Many Problems with P-values

General Article

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berk

ASA News AMERICA
Promoting t

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • To

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

nature

Explore content ⌄ Journal information ⌄ Publish with us ⌄

nature › news › article

Published: 27 August 2015

Over half of psychology studies fail reproducibility test

Monya Baker

Nature (2015) | Cite this article

Largest replication study to date casts doubt on many published positive results.

Many Problems with P-values

AND confidence intervals!

...nal information ⌄ Publish with us ⌄

... › news › article

Published: 27 August 2015

Over half of psychology studies fail reproducibility test

Monya Baker

Nature (2015) | Cite this article

3096 Accesses | 41 Citations | 1252 Altmetric | Metrics

Largest replication study to date casts doubt on many published positive results.

ASA News AMERICA... Promoting t...

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • To...

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

General Article

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berk...

A Particular Problem with P-values

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result (say, $p = 0.04$).
- ...**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?

ASA News
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

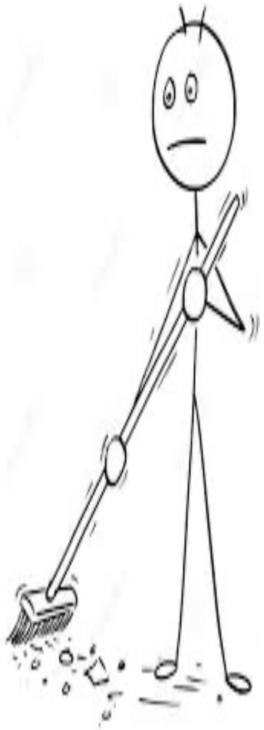
March 7, 2016

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

A 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:**

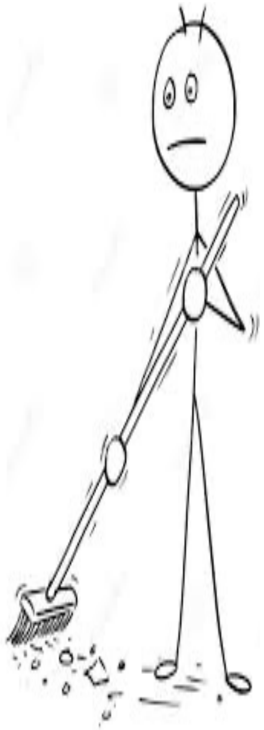
  sweep data together and re-calculate p-value
- **Is this p-hacking?** YES



A 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



Null Hypothesis Testing

H_0 represents null hypothesis

H_1 represents alternative hypothesis

...for data $X^n = (X_1, \dots, X_n)$

Both H_0 and H_1 are represented as (sets of) probability distributions

Example: z-test

Prototypical example: **z-test:**

$X_1, X_2, \ldots$ independently identically distributed (i.i.d.) $\sim N(\mu, 1)$
(Gaussian with mean μ and variance $\sigma^2 = 1$)

H_0: $\mu = \mu_0$

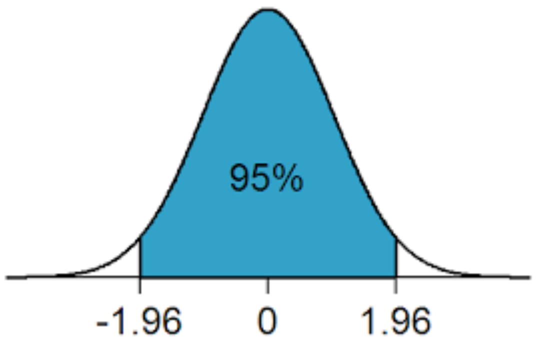
H_1: μ is "something else". comes in variations:

- $\mu = \delta$ for some given δ (simple-vs-simple)
- $\mu - \mu_0 \geq \delta$ (one-sided) or $|\mu - \mu_0| > \delta$ (two-sided)
- $\mu \neq \mu_0$ (two-sided)

