E is for Evidence

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

joint work with Rianne de Heide, Wouter Koolen, Alexander Ly, Muriel Pérez, Judith ter Schure, Rosanne Turner

E

is for Evidence

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

with Rianne de Heide,
Wouter Koolen,
Judith ter Schure,
Alexander Ly,
Rosanne Turner,
Muriel Perez






Replication Crisis in Science

somehow related to use of p-values and significance testing…

Replication Crisis in Science

somehow related to use of p-values and significance testing…

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

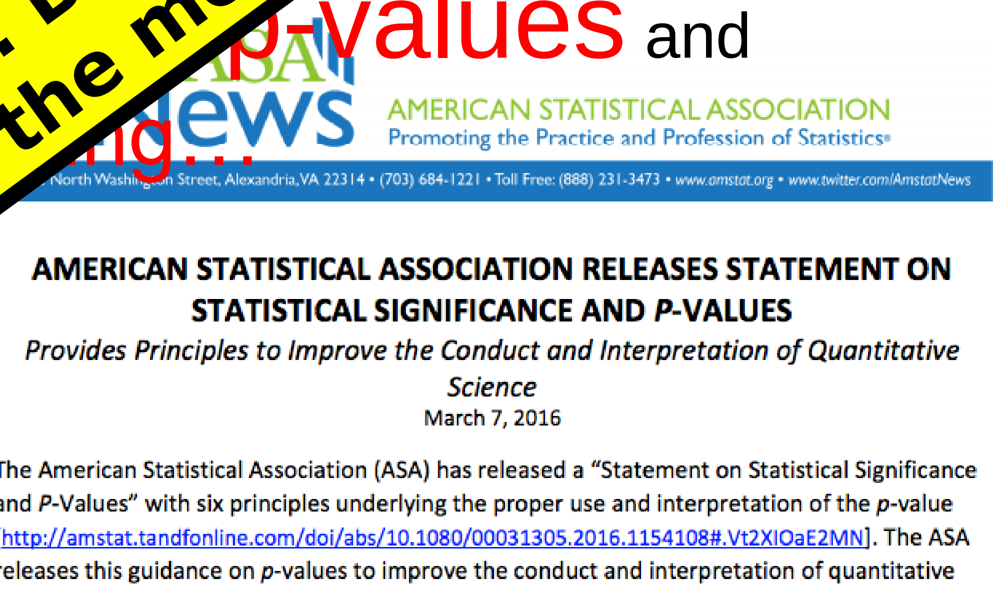
The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Replication … s in Science

somehow … p-values and

signi… …ing…

Redefine Statistical Significance (to $p < 0.005$): Benjamin et al. 2017, incl. some of the most famous statisticians



ASA News

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

… North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Abandon Significance: McShane et al. (including some of the most famous statisticians) 2019

Redefine Sta... Significance
(to $p < 0.005$) ...min et al. 2017,
incl. some of th... famous statisticians

...lication... s in Sc...

somehow ... and

sign... ...ing!

...IATION

AMERICAN STATISTICAL ASS...
STATISTICAL SIGNIFICA...

Provides Principles to Improve the Conduct and ...
Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical ... and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-va... [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Rise Up **Against Significance: 800 signatories (including some of the most famous statisticians) 2019**

Abandon Sign... (including ... statistici... McShane et al. ...e most famous

...ificance ...2017, ...ticians

Rede... (to p < 0.0... incl. some of th...

...lication...

...and

AMERICAN STATISTICAL ASS...

STATISTICAL SIGNIFICA...

Provides Principles to Improve the Conduct and ...

Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical ... and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-va... [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Menu

1. Null Hypothesis Testing, p-value
2. Two Problems with p-values
3. E is the new P
    - Conservative p-value interpretation
    - Likelihood ratio interpretation
    - How it solves the p-value problems
4. E-values and Bayes Factors
    - Main result of G., De Heide, Koolen '20 *SafeTesting*
5. E-Values and Evidence

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, today we assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{ \frac{1}{2} \right\}, \Theta_1 = [0,1] \setminus \left\{ \frac{1}{2} \right\}$

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}, \Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$ **Simple H_0**

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$$

  $$H_1 = \{ P_{\sigma,\mu} \big| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$$

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  Composite H_0

  $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$$
  $$H_1 = \{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \}$$

Standard Method: p-value, significance

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
- A ("nonstrict/conservative") **p**-value is a random **variable** (!) such that, for all $\theta \in \Theta_0$,

$$P_{\theta_0}(\mathbf{p} \leq \alpha) \leq \alpha$$

- ...with continuous-valued data we typically use strict p-values, i.e.

