Warning: These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès’s course from 2021.

Agenda:  E-values

1. E-values and corresponding tests
2. Bayes factors
3. Optional continuation problem
4. Example applications
5. FDR control with e-values (e-BH)

References:  Material for this lecture was based on [1] and [2].

15.1 Motivation

In this lecture, we study a replacement to the p-value called the e-value. The primary motivation for e-values is to address the optional continuation problem: deciding whether or not to collect new data and do further testing based on previous test outcomes. For example, suppose a research group A tests a new type of medication and obtains a "promising but inconclusive" result. Another research group B might see these results, and decide to conduct their own test with new data. Yet another group C might observe group B’s outcome, and collect data for further testing. To perform hypothesis testing in this setting, we would need to combine results from several tests in a statistically valid fashion. Attempting to use p-value based methods for this is unsatisfactory, because the experiments are not independent - each subsequent group decides to collect data and perform testing only after seeing the results of previous groups. Combining the data and re-calculating the p-value as if all the data were fixed in advance gives very wrong results [3], and can be considered p-hacking.

E-values give rise to safe tests: methods that are valid in the optional continuation setting. This allows researchers to monitor results and stop whenever they want, and still have statistically valid results, meaning that Type I error guarantees are preserved.

In this first section, we define e-values and show how to construct them using Bayes factors. Then, we discuss how to use them to construct safe tests.
15.2 E-values and corresponding tests

Suppose we have data \( X \) generated from probability distribution \( P \), and a hypothesis \( \mathcal{H}_0 \) (a set of probability measures).

**Definition 1.** A non-negative random variable \( E \) is called an **e-variable** for testing \( \mathcal{H}_0 \) if

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

Realized values of e-variables are called **e-values**. For simple hypotheses, e-variables are simply non-negative \( E \) with mean at most 1. To emphasize the difference between e-values and p-values, we can define p-variables as well.

**Definition 2.** A random variable \( P \) is called a **p-variable** for testing \( \mathcal{H}_0 \) if

\[
\sup_{P_0 \in \mathcal{H}_0} P_{P_0}(P(X) \leq \alpha) \leq \alpha \quad \text{for all} \quad \alpha \in (0, 1).
\]

Realized values of p-variables are p-values. From these definitions, we see a key difference between p-values and e-values: e-values control the expectation while p-values control the cdf.

We can relate e-values to p-values via the following transformation.

**Claim:** Let \( E \) be an e-value. Then \( E^{-1} \) is a conservative p-value. In other words, if \( P = E^{-1} \), then \( P(P \leq \alpha) \leq \alpha \).

**Proof:** Fix \( P_0 \in \mathcal{H}_0 \). By Markov’s inequality,

\[
P_{P_0}(1/E(X) \leq \alpha) = P_{P_0}(E \geq 1/\alpha) \leq \mathbb{E}_{P_0} \left[ \frac{E}{1/\alpha} \right] \leq \alpha.
\]

The p-value obtained from this transformation is conservative, because \( P(E \geq 1/\alpha) \) can be much smaller than \( \alpha \), as Markov’s inequality may not be tight. From this correspondence, we can use e-variables to test against \( \mathcal{H}_0 \) at level \( \alpha \), rejecting \( \mathcal{H}_0 \) if \( E(X) \geq \frac{1}{\alpha} \). For instance, the test that rejects \( \mathcal{H}_0 \) if and only if \( E(X) \geq 20 \), or if \( E^{-1}(X) \leq 0.05 \), has Type-I error bound 0.05. This is the **safe test** based on e-variable \( E \).

Next, we construct e-values using Bayes factor hypothesis testing.