Classical, p-value based testing

- I test new medication on n patients at level α
  n and α decided upon in advance
- p_n : p-value for null hypothesis H_0 at n

- If $p_n \leq \alpha$ I "reject" the null, otherwise I "accept" it
- two-sided z-test, standard $\alpha = 0.05 \Leftrightarrow$ "reject iff $|\bar{X} - \mu_0| \geq \frac{1.96}{\sqrt{n}}$ "



Simple H_0

Example: t-test

1-sample t-test:

$X_1, X_2, \ldots$ independently identically distributed (i.i.d.) $\sim N(\mu, \sigma^2)$
Gaussian with mean μ and variance σ^2. Define **effect size** $\delta := \mu/\sigma$

$H_0: \delta = \mu = 0$

H_1: δ is “something else”. Again comes in variations

Now H_0 is large set of distributions (one for each σ) : **Composite H_0**

Motivation Standard Procedure

- **Type-I error:** probability of rejecting null hypothesis even though it's true
  - false alarm; medication seems to work even though it doesn't

- By definition of p-value, for all $P \in H_0$,

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

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

Optional Continuation: what goes wrong?

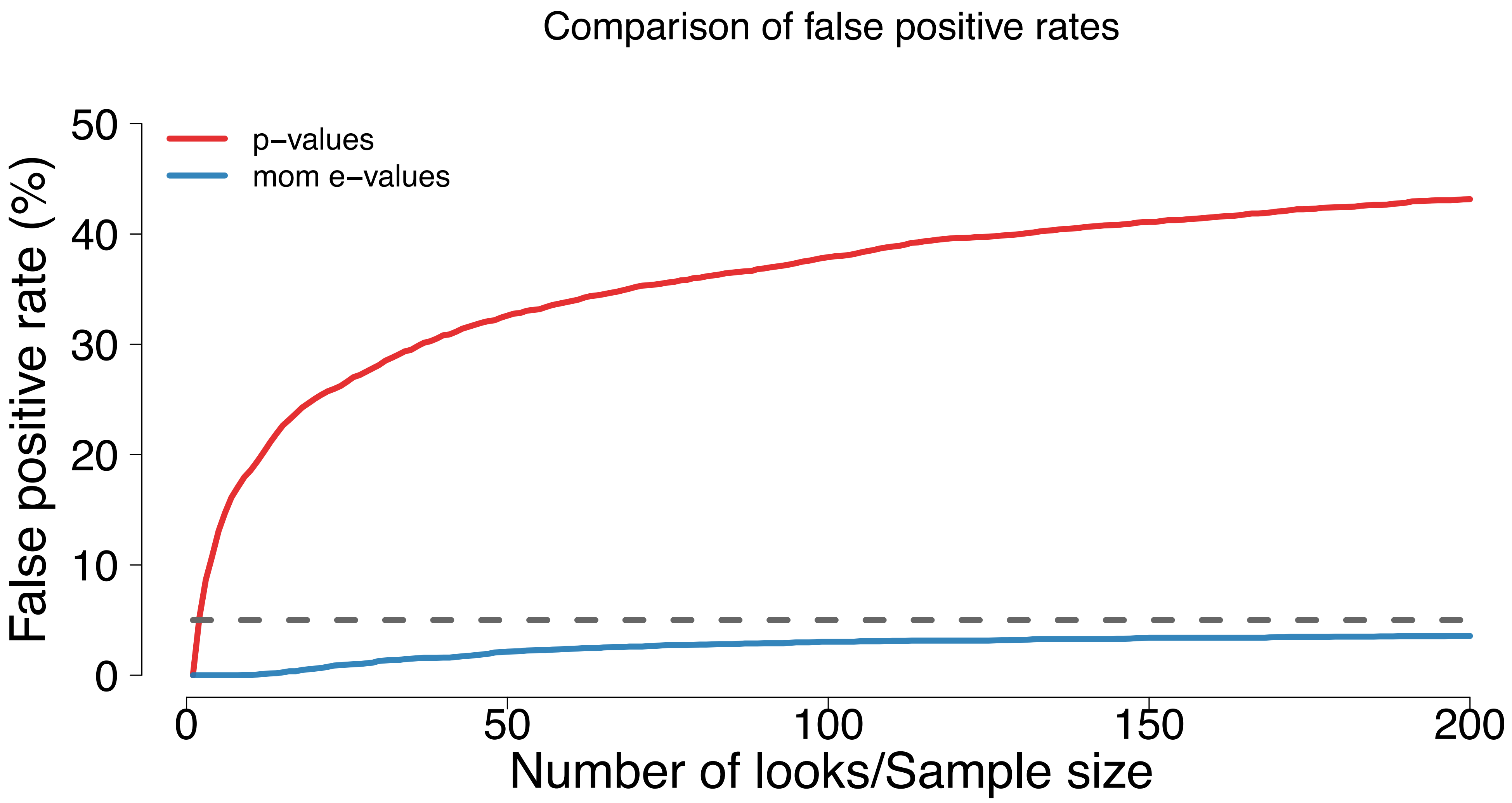
1. Do first test; observe $Y_{(1)} = (X_1, \ldots, X_{100})$
2. If significant ($p_{Y_{(1)}} < 0.05$) reject and stop
   else do 2nd test on 2nd batch $Y_{(2)} = (X_{101}, \ldots, X_{200})$
3. If significant ($p_{(Y_{(1)}, Y_{(2)})} < 0.05$) reject else accept