$$P_{\theta_0}(\mathbf{p} \leq \alpha) = \alpha$$

Standard Methodology of Neyman-Pearson testing

1. We fix H_0 (and H_1) and significance level α (e.g. 0.05)
2. We set a sample plan
   - e.g. $n = 100$, or 'stop as soon as you have seen three 1s in a row'
3. This determines random variable $Y = \mathrm{X}^{\tau} = (\mathrm{X}_1, \ldots, X_{\tau})$
   - e.g. $\tau = n = 100$ or $\tau = \min\{n: X_{n-2} = X_{n-1} = X_n = 1\}$
4. We define a *p*-value on Y
5. We observe Y. If $p < \alpha$: **reject $\boldsymbol{H_0}$** , otherwise accept

Motivation behind Neyman-Pearson Test

- The **Type-I error** is the probability that we reject the null hypothesis even though it is true.
  - False alarm; medication seems to work even though it doesn't
- By the definition of p-value, for all $P \in H_0$,

$$P(\text{reject}) = P(p < \alpha) \leq \alpha$$

- Hence Type-I error is bounded by significance level α

Long-Run Rationale

- We determine (before experiment!) a significance level α and we 'reject' the null hypothesis iff $p < \alpha$
- This gives a **Type-I Error Probability bound α**
- **If we follow this decision rule consistently throughout our lives and set e.g. $\alpha = 0.05$, then in long run we reject nulls while they are correct at most 5% of the time**

Neyman's **Inductive Behaviour** Philosophy

Long-Run Rationale

- We determine (before experiment!) a significance level α and we 'reject' the null hypothesis iff $p < \alpha$
- This gives a **Type-I Error Probability bound α**
- **If we follow this decision rule consistently throughout our lives and set e.g. $\alpha = 0.05$, then in long run we reject nulls while they are correct at most 5% of the time**
- **Strict** Neyman-Pearson: do not mention p-value itself only decide reject or accept!

Standard Methodology of Neyman-Pearson testing

1. We fix H_0 (and H_1) and significance level α (e.g. 0.05)
2. We set a sample plan
   - e.g. $n = 100$, or 'stop as soon as you have seen three 1s in a row'
3. This determines a random variable $Y = \mathrm{X}^{\tau}$
   - e.g. $\tau = n = 100$ or $\tau = \min\{n: X_{n-2} = X_{n-1} = X_n = 1\}$
4. We define a *p*-value on Y
5. We observe data. If $p < \alpha$ we **reject $\boldsymbol{H_0}$** , otherwise we accept

Standard Methodology of ~~Neyman-Pearson~~ testing in practice

1. We fix H_0 (and H_1) and significance level α (e.g. 0.05)
2. We set a sample plan
   - e.g. $n = 100$, or 'stop as soon as you have seen three 1s in a row'
3. This determines a random variable $Y = \mathrm{X}^{\tau}$
   - e.g. $\tau = n = 100$ or $\tau = \min\{n: X_{n-2} = X_{n-1} = X_n = 1\}$
4. We define a *p*-value on Y
5. Observe data. If $p < \alpha$ **reject $\boldsymbol{H_0}$** , otherwise accept
6. Also report p-value as indication of **strength of evidence against $\boldsymbol{H_0}$**

Two Problems with p-values

1. Type-I error guarantee not preserved under **optional continuation** – something we do all the time in modern practice!
   - note: I do think the Type-I error guarantee is highly desirable! The problem is that it does not hold

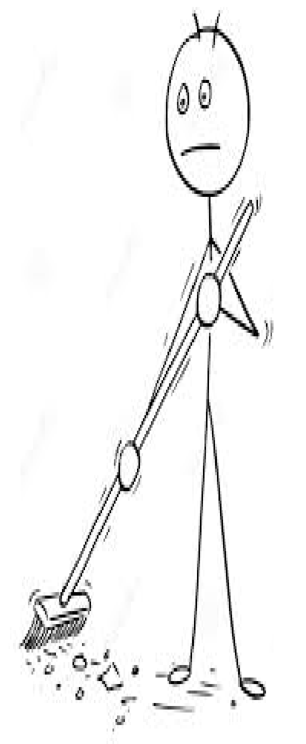
2. Evidential Meaning is compromised by p-values dependence on counterfactual decisions

First Problem with P-values

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?

