_0[V_\tau|\tau < \infty] \geq 1/\alpha$. Therefore,

$$\mathbb{P}_0\left(\sup_t V_t \geq 1/\alpha\right) \leq \mathbb{P}_0(\tau < \infty) \leq \frac{1}{\mathbb{E}_0[V_\tau|\tau < \infty]} \leq \alpha.$$

In summary, combining e-values with arbitrary stop/continue strategy and rejecting the null when the final $V_T$ has $V_T \geq 1/\alpha$ is safe since

$$\mathbb{P}_0(V_T \geq 1/\alpha) \leq \mathbb{P}_0\left(\sup_t V_t \geq 1/\alpha\right) \leq \alpha,$$

which suggests that the Type-I error is at most $\alpha$.

# 2 E-values in multiple testing

Suppose we observe $n$ e-values $e_1, \ldots, e_n$ corresponding to the hypotheses $H_1, \ldots, H_n$. The $\alpha$-level e-BH procedure involves sorting the e-values in decreasing order as $e_{(1)} \geq \cdots \geq e_{(n)}$ and rejecting the hypotheses associated with the $\hat{k}$ largest e-values, where

$$\hat{k} := \max\left\{1 \leq i \leq n : e_{(i)} \geq n/(i\alpha)\right\}.$$

Notice that $P(1/e_i \leq t) \leq t$ by Markov's inequality, which indicates that $1/e_i$ is super-uniform. Thus, the e-BH procedure is simply the BH procedure applied to the p-values $\{1/e_i\}_{i=1}^n$. An advantage of the e-BH procedure is that it controls FDR at level $\alpha$ even under unknown arbitrary dependence among the e-values.

**Theorem.** The e-BH procedure has FDR at most $n_0\alpha/n$ regardless of the dependence among the e-values.

*Proof.* Note that

$$\begin{aligned}
\text{FDP} =& \sum_{i=1}^n \frac{\mathbf{1}\{ie_{(i)} \geq n/\alpha, H_{(i)} \text{ is under the null}\}}{1 \vee \hat{k}} \\
\leq& \sum_{i=1}^n \frac{\mathbf{1}\{ie_{(i)} \geq n/\alpha, H_{(i)} \text{ is under the null}\}}{1 \vee i} \\
\leq& \sum_{i=1}^n \mathbf{1}\{H_{(i)} \text{ is under the null}\}\frac{\alpha e_{(i)}}{n} = \frac{\alpha}{n}\sum_{i \in \mathcal{N}_0} e_i.
\end{aligned}$$

| method | $m(t)$ | $R_i(t)$ | method | $m(t)$ | $R_i(t)$ |
|---|---|---|---|---|---|
| BH | $nt$ | $\mathbf{1}\{p_i \le t\}$ | BC | $1+\sum_{i=1}^n \mathbf{1}\{p_i \ge 1-t\}$ | $\mathbf{1}\{p_i \le t\}$ |
| GBH | $ng(t)$ | $\mathbf{1}\{\varphi_i(p_i) \le t\}$ | GBC | $1+\sum_{i=1}^n \mathbf{1}\{\varphi_i(1-p_i) \le t\}$ | $\mathbf{1}\{\varphi_i(p_i) \le t\}$ |
| ST | $n\pi_0^\lambda t$ | $\mathbf{1}\{p_i \le t\}$ | | | |

Table 1: The selections of $m(t)$ and $R_i(t)$ for different methods.

As $\mathbb{E}_0[e_i] \le 1$, we have

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] \le n_0\alpha/n.$$

**Remark.** From the proof, we see that to control the FDR, we only require

$$\mathbb{E}\left[\sum_{i\in\mathcal{N}_0} e_i\right] \le n,$$

which is weaker than $\mathbb{E}_0[e_i] \le 1$ for all $i \in \mathcal{N}_0$.

**Exercise 9.3:** There is an interesting connection between the BH and e-BH procedures. Suppose we observe a p-value $p_i$ for hypothesis $H_i$. Recall the threshold in the BH procedure:

$$T = \sup\left\{0 < t \le 1 \colon \frac{nt}{1 \vee R(t)} \le \alpha\right\}.$$

Define the e-value associated with $H_i$ to be

$$e_i = \frac{1}{T}\mathbf{1}\{p_i \le T\}.$$

Let $\mathcal{S}_{\mathrm{BH}}$ be the set of rejections obtained through the BH procedure at the FDR level $\alpha$, and let $\mathcal{S}_{\mathrm{eBH}}$ represent the set of rejections obtained from the e-BH procedure at the same FDR level $\alpha$, with the e-values defined above. Show that $\mathcal{S}_{\mathrm{BH}} = \mathcal{S}_{\mathrm{eBH}}$.

More generally, suppose we reject the $i$th hypothesis if $R_i(T) = 1$ with

$$T = \sup\left\{t \in \mathcal{D} : \frac{m(t)}{1 \vee \sum_{j=1}^n R_j(t)} \le \alpha\right\}.$$

Here $\mathcal{D}$ denotes the domain of the threshold, $m(t)$ is an estimate of the number of false discoveries, and $\sum_{j=1}^n R_j(t)$ is the total number of rejections, with $R_j(t)$ being the indicator function that indicates whether the $j$th hypothesis should be rejected or not at the threshold $t$. The corresponding e-BH procedure is defined based on the $e$-values $e_i = nR_i(T)/m(T)$ for $1 \le i \le n$. The selections of $m(t)$ and $R_i(t)$ for different methods are summarized in Table 1.