15.2.1 Constructing e-values with Bayes factors

In Bayes factor hypothesis testing (Jeffreys ’39), we have two hypotheses

\[
\mathcal{H}_0 = \{ p_\theta \mid \theta \in \Theta_0 \} \text{ vs } \mathcal{H}_1 = \{ p_\theta \mid \theta \in \Theta_1 \}.
\]

Evidence in favor of \( \mathcal{H}_1 \) is measured by the **Bayes factor**

\[
\frac{p_{W_1}(X)}{p_{W_0}(X)}.
\]
where
\[
p_{W_1}(X) := \int_{\theta \in \Theta_1} p_\theta(X) dW_1(\theta) \\
p_{W_0}(X) := \int_{\theta \in \Theta_0} p_\theta(X) dW_0(\theta).
\]

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 \(H_0 = \{p_0\}\) and \(H_1 = \{p_\theta \mid \theta \in \Theta_1\}\). The Bayes factor simplifies to
\[
M(X) := \frac{p_{W_1}(X)}{p_0(X)}
\]
No matter what prior \(W_1\) we choose, we have
\[
\mathbb{E}_{X \sim p_0} [M(X)] = 1.
\]
This shows that for simple nulls, the Bayes factor is an e-value. In the even simpler case where both \(H_0\) and \(H_1\) are point hypotheses,
\[
E(X) = \frac{p_1(X)}{p_0(X)}
\]
is an e-value. Thus, Bayes factors can be used to obtain e-values for safe testing.

Note that safe testing is not Neyman-Pearson (NP) testing. The safe test rejects if \(E(X) \geq 1/\alpha\). Compared to the NP test, which rejects if \(E(X) \geq 1/B\), with \(B\) chosen such that \(P_{X \sim p_0} (E(X) \geq B) = \alpha\), the safe test is more conservative and typically results in a loss of power.

**Example 1.** Suppose we have \(X = (X_1, X_2, \ldots, X_n)\) with \(X_i\) iid \(N(\mu, 1)\). We consider simple hypotheses
\[
H_0 : \mu = 0, H_1 : \mu = \mu_1.
\]
The e-variable is
\[
E = \prod_{i=1}^{n} \exp \left( \mu_1 X_i - \frac{\mu_1^2}{2} \right)
\]
which corresponds to a rejection region for the safe test of
\[
\sum_{i=1}^{n} \mu_1 X_i - \frac{\mu_1^2}{2} > \log 20 \approx 3,
\]
much more conservative than the rejection region \(X \geq \frac{1.64}{\sqrt{n}}\) given by the Neyman-Pearson test.
Example 2 (Gaussian location family). Suppose we have $X = (X_1, X_2, \ldots, X_n)$ with $X_i$ iid $\mathcal{N}(\mu, 1)$ and the hypotheses

$$H_0 : \mu = 0, H_1 : \mu \in \Theta_1.$$ 

Assume a prior on the alternative $w(\mu) \propto \exp (-\mu^2/2)$. The Bayes factor is given by

$$E := \frac{p_W(X)}{p_0(X)} = \frac{\int_{\mu \in \mathbb{R}} p_\mu(X) w(\mu) d\mu}{p_0(X)}$$

This is an e-value. After some calculation we get

$$\log E = -\frac{1}{2} \log(n + 1) + \frac{1}{2} (n + 1) \hat{\mu}_n^2$$

where $\hat{\mu}_n = \frac{n}{n + 1} \bar{X}$ is the Bayes MAP estimator. The safe test thus rejects the null when $E \geq 20$, or when

$$|\hat{\mu}_n| \geq \sqrt{\frac{5.99 + \log(n + 1)}{n + 1}}$$

where we used $2 \log 20 \approx 5.99$. Again, this is more conservative and less powerful than NP, which rejects when $|\hat{\mu}_n| \geq \frac{1.96}{\sqrt{n}}$.

While safe testing is more conservative relative to NP, it offers a host of other advantages.

15.2.2 Advantages of e-values

There are various statistical advantages of e-values:

1. We know how to construct e-values for high-dimensional problems, whereas it can be hard to do the same with p-values (e.g. high-dimensional logistic regression)

2. They allow us to perform sequential inference and gradual appraisal of information and evidence

3. P-values, when small (e.g. on order of $10^{-10}$), rely heavily on the tail distribution of the model. E-values are more robust to model misspecification

4. E-values concern expectations, which are robust to data dependence, whereas tail bounds are not.

5. Non-asymptotic and often model-free

Moreover, the theory of martingales gives e-values validity for optional stopping times. In the next section, we will see that e-values are easy to combine and give us flexibility to stop/continue in data collection (online testing; unfixed sample size), allowing for safe tests for optional continuation.
15.3 Safety under optional continuation

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.