First Problem with P-values

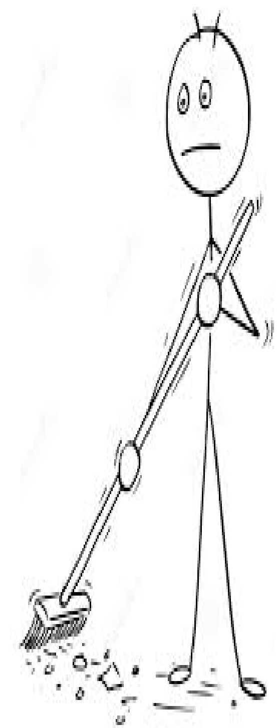
- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method, more often than not:**
  sweep data together and re-calculate p-value
- **Is this p-hacking?** **YES**



First Problem with P-values

- Suppose reseach group A tests medication, gets 'promising but not conclusive' result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method:**

  sweep data together and re-calculate p-value
- **Is this p-hacking? YES**
- **Does meta-analysis have the tools to do this much better? NO**



What can go wrong if you re-calculate p-values like this?

1. Do first test; observe $Y_{(1)} = (\mathrm{X}_1, \ldots, X_{100})$
2. If significant ($p_{Y_{(1)}} < 0.05$) reject and stop
   else do 2nd test on 2nd batch $Y_{(2)} = (X_{101}, \ldots, X_{200})$
3. If significant ($p_{(Y_{(1)}, Y_{(2)})} < 0.05$) reject else accept

$p_{Y_{(1)}, Y_{(2)}}$ is a p-value defined on X^{200} which is the wrong sample space. In X^{200} each outcome is vector of 200 X_i's

We should instead calculate a p-value on a sample space in which some outcomes have length 100 and other 200

What can go wrong?

1. Do first test; observe $Y_{(1)}$
2. If significant ($p_{Y_{(1)}} < 0.05$) , reject and stop
   else…
3. …Do second test on second batch $Y_{(2)}$
4. If significant ($p_{(Y_{(1)},Y_{(2)})} < 0.05$) , reject and stop
   else…
5. …Do third test on third batch $Y_{(3)}$…

…if you keep doing this long enough, the Type-I error probability goes to 1 instead of 0.05 !

Second problem: p-values rely on counterfactuals

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). I want to write a nice paper about this...But just to make sure I ask a statistician whether I did everything right.

p-values depend on counterfactuals

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n = 100$). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been ‘almost-but-not-quite’ significant?

p-values depend on counterfactuals

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been ‘almost-but-not-quite’ significant?
- I say “Well I never thought about that. Well, perhaps, but I’m not sure, I would have asked my boss for money to test another 50 patients”

p-values depend on counterfactuals

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result (*p =0.03* based on fixed *n=100*). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been 'almost-but-not-quite' significant?
- I say "Well I never thought about that. Well, perhaps, but I'm not sure, I would have asked my boss for money to test another 50 patients".
- *Now the statistician says: that means your result is invalid!*

p-values depend on counterfactuals

- Wheter or not a test based on p-values is valid depends on what you **would have done in situations that did not occur!**
- This is weird, both philosophically but also practically. In many testing situations it is **simply impossible to know** in advance what *would* have happened if the data had been different
- It also shows that it's really problematic to think of p-values as measuring evidence against the null!

Menu

1. Null Hypothesis Testing, p-value
2. Two Problems with p-values
3. **E is the new P**
    - Conservative p-value interpretation
    - Likelihood ratio interpretation
    - How it solves the p-value problems
4. E-values and Bayes Factors
    - Main result of G., De Heide, Koolen ‘20 *SafeTesting*
5. E-Values and Evidence

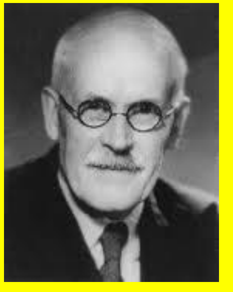
E is the new P

- We propose a generic replacement of the p-value that we call the e-value
- e-values handle **optional continuation** (to the next test (and the next, and ..)) without any problems

(simply multiply e-values of individual tests, despite dependencies)

E is the new P

E-variables have Fisherian, Neymanian and Bayes-Jeffreys' aspects to them, all at the same time

Cf. J. Berger (2003, IMS Medaillion Lecture): *Could Neyman, Fisher and Jeffreys have agreed on testing?*

individual tests, despite dependencies)

e-variables/e-values: General Definition

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis
- An **e-variable** for sample size n is a function $S : \mathcal{X}^n \to \mathbb{R}_0^+$ such that for **all** $P_0 \in H_0$, we have

$$\mathbf{E}_{P_0}[S(X^n)] \leq 1$$

First Interpretation: p-values

- Proposition: Let S be an e-variable. Then $S^{-1}(X^n)$ is a conservative p-value, i.e. p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- Proof: just Markov's inequality!