What goes wrong?

1. Do first test; observe $Y_{(1)} = (X_1, \dots, X_{100})$
2. If significant ($p_{Y_{(1)}} < 0.05$) reject and stop
   else do 2nd test on 2nd batch $Y_{(2)} = (X_{101}, \dots, X_{200})$
3. If significant ($p_{(Y_{(1)}, Y_{(2)})} < 0.05$) reject else accept

If $p_{Y_{(1)}}$ is strict p-value then $P_0\left(p_{Y_{(1)}} \leq 0.05\right) = 0.05$

total probability of rejecting under the null strictly larger than 0.05
⇒ Type-I error guarantee violated



This and subsequent graphs made by **Alexander Ly** – *thanks!*

E is the new P

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*-values solve **additional issues** related to p-values:

- *p*-values' reliance on **counterfactual knowledge**
- changing α **after the fact**
  - G., *Beyond* Neyman-Pearson, *PNAS*, 2024, Hemerik and Koning, *Stat. Science* 2025

- interpretation

…*but not all issues* ("e-hacking" is harder, but possible…)

E-values appear implicitly, without a name, in work of H. Robbins and students (late 1960s). Then nothing much happens until **2019** when following papers appear on arXiv:

Safe Testing

(G., De Heide, Koolen, now *Journal Royal Stat. Soc. B*)

E-Values: Calibration, Combination and Applications

(V. **Vovk**, R. Wang, now *Annals of Statistics)*

Testing by Betting

(G. Shafer, now *Journal Royal Stat. Soc. A*)

Universal Inference

L. Wasserman, A. **Ramdas**, S. Balakrishnan, now *PNAS*)

2025: 100s of papers on e-processes, anytime-valid confidence intervals, sequential testing by betting, with optional stopping, multiple testing,…

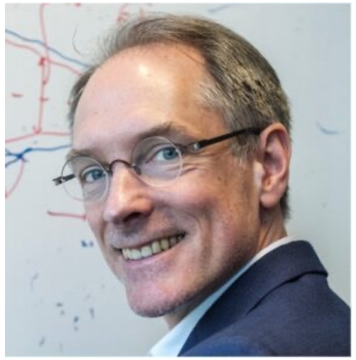
3 international workshops, attendants from Netflix, Booking, ...