The safe test will run as follows. We first evaluate some e-value \(E_1\) on the first batch \((X_1, \ldots, X_{n_1})\). If the outcome is in a certain range (e.g. promising but not conclusive) and \(Z_{n_1}\) has certain values (e.g. 'boss has money to collect more data') then we move to evaluate some e-value \(E_2\) on the next batch \((X_{n_1+1}, \ldots, X_{N_2})\). Otherwise, we stop. Let \(T\) be the number of data batches collected when we do stop. We report as the final result

\[
E := \prod_{i=1}^T E_i
\]

Claim: \(E\) is itself an e-value, irrespective of the stop/continue rule used.

To formalize this, define filtration \(\mathcal{F}_t, t = 0, 1, 2, \ldots\) Define a conditional e-variable \(E_t\) as a non-negative RV which is \(\mathcal{F}_t\) measurable, such that for all \(P_0 \in \mathcal{H}_0\),

\[
\mathbb{E}_{P_0}[E_t | \mathcal{F}_{t-1}] \leq 1.
\]

**Proposition 1.** With \(E_1, E_2, \ldots\) as above, the process

\[
V_t = \prod_{i \leq t} E_i
\]

is a non-negative supermartingale (under the null).

Proof: Computing the conditional expectation of \(V_t\), we get

\[
\mathbb{E}[V_t | F_{t-1}] = \mathbb{E}[E_t V_{t-1} | F_{t-1}]
= V_{t-1} \mathbb{E}[E_t | F_{t-1}]
\leq V_{t-1}.
\]

Now suppose \(\tau\) is a stopping time. By Doob's optional stopping theorem,

\[
\mathbb{E}(V_\tau) \leq 1.
\]

In particular, \(V_\tau\) is an e-value, and thus we can use it for testing.

As a consequence of this, we have the following result.

**Claim (Ville's Inequality):** Under any \(P_0 \in \mathcal{H}_0\),

\[
\mathbb{P}_{P_0}\left( \sup_t V_t \geq 1/\alpha \right) \leq \alpha
\]

Proof: Define the stopping time \(\tau = \inf\{t | V_t \geq 1/\alpha\}\). By Doob's optional stopping theorem, \(P(\tau < \infty) \leq \alpha\).

In summary, under any stopping time \(\tau\), the end-product \(V_\tau\) of all employed e-values is itself an e-value even if \(E_i\) depends on the past. Thus, Type-I error is guaranteed to be preserved under optional continuation. Combining e-values with arbitrary stop/continue strategy and rejecting \(\mathcal{H}_0\) when final \(V_\tau\) has \(V_\tau \geq 20\) is safe, since Type-I error at most 0.05.
### 15.4 Examples in testing multiple hypotheses

**Detecting trading skills.** There are \( K \) traders who each manage a fund. For each fund \( k \), we observe the monthly returns \( X_{k,j}, j = 1, \ldots, n_k \). Null hypothesis \( k \) is that trader \( k \) is not skillful, i.e. that

\[
\mathbb{E} [X_{k,j} \mid \mathcal{F}_{j-1}] \leq 1
\]

for \( j = 1, \ldots, n_k \).

The problem is the test statistics (performance of funds) have complicated serial and cross dependence, making it hard to construct p-values and perform classical testing. However, we can easily construct e-values as

\[
E_k = \prod_{j=1}^{n_k} X_{k,j}.
\]

**Multi-armed bandit problems.** In this setting, there are \( K \) arms, with null hypothesis \( k \) being that arm \( k \) has mean reward at most 1. We employ strategy \( (k_t) \), which at time \( t \geq 1 \) pulls arm \( k_t \), obtaining an iid reward \( X_{k_t,t} \geq 0 \). The goal is to quickly detect arms with mean greater than 1 (or maximize profit, minimize regret, etc). The running reward for arm \( k \) at time \( j \) is

\[
M_{k,t} = \prod_{1 \leq j \leq t \atop k_j = k} X_{k,j}
\]

There is complicated dependence due to exploration/exploitation, but we can construct e-values \( M_{1,\tau}, \ldots, M_{K,\tau} \) for any stopping time \( \tau \).