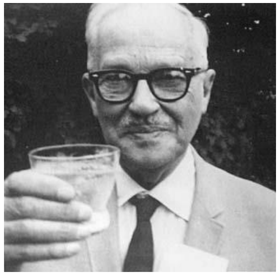
$$P\left(\, S(X^n) \geq \alpha^{-1}\right) \leq \frac{\mathbf{E}[S(X^n)]}{\alpha^{-1}} = \alpha$$

“Safe” Tests

- The test against H_0 at level α based on e-variable S is defined as the test which rejects H_0 if $S(X^n) \geq \frac{1}{\alpha}$

- Since S^{-1} is a conservative p-valuue...

- ....the test which rejects H_0 iff $S(X^n) \geq 20$, i.e. $S^{-1}(X^n) \leq 0.05$, has **Type-I Error** Bound of 0.05



Second Interpretation: Likelihoods (when H_0 and H_1 are simple)

Consider $H_0 = \{ p_0 \}$ and $H_1 = \{p_1\}$ .Then likelihood ratio given by

$$S(X^n) := \frac{p_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

But then S is also an E-variable!

$$\mathbf{E}_{X^n \sim P_0}[S(X^n)] =$$

$$\int p_0(x^n) \cdot \frac{p_1(X^n)}{p_0(x^n)} dx^n = \int p_1(x^n) dx^n = 1$$

The Main Theorem of Safe Testing (G., De Heide, Koolen, '20)

- Let H_0 and H_1 be (essentially) arbitrary.
  - In particular, they can both be composite
- …and let Y represent the data from our experiment.
- A **non-trivial** E-variable for H_0 that tends to take on large values if H_1 is true **always** exists!

  - such E-variables often take on the form of Bayes factors; however, not all Bayes factors are E-variables, and there are very useful E-variables that are not Bayes factors

Menu

1. Null Hypothesis Testing, p-value
2. Two Problems with p-values
3. E is the new P
   - Conservative p-value interpretation
   - Likelihood ratio interpretation
   - How it solves the p-value problems
4. E-values and Bayes Factors
   - Main result of G., De Heide, Koolen ‘20 *SafeTesting*
5. E-Values and Evidence

e-value based tests are safe under optional continuation

- Suppose we observe data $(X_1, Z_1), (X_2, Z_2), \ldots$
  - Z_i: side information

  ...coming in batches of size $n_1, n_2, \ldots, n_k$. Let $N_j := \sum_{i=1}^{j} n_i$
- We first evaluate some e-value S_1 on $(X_1, \ldots, X_{n_1})$.
- If outcome is in 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 evaluate some e-value S_2 on $\left(X_{n_1+1}, \ldots, X_{N_2}\right)$,
  otherwise we **stop.**

Safe Tests are Safe

- We first evaluate S_1.
- If outcome is in certain range and Z_{n_1} has certain values then we evaluate S_2 ; otherwise we **stop.**
- If outcome of S_2 is in certain range and Z_{N_2} has certain values then we compute S_3 , else we **stop**.
- ...and so on
- ...when we finally stop, after say K data batches, we report as final result the product $S := \prod_{j=1}^{K} S_j$
- **First Result, Informally: any S composed of e-values in this manner is itself an e-value, irrespective of the stop/continue rule used!**

Formalizing First Result

- Let $\left(Y_{(i)}\right)_{i\in\mathbb{N}}$ represent some random process.
- A **conditional** e-variable $S_{(i)}$ for $Y_{(i)}$ given $Y^{(i-1)} = (Y_1, \dots, Y_{(i-1)})$ is a nonnegative RV that is determined by $Y^{(i)}$ (i.e. it can be written as a fn $S_{(i)} = f\left(Y^{(i)}\right)$) and that satisfies, for all $P_0 \in H_0$:

$$\mathbf{E}_{P_0}\left[S_{(i)} \mid Y_{(1)}, \ldots, Y_{(i\text{-}1)}\right] \leq 1$$

Formalizing First Result

- **Conditional** e-variable:

$$\mathbf{E}_{P_0}\left[S_{(i)} \mid Y_{(1)}, \ldots, Y_{(i-1)}\right] \leq 1$$

- **Proposition:** Let $S_{(1)}, S_{(2)}, \ldots$ be e-variables for $Y_{(i)}$ conditional on $Y^{(i-1)}$. Then the process $\left(S^{(i)}\right)_{i\in\mathbb{N}}$ with $S^{(n)} = \prod_{i=1..n} S_{(i)}$ is a nonnegative supermartingale

