STATS 300C: Theory of Statistics **Spring 2022**

Lecture 14 — May 12, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: William Hartog*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and 2021 scribe notes written by Yujin Jeong and Cheuk To Tsui.*

Reading: Include Reading

Agenda: Selective Inference

1. Verifying the Winner
2. E-values and corresponding tests
3. Bayes factors
4. Optional continuation problem
5. Example applications

14.1 Verifying the Winner

We now discuss another selective inference problem, which has a different flavor. The content below is from Will Fithian's PhD dissertation (Stanford University, 2015.) In his dissertation, Will extended location family results of Gutman and Mymin (1987).

14.1.1 The Iowa Republican Poll (May, 2015)

The following table lists the results of a May 2015 Iowa Republican poll of 667 samples in which Scott Walker leads with 21% of the vote.

The question we want to ask is, "Is Scott Walker truly winning?" This question depends on the data we observed as we don't know who is in first until we see the data. Rather than having an issue with model selection, we have an issue with question selection.

To answer the question, we want to construct a test for the hypothesis that the winning candidate in a poll is truly in the lead. Further, we also want to be able to make a claim like "I am 95% confident that Scott Walker is leading by at least x%."

Rank	Candidate	Result	Votes
1	Scott Walker	21 %	140
2	Rand Paul	13 %	87
3	Marco Rubio	13 %	87
4	Ted Cruz	12 %	80
⋮	⋮	⋮	⋮
14	Bobby Jindal	1 %	7
15	Lindsey Graham	0 %	0

Table 14.1. Iowa Republican vote

14.1.2 Selective Hypothesis Testing

We model the number of votes that the candidates received as multinomial:

$$X = (X_1, \dots, X_{15}) \sim \text{Multinomial}(n, \pi).$$

We want to ask whether the winning candidate i in a poll is actually in the lead after seeing the data. Note that the question we are asking is data dependent and so is random. We test

$$\begin{aligned} H_i &= \pi_i \leq \max_{j \neq i} \pi_j \\ &= \cup_{j \neq i} H_{i \leq j} : \pi_i \leq \pi_j \end{aligned}$$

on the event $A_i = \{X_i > \max_{j \neq i} X_j\}$. We want to construct a selective level α test $\phi_i(X)$. Test $\phi_i(X)$ is a selective level α-test if

$$E[\phi_i(X)|A_i] \leq \alpha \quad \text{for any distribution in } H_i.$$

14.1.3 Construction of a selective test

We can construct a selective level α-test with the following procedure.

1. Construct a selective p-value $p_{i,j}$ for $H_{i \leq j}$ on A_i. For example with $i = 1, j = 2$, we will construct $p_{1,2}$ based on the conditional distribution

   $$X_1 | X_1 + X_2, X_{3:15}, A_1$$

   which follows truncated binomial distribution of $\text{Bin}(X_1 + X_2, \frac{\pi_1}{\pi_1 + \pi_2})$. Note that

   $$H_{1 \leq 2} : \pi_1 \leq \pi_2 \iff \pi_1 / (\pi_1 + \pi_2) \leq 1/2.$$

   Therefore, we are testing whether $X_1 \sim \text{Bin}(m, p)$ with $p \leq 1/2$ and $m = X_1 + X_2$ conditioned on $X_1 > m/2$. Then, $p_{1,2}$ is computed as

   $$\begin{aligned} p_{1,2} &= \mathbb{P}(\text{Bin}(m, 1/2) \geq X_1 | \text{Bin}(m, 1/2) > m/2) = \frac{\mathbb{P}(\text{Bin}(m, 1/2) \geq X_1)}{\mathbb{P}(\text{Bin}(m, 1/2) > m/2)} \\ &= 2\mathbb{P}(\text{Bin}(m, 1/2) \geq X_1). \end{aligned}$$

2. Combine p-value as
$$p_i = \max_{j\neq i} p_{i,j}.$$
In this way, we have a valid selective α test since
$$P(p_i \leq \alpha | A_i) = P(\cap_{j\neq i}\{p_{i,j} \leq \alpha\}|A_i) \leq \min_{j\neq i} P(p_{i,j} \leq \alpha | A_i) \leq \alpha$$
as we construct level-α tests for each $H_{i\leq j}$. In our 2015 Iowa Primary poll example, the optimal j is the runner up in the poll.