Central Players

Aaditya Ramdas and his group at CMU

2023 Institute of Mathematical Statistics
Peter Hall Early Career Prize *“recognizing Dr. Ramdas’ outstanding potential to* ***shape the future of statistics****”*

2024 Presidential Early Career Award for Scientists and
. engineers (**PECASE**)



Yours truly, and my group at CWI and Leiden

2024 ERC Advanced Grant

The Early Pioneer: **Volodya Vovk**

Also Johanna Ziegel (ETH), Ruodu Wang (Waterloo), Glenn Shafer (Rutgers), W. Koolen (CWI), M. Larsson (CMU), J. Ruf (LSE), R. de Heide (Twente), M. Jordan (Berkeley), N. Koning (Erasmus), many others…

Paradigms

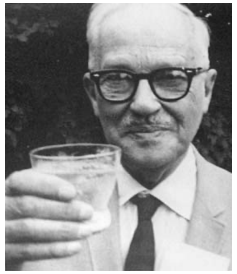
Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"

  Type-I **error guarantees/confidence intervals**

- **Fisher**ian/ "evidential"

  focus on **evidence** (as measured by p-values)



"standard" method: funny mix between these two

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"
  Type-I **error guarantees/confidence intervals**

- **Fisher**ian/ "evidential"
  focus on **evidence** (as measured by p-values)

E-values:

generalize α-validity methods to OC, OS, roving α

replace p- by e-value

E-values are a **frequentist** paradigm with α-validity and with evidence interpretation. **Now combination is natural!**

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"

  Type-I **error guarantees/confidence intervals**

- **Fisher**ian/ "evidential"

  focus on **evidence** (as measured by p-values)

E-values:

generalize α-validity methods to OC, OS, roving α

replace p- by e-value

Myth Nr 2:

"It is impossible to combine frequentist, α-validity methods with OS (optional stopping)/full sample plan flexibility"

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"
  Type-I **error guarantees/confidence intervals**

- **Fisher**ian/ "evidential"
  focus on **evidence** (as measured by p-values)

E-values:

generalize α-validity methods to OC, OS, roving α

replace p- by e-value

Myth Nr 2:

~~"It is impossible to combine frequentist α-validity methods with OS (optional stopping)/full sample plan flexibility"~~

Menu

1. **e-values and corresponding tests**
   - definition, likelihood ratio/Bayes factor interpretation (simple H_0)
   - solve optional continuation (OC) problem

2. from e-values to e-processes, from OC to OS

3. composite H_0

4. e-based confidence intervals

E stands for Expectation

An e-**variable** S for data Y is a nonnegative statistic, i.e. a nonnegative function of the data, such that for **all** $P_0 \in H_0$, we have

$$\mathbf{E}_{P_0}[S(Y)] \leq 1$$

The value $S(y)$ taken by S with data $Y = y$ is called the **e-value**

First Interpretation: Likelihood Ratios

Let $H_0 = \{P_0\}, H_1 = \{P_1\}, Y = X^n$

"Likelihoodists" measure evidence in favour of H_1 by

$$L(Y) = L(X^n) = \frac{p_1(X^n)}{p_0(X^n)}$$

$$\mathrm{E}_{P_0}[L(Y)] = \sum_{x^n} p_0(x^n) L(x^n) = \sum_{x^n} \cancel{p_0(x^n) \frac{p_1(x^n)}{p_0(x^n)}} = \sum_{x^n} p_1(x^n) = 1$$

… so L is an e-variable!

Example LR as E-Value

Bernoulli test: $p_\theta(X^n) = \theta^{n_1}(1-\theta)^{n-n_1}$

...for data $X^n = (X_1, \ldots, X_n)$ with each X_i either 0 or 1, $n_1 = \sum_{i=1}^{n} X_i$
(independent **coin tosses** with bias θ)

Example:

$H_0 \colon \theta = \frac{1}{2}$, $H_1 \colon \theta = \frac{3}{4}$

$L(X^n) = \frac{\left(\frac{3}{4}\right)^{n_1} \cdot \left(\frac{1}{4}\right)^{n_1 - 1}}{\left(\frac{1}{2}\right)^{n}}$ is an e-variable

First Interpretation: Likelihood Ratios

We may think of e-variables as (vast) generalizations of likelihood ratios

simple null: every likelihood ratio q/p_0 **is** an e-variable

composite null (as in t-test):

- likelihood ratios as usually defined are usually not e-variables
- e-variables may look completely different from likelihood ratios.