- **Consequence: Ville's Inequality:**

$$P_0(\exists i: \; S^{(i)} \geq 1/\alpha) \leq \alpha.$$

“Safe” Tests are Safe

Pre-Ville’s Inequality:

Under any stopping time τ, the end-product of all employed e-values $\prod_{i=1..\tau} S_{(i)}$is **itself an e-value** even if defn of $S_{(i)}$ depends on past (then $S_{(i)}$ is conditional e-value)

Corollary: Type-I Error Guarantee Preserved under Optional Continuation

Suppose we combine e-values with arbitrary stop/continue strategy and reject H_0 when final $S^{(\tau)}$ has $1/S^{(\tau)} \leq 0.05$. Then resulting test is “safe for optional continuation”: Type-I Error ≤ 0.05

Safe Tests are Safe

Pre-Ville:

Under any stopping time τ, the end-pr… of all employed e-values $\prod_{i=1..\tau}$ … **…self an e-value** even if defn of S… …pends on past (then $S_{(i)}$ is cond… …e-value)

Corollary: Type-I E… …uarantee Preserved under Optio… …tinuation

Suppos… …ombine e-values with arbitrary sto… …inue strategy and reject H_0 when final $S^{(\tau)}$ has $1/S^{(\tau)} \leq 0.05$. Then resulting test is "safe for optional continuation": Type-I Error ≤ 0.05

e-values solve a central problem of p-values!

E-Values do not rely on counterfactual OC decisions

- Let $Y_{(1)}$ be a random variable representing my first batch of data.
- I quantify the evidence against H_0 in $Y_{(1)}$ by an E-variable $S_{(1)} = s(Y_{(1)})$. Say it is 10

E-Values do not rely on counterfactual OC decisions

- Let $Y_{(1)}$ be a random variable representing my first batch of data.
- I quantify the evidence against H_0 in $Y_{(1)}$ by an E-variable $S_{(1)} = s(Y_{(1)})$. Say it is 10
- Now my boss tells me: ah if it would have been ≥ 18 I would have given you some money to organize a second study, and you could have calculated $S_{(2)} = s(Y_{(2)})$ and report $S^{(2)} = s\left(Y_{(1)}\right) \cdot s(Y_{(2)})$
- Does this mean your E-value is not valid any more?

E-Values do not rely on counterfactual OC decisions

- Let $Y_{(1)}$ be a random variable representing my first batch of data.
- I quantify the evidence against H_0 in $Y_{(1)}$ by an E-variable $S_{(1)} = s(Y_{(1)})$. Say it is 10
- Now my boss tells me: ah if it would have been ≥ 18 I would have given you some money to organize a second study, and you could have calculated $S_{(2)} = s(Y_{(2)})$ and report $S^{(2)} = s\left(Y_{(1)}\right) \cdot s(Y_{(2)})$
- Does this mean your E-value is not valid any more?
- No! …because $S^* = S_{(1)}$ if $S_{(1)} < 18$ and $S_{(1)} \cdot S_{(2)}$ otherwise is still an E-value!

Menu

1. Null Hypothesis Testing, p-value
2. Two Problems with p-values
3. E is the new P
   - Conservative p-value interpretation
   - Likelihood ratio interpretation
   - How it solves the p-value problems
4. E-values and Bayes Factors
   - Main result of G., De Heide, Koolen '20 *SafeTesting*
5. E-Values and Evidence

E-Values, Likelihood Ratios, Bayes

- **Bayes factor hypothesis testing (Jeffreys '39)** with $H_0 = \{ p_\theta | \theta \in \Theta_0 \}$ vs $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:
  Evidence in favour of H_1 measured by

$$\frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$$

where

$$p_{W_1}(X_1, \ldots, X_n) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) dW_1(\theta)$$

$$p_{W_0}(X_1, \ldots, X_n) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) dW_0(\theta)$$

E-values, LRs, Bayes, simple $\boldsymbol{H_0}$

Bayes factor hypothesis testing

between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

Bayes factor of form

$$M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

Note that (no matter what prior W_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] =$$

$$\int p_0(x^n) \cdot \frac{p_{W_1}(X^n)}{p_0(x^n)} dx^n = \int p_{W_1}(x^n) dx^n = 1$$

E-values, LRs, Bayes, simple H_0

Bayes factor hypothesis testing

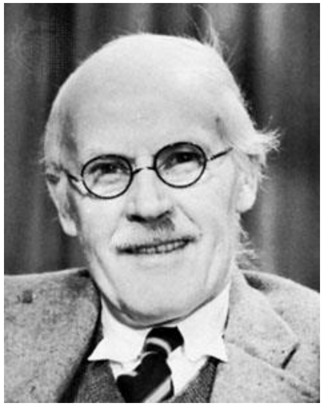
between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

Bayes factor of form

$$M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

Note that (no matter what prior W_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] = 1$$



The Bayes Factor for Simple H_0

is an e-value!