## 2.1 Combining e-values

As we have seen, many procedures are equivalent to the e-BH procedure with a suitably defined set of e-values. Suppose we have $L$ such procedures. Each procedure (performed at level $\alpha$) is equivalent to the corresponding e-BH procedure applied to the set of e-values $\{e_i^l : i \in [n]\}_{l=1}^L$ for $l = 1, 2, \ldots, L$ and $[n] = \{1, 2, \ldots, n\}$. Here $e_i^l$ is the e-value associated with the hypothesis $H_i$ from the $l$th procedure. Suppose

$$\sum_{i\in\mathcal{N}_0} \mathbb{E}[e_i^l] \le n.$$

So, each of these procedures controls the FDR at level $\alpha$. Now let

$$e_i = \sum_{l=1}^L w_{l,i} e_i^l$$

be the weighted e-value, where $w_{l,i} \geq 0$ is the aggregating weight. If $\sum_{l=1}^{L} \max_i w_{l,i} \leq 1$, the weighted e-values satisfy

$$\sum_{i \in \mathcal{N}_0} \mathbb{E}[e_i] \leq n.$$

As a result, the e-BH procedure applied to $\{e_i\}$ controls the FDR at the desired level.

*Proof.*

$$\begin{aligned}
\sum_{i \in \mathcal{N}_0} \mathbb{E}[e_i] &= \sum_{i \in \mathcal{N}_0} \sum_{l=1}^{L} w_{l,i} \mathbb{E}[e_i^l] \\
&\leq \sum_{i \in \mathcal{N}_0} \sum_{l=1}^{L} \max_i w_{l,i} \mathbb{E}[e_i^l] \\
&= \sum_{l=1}^{L} \max_i w_{l,i} \sum_{i \in \mathcal{N}_0} \mathbb{E}[e_i^l] \\
&\leq \sum_{l=1}^{L} \max_i w_{l,i} n \\
&\leq n.
\end{aligned}$$

## 2.2 Assembling e-values

Suppose we have $L$ sets of e-values $\{e_i^l \colon i \in \mathcal{G}_l, |\mathcal{G}_l| = n_l\}$ from $L$ different datasets, where $\cup_l \mathcal{G}_l = [n]$, $\mathcal{G}_{l_1} \cap \mathcal{G}_{l_2} = \emptyset$ if $l_1 \neq l_2$, $e_i^l$ is associated with the hypothesis $H_i$ and

$$\sum_{i \in \mathcal{G}_l \cap \mathcal{N}_0} \mathbb{E}[e_i^l] \leq n_l.$$

Thus, the e-BH procedure applied to $\{e_i^l : i \in \mathcal{G}_l\}$ would control the FDR within the $l$th dataset.

Let $e_i = w_{l,i} e_i^l$ be the weighted e-value, where $w_{l,i} \geq 0$ is the assembling weight. If $\sum_{l=1}^{L} n_l \max_{i \in \mathcal{G}_l} w_{l,i} \leq n$, the weighted e-values satisfy

$$\sum_{i \in \mathcal{N}_0} \mathbb{E}[e_i] \leq n.$$

As a result, the e-BH procedure applied to $\{e_i : i \in [n]\}$ would control the overall FDR within the whole dataset (which combines the $L$ datasets together).

*Proof.*

$$\begin{aligned}
\sum_{i \in \mathcal{N}_0} \mathbb{E}[e_i] &= \sum_{l=1}^{L} \sum_{i \in \mathcal{G}_l \cap \mathcal{N}_0} w_{l,i} \mathbb{E}[e_i^l] \\
&\leq \sum_{l=1}^{L} \sum_{i \in \mathcal{G}_l \cap \mathcal{N}_0} \max_{i \in \mathcal{G}_l} w_{l,i} \mathbb{E}[e_i^l] \\
&= \sum_{l=1}^{L} \max_{i \in \mathcal{G}_l} w_{l,i} \sum_{i \in \mathcal{G}_l \cap \mathcal{N}_0} \mathbb{E}[e_i^l] \\
&\leq \sum_{l=1}^{L} \max_{i \in \mathcal{G}_l} w_{l,i} n_l \\
&\leq n.
\end{aligned}$$

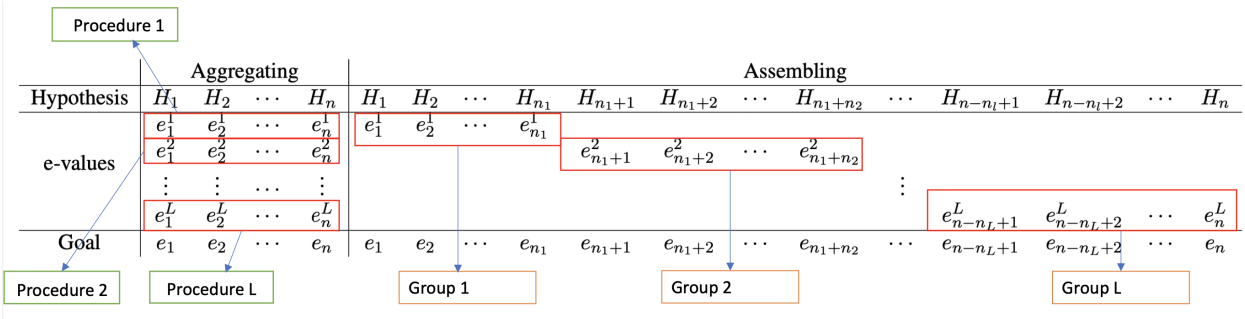


Figure 1: An illustration of aggregating e-values from different procedures and assembling e-values from different datasets.