Yet still helps to think about e-variables as "something like likelihood ratios", measuring "evidence against null"

E also stands for Evidence

The Fundamental Property

Let S be an arbitrary e-variable, i.e. for $P \in H_0$, $\mathbf{E}_P[S(Y)] \leq 1$. Then also, for all $0 < \alpha \leq 1$,

$$P\left(S(Y) \geq \frac{1}{\alpha}\right) \leq \alpha$$

The probability, under the null, that an e-variable is large, is small

e.g. the probability that it exceeds 20, is bounded by 0.05

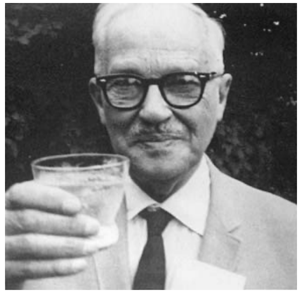
E-values behave “a bit” like reciprocals of p-values

E-value based Tests

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

- By the Fundamental Property, this Test has α-validity. For example:

- Test which rejects H_0 iff $S(X^n) \geq 20$ has **Type-I Error Bound** of 0.05



E-value based tests remain valid under optional continuation

- Suppose we observe $Y_{(1)}, Y_{(2)}, \ldots$
  - $Y_{(j)}$: data from j-th study (itself a sample or a summary statistic)

Optional Continuation

- Suppose we observe $Y_{(1)}, Y_{(2)}, \ldots$
  - $Y_{(j)}$: data from j-th study

- We first evaluate some e-variable $S_{(1)}$ on $Y_{(1)}$.
- If outcome in certain range (e.g. boss thinks result promising enough to collect more data) then....

  we evaluate some e-variable $S_{(2)}$ on $Y_{(2)}$,

  otherwise we **stop.**

- We first evaluate $S_{(1)}$.
- If outcome is in certain range then we evaluate $S_{(2)}$; otherwise **stop.**
- If outcome of $S_{(2)}$ is in certain range we compute $S_{(3)}$, else **stop**.
- ...and so on
- ...when we finally stop, after say τ studies, we report as final result the product $S^{(\tau)} := \prod_{j=1}^{\tau} S_{(j)}$

First insight, informally: the product $S^{(\tau)}$ is itself an e-variable, i.e. $\mathbf{E}\left[S^{(\tau)}\right] \leq 1$ irrespective of the stop/continue-rule τ used

- Procedure "reject H_0 iff $S^{(\tau)} \geq \alpha^{-1}$" has Type-I error bounded by α
- To implement procedure we do not need to know definition of stop/continue-rule. We only need to know if we actually stop or not

- We first evaluate $S_{(1)}$.
- If outcome is in certain range then we evaluate $S_{(2)}$; otherwise **stop.**
- If outcome of $S_{(2)}$ is in certain range we compute $S_{(3)}$, else **stop**.
- ...and so on
- ...when we finally stop, after say τ studies, we report as final result the product $S^{(\tau)} := \prod_{j=1}^{\tau} S_{(j)}$

We solved Optional Continuation Problem!

First insight, informally: the product $S^{(\tau)}$ is itself an e-variable, i.e. $\mathbf{E}\left[S^{(\tau)}\right] \leq 1$ irrespective of the stop/continue-rule τ used

- Procedure "Reject H_0 iff $S^{(\tau)} \geq \alpha^{-1}$" has Type-I error bounded by α
- To implement procedure we do not need to know definition of stop/continue-rule. We only need to know if we actually stop or not

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 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 $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

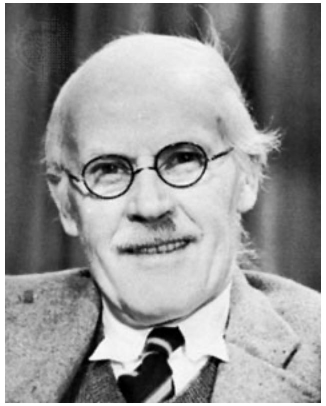
Bayes factor of form

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

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!