Composite H_0: Bayes may not be Safe!

Bayes factor given by $M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$

E-value requires that **for all** $P_0 \in H_0$:

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] \leq 1$$

...but for a Bayes factor we can only guarantee that

$$\mathbf{E}_{X^n \sim P_{W_0}}[M(X^n)] \leq 1$$

Composite H_0: Bayesian testing can be unsafe!

- ...for Bayes factor we can in general only guarantee

$$\mathbf{E}_{X^n \sim P_{W_0}} [M(X^n)] \leq 1$$

- In general Bayesian tests with composite H_0 are not safe ...which means that they loose their Type-I error guarantee interpretation when we combine Bayes factors on different studies
- Bayesian tests with composite H_0 ***are*** safe if you really believe your prior on H_0
- I usually don't believe my prior, so no good for me!

Composite H_0: Bayes may not be Safe!

Bayes factor given by $M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$

- In general Bayes factors with composite H_0 are not E-values
- ...but there do exist *very special priors* W_1^* **,** W_2^* **(sometimes highly unlike priors that "Bayesian" statisticians tend to use!)** for which Bayes factors become E-values and even very good E-values
- Main Theorem of G., De Heide, Koolen '20, *safe testing* shows how to construct such priors

E-Values vs Bayes, Part II: nonparametric H_0

- There is another issue with Bayesian testing:

- At least when n is small, not clear how to do a Bayesian test of a nonparametric null…

E-Values vs Bayes, Part II: nonparametric $\boldsymbol{H_0}$ - Example

We observe independent data $(X_{1a}, X_{1b}), (X_{2a}, X_{2b}), \ldots$

- H_0: for all i, distribution of X_{1i} and X_{2i} is the same
- H_1: (e.g.) for at least some i, they are different!

We make **no further assumptions on $\boldsymbol{H_0}$**: could be Gaussian, Bernoulli, heavy-tailed, …. So: H_0 is huge!

A classic p-value based test for this is Wilcoxon's (1945!) signed-rank test – used 10000s of times

E-Values vs Bayes, Part II: nonparametric $\boldsymbol{H_0}$ - Example

We observe independent data $(X_{1a}, X_{1b}), (X_{2a}, X_{2b}), \ldots$

- H_0: for all i, distribution of X_{1i} and X_{2i} is the same
- H_1: (e.g.) for at least some i, they are different!

A classic p-value based test for this is Wilcoxon's (1945!) signed-rank test – used 10000s of times

As a Bayesian you either have to make parametric assumptions or use a prior on a nonparametric set – which (a) still will not cover all of H_0 - and (b) which may need a large sample before it starts to work

E-Values vs Bayes, Part II: nonparametric $\boldsymbol{H_0}$ - Example

We observe independent data $(X_{1a}, X_{1b}), (X_{2a}, X_{2b}), \ldots$

- H_0: for all i, distribution of X_{1i} and X_{2i} is the same
- H_1: (e.g.) for at least some i, they are different!

For E-variable methodology, this setting is perfectly fine. Use for example the **Efron-De la Pena** E-Variable:

$$S_\lambda \ := \ \exp\left(\lambda \sum_{i=1..n} Z_i \ - \left(\frac{\lambda^2}{2} \sum_{i=1..n} Z_i^2 \right) \right)$$

where $Z_i = X_{ia} - X_{ib}$. Or better: $S_w \ := \int S_\lambda d\lambda$

We use a prior but we are still not Bayesian (at least not in the classic sense!)