14.1.4 Is Scott Walker truly winning?

Let's apply this selective test to the Iowa Republican Poll data and consider the question of whether Scott Walker is really more likely to win than Rand Paul and by how much. It follows from our analysis that we only need to look at $p_{SW,RP}$ based on the distribution

$$(X_{SW} \mid X_{SW} + X_{RP} = 227, X_{others}, SW\text{wins}) \overset{d}{=} (X_{SW} \mid X_{SW} + X_{RP} = 227, X_{SW} \geq 114)\,.$$

Under $H_{SW} : \pi_{SW} \leq \max_{j\neq SW} \pi_j$, we obtain the p-value

$$p_{SW} = \max_{j\neq SW} p_{SW,j} = p_{SW,RP} = 2P(\text{Bin}(227, 1/2) \geq 140) = 0.00053.$$

As a result, we reject the null H_{SW} and conclude that Scott Walker is actually winning.

Furthermore, we want to make a claim like we are 95% confident that Scott Walker is leading by at least x%, which means $\pi_{SW} \geq (1 + x/100) \max_{j\neq SW} \pi_j$. Note that we have a test that is valid for each $\theta = \frac{\pi_{SW}}{\pi_{SW}+\pi_{RP}}$. Therefore, X rejects whenever it is in some region $R(\theta)$ such that $P_\theta(X \in R(\theta)) \leq \alpha$. Then, if we define $C(X)$ as the set of all θ for which X is not in $R(\theta)$,
$$P_\theta(\theta \in C(X)) = P_\theta(X \notin R(\theta)) \geq 1 - \alpha.$$
Essentially, our 95% confidence intervals in this case are of the form of

$$\theta \geq \theta_0 \iff \pi_{SW} \geq (1 + x/100)\pi_{RP}$$

for some θ_0 and x. We get θ_0 by solving

$$P_{\theta_0}[X_{SW} \geq 140 | X_{SW} + X_{RP} = 227, X_{SW} > 114] = 0.05.$$

With further calculations, we are 95% confident that Scott Walker is leading by at least 22%.

14.2 E-values 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.

14.3 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

$$\sup_{P_0 \in \mathcal{H}_0} \mathbb{P}_{P_0} \left(P(X) \leq \alpha\right) \leq \alpha \text{ for all } \alpha \in (0,1).$$

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 $\mathbb{P}(P \leq \alpha) \leq \alpha$.
Proof: Fix $P_0 \in \mathcal{H}_0$. By Markov's inequality,

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

The p-value obtained from this transformation is conservative, because $P(E \geq 1/\alpha)$ can be much smaller than α, as Markov's inequality may not be tight. From this correspondence, we can use e-variables to test against $\mathcal{H}_0$ at level α, 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.

14.3.1 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 $\mathbb{P}_{X \sim p_0}(E(X) \geq B) = \alpha$, 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 $\mathcal{N}(\mu, 1)$. 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 $\bar{X} \geq \frac{1.64}{\sqrt{n}}$ 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 $w(\mu) \propto \exp\left(-\mu^2/2\right)$. 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

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

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

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

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

14.3.2 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.

14.4 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

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

<u>Claim:</u> 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 $P_0 \in \mathcal{H}_0$,

$$\mathbb{E}_{P_0}\left[E_t \mid \mathcal{F}_{t-1}\right] \leq 1.$$

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

$$\begin{aligned}\mathbb{E}\left[V_t \mid F_{t-1}\right] &= \mathbb{E}\left[E_t V_{t-1} \mid F_{t-1}\right] \\ &= V_{t-1}\mathbb{E}\left[E_t \mid F_{t-1}\right] \\ &\leq V_{t-1}.\end{aligned}$$

Now suppose τ is a stopping time. By Doob's optional stopping theorem,

$$\mathbb{E}\left(V_\tau\right) \leq 1.$$

In particular, V_τ 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 $\tau = \inf\{t \mid V_t \geq 1/\alpha\}$. By Doob's optional stopping theorem, $P(\tau < \infty) \leq \alpha$.

In summary, under any stopping time τ, the end-product V_τ 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_τ has $V_\tau \geq 20$ is safe, since Type-I error at most 0.05.

14.5 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

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

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

$$M_{k,t} = \prod_{\substack{1 \leq j \leq t \\ k_j = k}} X_{k,j}$$

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 τ.

Bibliography

[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

[2] Ruodu Wang. Game-theoretic statistical inference E-values vs p-values: calibration, combination, and closed testing. Game-theoretic statistical inference, CMU February 2021

[3] Peter Grünwald, Rianne de Heide, Wouter Koolen. Safe testing. arXiv preprint, 2019.

[4] William Fithian. "Topics in adaptive inference." PhD thesis, Stanford University, 2015.