Optional Continuation Revisited

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

$$S_{(1)} = \frac{\int_{\Theta_1} p_\theta(X_1, \ldots, X_{n_1}) dW(\theta)}{p_0(X_1, \ldots, X_{n_1})} \quad S_{(2)} = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta)}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

Optional Continuation Revisited

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

$$S_{(1)} = \frac{\int_{\Theta_1} p_\theta(X_1, \ldots, X_{n_1}) dW(\theta)}{p_0(X_1, \ldots, X_{n_1})} \quad S_{(2)} = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta)}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

- But choice of jth function $S_{(j)}$ 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)

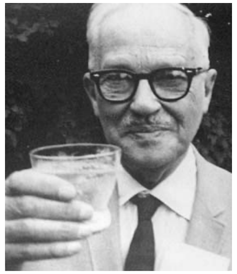
$$S_{(1)} \cdot S_{(2)} = \frac{\int p_\theta(X_1, \ldots, X_{N_2}) dW(\theta)}{p_0(X_1, \ldots, X_{N_2})}$$

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"

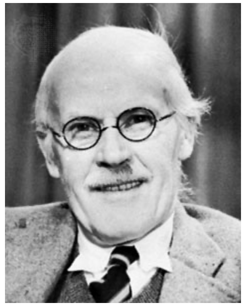
  Type-I **error guarantees/confidence intervals**



- **Fisher**ian/ "evidential"

  focus on evidence (as measured by p-values)

- Bayesian methods – use of priors
  - **Jeffreys**ian
  - General



Menu

1. e-values and corresponding tests
    - definition, likelihood ratio/Bayes factor interpretation (simple H_0)
    - solve optional continuation (OC) problem

2. **from e-values to e-processes, from OC to OS**

3. composite H_0

4. e-based confidence intervals

E-Processes

An **e-process** is a sequence of functions $S_1, S_2, \ldots$ with $S_i \geq 0$ a statistic of first i data points (i.e. S_i is a function of X^i) s.t. for **all** $P_0 \in H_0$, and **every*** stopping **rule**/time $\boldsymbol{\tau}$ we have

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

i.e. under arbitrary τ, process turns into e-variable.

Example of stopping time:

$\tau = n$ for fixed n

E-Processes

An **e-process** is a sequence of functions $S_1, S_2, \ldots$ with $S_i \geq 0$ a statistic of first i data points (i.e. S_i is a function of X^i) s.t. for **all** $P_0 \in H_0$, and **every*** stopping **rule**/time $\boldsymbol{\tau}$ we have

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

i.e. under arbitrary τ, process turns into e-variable.

Examples of stopping times:

$\tau = n$ for fixed n

τ: stop at smallest n such that X^n contains three values ≥ 1

τ: stop at smallest n at which $S_n(X^n) \geq 20$

Likelihood Ratios are E-Processes (simple H_0)

Let $H_0 = \{P_0\}, H_1 = \{P_1\}$, and $L_1, L_2, \ldots$ defined by

$$L_n(X^n) := \frac{p_1(X^n)}{p_0(X^n)}$$

- For every stopping time/rule τ,

$$\mathrm{E}_{P_0}[L_\tau(X^\tau)] = \sum_{x^\tau \text{ for which rule stops}} p_0(x^\tau) L_\tau(x^\tau) = \sum_{x^\tau} \cancel{p_0(x^\tau)} \frac{p_1(x^\tau)}{\cancel{p_0(x^\tau)}} = 1,$$

so $L_1, L_2, \ldots$ is an e-process!