Menu

1. Null Hypothesis Testing, p-value
2. Two Problems with p-values
3. E is the new P
   - Conservative p-value interpretation
   - Likelihood ratio interpretation
   - How it solves the p-value problems
4. E-values and Bayes Factors
   - Main result of G., De Heide, Koolen ‘20 *SafeTesting*
5. E-Values and Evidence

E-Values as Evidence

- p-values are still often used as evidence in data (small p-value means large evidence against the null)
  - Bayesians and likelihoodists have severely attacked this interpretation (see e.g. *Statistical Evidence: a Likelihood Paradigm* by R. Royall)

E-Values as Evidence

- p-values are still often used as evidence in data (small p-value means large evidence against the null)
  - Bayesians and likelihoodists have severely attacked this interpretation (see e.g. *Statistical Evidence: a Likelihood Paradigm* by R. Royall)
- Likelihood ratios are now the standard way to represent evidence in **Courts of Law** worldwide, e.g.

  H_0: incomplete DNA sample from defendant
  H_1: DNA sample not from defendant

  - p-values have been more or less banned in court

E-Values as Evidence

- p-values are still often used as evidence in data (small p-value means large evidence against the null)
  - tenuous!
- Likelihood ratios are now the standard way to represent evidence in **Courts of Law** worldwide, e.g.

  H_0: incomplete DNA sample from defendant
  H_1: DNA sample not from defendant

Idea: completely separate decision ('testing' in stats, 'verdict' in court – supplied by judge) from pieces of evidence (supplied by domain expert)

E-Values as Evidence

- p-values are still often used as evidence in data (small p-value means large evidence against the null)
  - Interpretation very tenuous
- likelihood ratios: uncontroversial when H_0 and H_1 are simple…

  ….and then E-values, likelihoods and Bayes factors coincide

…so can we view E-values or Bayes factors or neither as a proper generalization of evidence for composite H_0 and H_1 ?

Bayes vs E

- Likelihoodist and Bayesian evidence against H_0 always evidence for H_1
  - problems** if H_0 or H_1 composite/nonparametric
- E-value can quantify evidence against H_0 without this being evidence for a specific H_1
  - Like the p-value, but avoids problems such as OC and counterfactual dependence
  - fine for composite H_0 and H_1.

Bayes vs E: Luckiness Principle

- E-value can quantify evidence against H_0 without this being evidence for a specific H_1
- Composite H_0, H_1: we do get subjective component
  - different *E*-variables exist for same problem
  - they also involve priors

…bit this refers to **luckiness** rather than **belief** :

- H_0 false: the 'better' your prior, the more evidence against the null you get
- H_0 true: no matter what prior chosen, it is extremely unlikely that you get substantial evidence against null

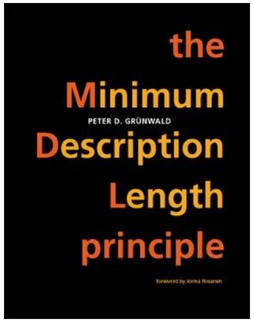
E-Values and evidence for $\boldsymbol{H_1}$

- In some special cases, we can use E-values in combination with composite H_0 and H_1 also to gain evidence for H_1
  - still different from Bayes
- These require a certain symmetry between H_0 and H_1 This works e.g. in t-test setting, with $\delta = \mu/\sigma$, if for some $\delta_1 \geq \delta_0$ we have:

$$H_0 : \delta \leq \delta_0 \ , \ H_1 : \delta > \delta_0$$

Evidence against?

- Does it even make sense to have evidence against H_0 without clear evidence for a specific H_1 ?
- Age-old debate. Likelihoodists think not. I disagree!
- Consider Quantum Random Number Generators. Ryabko and Monarev (2006) suggested to try to compress their output using **WinZip**
- If we can compress it by 200 bits, the null hypothesis of randomness (fair coin flips) gets an E-value of 2^{-200} . I think that pretty much disproves H_0 !
- more generally, there is a 1-1 correspondence between E-values and codelenghts **using a specific type of codes**



there’s so much more…

- **betting** interpretation (Shafer 2020, JRRS A)
  - Are all “unproblematic” extensions of likelihood (partial/conditional likelihood) really e-variables?
- can use e-values to build **always valid confidence sequences** (Howard, Ramdas et al. – many papers)
  - Our work is orthogonal to the discussion of ‘whether testing makes sense at all’!
- **e-values vs p-values:** calibration, merging by mixing etc (Vovk, Wang, Shafer – several papers)
- Practical applications developed in our group: Cox regression with optional stopping, 2x2 tables, …

Who did what?