### 15.5 Composite nulls and Bayes factors

Suppose we now have composite nulls. The Bayes factor is given by

\[
M(X) := \frac{p_{W_1}(X)}{p_{W_0}(X)}
\]

For this ratio to be an e-value, we require that for \( P_0 \in \mathcal{H}_0 \), \( \mathbb{E}_{X \sim P_0} [M(X)] \leq 1 \), i.e. that for every null the expectation is less than 1. But we can only guarantee that \( \mathbb{E}_{X \sim P_{W_0}} [M(X)] \leq 1 \). Bayes factors with composite nulls are therefore not e-values in general.

However, even if we have a composiite null, we can sometimes create e-values using Bayes factors. Suppose we are given a prior \( W_1 \) on \( \Theta_1 \). We solve the convex optimization problem

\[
W_0^* := \arg\min_{W_0 \text{ distr on } \Theta_0} D (P_{W_1} \| P_{W_0})
\]

where \( D \) is the Kullback-Leibler divergence

\[
D (P \| Q) := \mathbb{E}_{X \sim P} \left[ \log \frac{p(X)}{q(X)} \right]
\]
This finds the prior on $\Theta_0$ that is closest to $P_{W_1}$, provided it exists. $P_{W_0}$ is called the reverse information projection of $P_{W_1}$ on the set $\overline{H}_0 = \{P_W \mid W \text{ is distribution on } \Theta_0\}$.

We define the notion of GROW, a measure of the ”best” e-value. A GROW e-variable relative to $P_{W_1}$ is defined to achieve

$$\sup_E \mathbb{E}_{X \sim P_{W_1}}[\log E]$$

where the supremum is over all e-values relative to $H_0$. The result below says that $P_{W_0^*}$ gives us a GROW e-variable.

**Theorem 1** (Li ’99, Barron & Li ’00, Grünwald et al. ’19). If $W_0^*$ exists, then

$$\frac{p_{W_1}(X)}{p_{W_0^*}(X)}$$

is an e-variable. Moreover, it is the Bayes GROW e-variable relative to $W_1$, achieving

$$\max_{E\text{-var for } H_0} \mathbb{E}_{X \sim P_{W_1}}[\log E].$$

We want our e-variable to be large under the alternative, since we want to reject the null. This theorem says that $W_0^*$ gives us the best e-value in the sense that it maximizes $\log E$ under $W_1$, while being small under $H_0$, as $\mathbb{E}[E] \leq 1$. It is important to make $\log E$ large rather than $E$, because we want to avoid $E$ taking on 0 values, since we will multiply these $E$s together for safe testing.

### 15.6 FDR control with e-values

E-values can also be used for FDR control. Suppose we have $n$ hypothesis, realized e-values $e_1, \ldots, e_n$ associated with $H_1, \ldots, H_n$, and FDR level $\alpha \in (0, 1)$. We order them $e_{(1)} \geq \cdots \geq e_{(n)}$ The e-BH procedure rejects hypotheses with the largest $\hat{k}$ e-values, where

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

This procedure has parallels to BH: writing $1/e_{(i)} := p(i)$ gives us the thresholds $\frac{i}{n} \frac{1}{p(i)} \geq \frac{1}{\alpha}$, which rearranges to $p(i) \leq \frac{\alpha i}{n}$. It controls FDR without any assumption on dependence of e-values.

**Theorem 2** (Wang & Ramdas ’20 [4]). The e-BH procedure has FDR at most $n_0\alpha/n$. 

15-7
References

[1] Peter Grünwald. E is the New P: Tests that are safe under optional stopping, with an application to time-to-event data. International Seminar on Selective Inference, November 2020