- G., De Heide, Koolen. ***Safe Testing,*** *Arxiv 2020*
  shows e-values always exist and relation Bayes factors
- *evidence Interpretation of E-values: not written down yet*
- *All other stuff you have seen is not really new!*

Development of E-variables and the like: Glenn Shafer, Volodya Vovk

(game theoretic probability)

Aaditya Ramdas

- counterfactual issues p-values:
- 1960s (e.g. Pratt, Birnbaum), 1980s (prequential, Dawid)
- Type I errors with optional stopping: Robbins+students (+-1970). First appearance of E-variable: Levin (1975)

Optional Continuation, simple H_0

- S_j may be same function as S_{j-1}, e.g. (simple H_0)

$$S_1 = \frac{\int_{\Theta_1} p_\theta(X_1, \ldots, X_{n_1}) dW(\theta)}{p_0(X_1, \ldots, X_{n_1})} \qquad S_2 = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta)}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

- But choice of jth e-value S_j may also depend on previous X^{N_j}, Y^{N_j} , e.g.

$$S_2 = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta | X_1, \ldots, X_{n_1})}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

and then (full compatibility with Bayesian updating)

$$S_1 \cdot S_2 = \frac{\int p_\theta(X_1, \ldots, X_{N_2}) dW(\theta)}{p_0(X_1, \ldots, X_{N_2})}$$

I’ll only explain a special case: separated hypotheses

- Suppose we are willing to admit that we’ll only be able to tell H_0 and H_1 apart if $P \in H_0 \cup H_1'$ for some $H_1' \subset H_1$ that excludes points that are ‘too close’ to H_0 e.g.

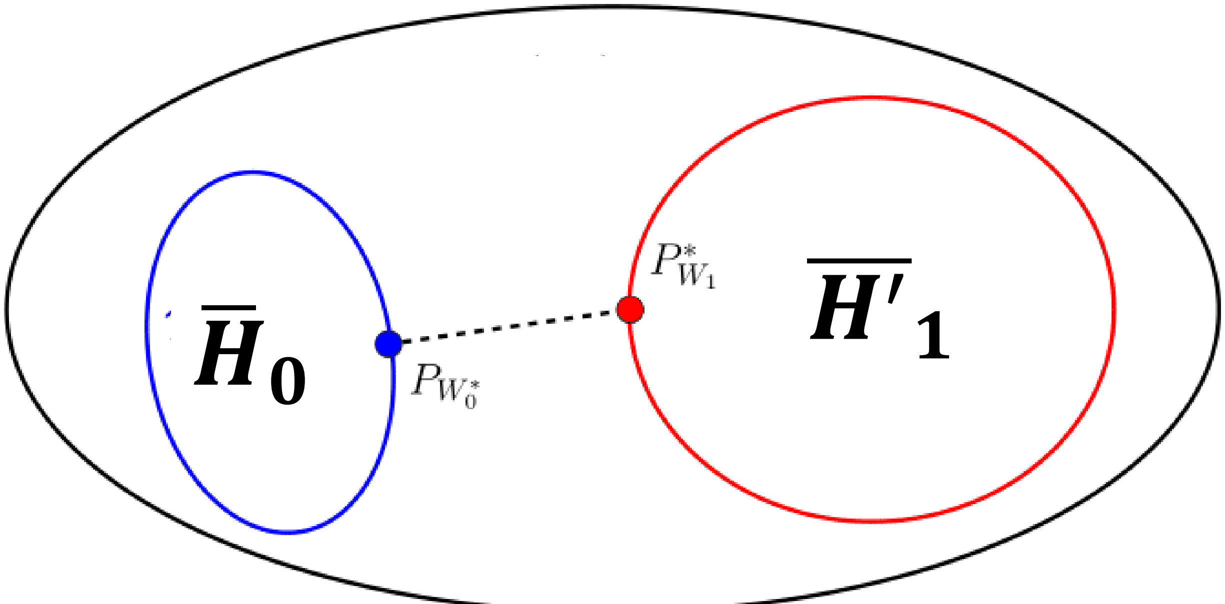
$$H_1' = \{P_\theta : \theta \in \Theta_1'\}, \Theta_1' = \{\theta \in \Theta_1 : \inf_{\theta_0 \in \Theta_0} \|\theta - \theta_0\|_2 \geq \delta\}$$

The best S-Value is given by the **Joint Information Projection (JIPr)**

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$\mathcal{W}_1$ set of all priors (prob distrs) on Θ'_1

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$



Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Here D is the relative entropy or Kullback-Leibler divergence, the central divergence measure in information theory:

$$D(P \| Q) := \mathbf{E}_{X^n \sim P} \left[\log \frac{p(X^n)}{q(X^n)} \right]$$

Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \text{arg} \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose (W_1^*, W_0^*) exists. Then $S^* := \frac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-variable relative to H_0. (b) it is in some special sense the 'best' E-variable!