Example

Bernoulli test: $p_\theta(X^\tau) = \theta^{\tau_1}(1-\theta)^{\tau-\tau_1}$

$H_0: \theta = \frac{1}{2}$, $H_1: \theta = \frac{3}{4}$

$L(X^\tau) = \frac{\left(\frac{3}{4}\right)^{\tau_1} \cdot \left(\frac{1}{4}\right)^{\tau-\tau_1}}{\left(\frac{1}{2}\right)^{\tau}}$ is an e-variable for any stopping time/rule τ

τ: fixed n

τ: stop as soon as you've seen 3 ones and then a zero

τ: stop as soon as $L(X^\tau) \geq 20$

E-process-based tests are α-valid under OS

Suppose $S_1, S_2, \ldots$ is an e-process. Then S_τ is an e-variable. By the fundamental property, we have Type-I error guarantee,

$$P_0\left(S_\tau(X^\tau) \geq \frac{1}{\alpha}\right) \leq \alpha$$

Suppose we reject H_0 iff $S_\tau(X^\tau) \geq \frac{1}{\alpha}$ for some stopping time τ . Then we have a Type-I error guarantee of α : validity under **OS (Optional Stopping)**

E-process-based tests are α-valid under OS

Suppose we reject H_0 iff $S_\tau(X^\tau) \geq \frac{1}{\alpha}$ for some stopping time τ . Then we have a Type-I error guarantee of α : validity under **OS (Optional Stopping)**

This works for every τ, i,e. irrespective of when and for what reason we stopped - even if we do not know for what reason we stopped!

Bayes factors with simple $\boldsymbol{H_0}$ provide e-processes

Just like Bayes factors with simple H_0 for data $Y = X^n$ is an e-variable, the Bayes factor process $B_1, B_2, \ldots$ with $B_n = \frac{p_{W_1}(X^n)}{p_0(X^n)}$ is an e-process.

Menu

1. e-values and corresponding tests
   - definition, likelihood ratio/Bayes factor interpretation (simple H_0)
   - solve optional continuation (OC) problem

2. from e-values to e-processes, from OC to OS

3. **composite $\boldsymbol{H_0}$**

4. e-based confidence intervals

Composite H_0:
Bayes may not give e-variable!

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** $P_0 \in H_0$:

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

If H_0 composite then likelihood cancellation argument does not work any more, and Bayes factors usually don't give e-values any more

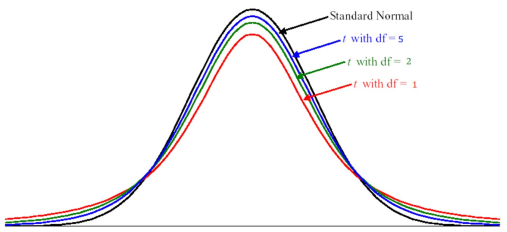
$$M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$$

- ...but (special case of Theorem 1 of G., De Heide, Koolen '24 JRSSB):

For every $\boldsymbol{W_1}$ there exists a special, unique prior $\boldsymbol{W_0^*}$ (sometimes highly 'nonstandard') for which Bayes factors do become e-values

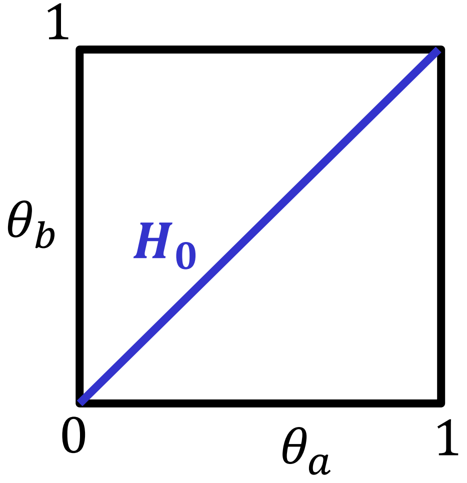
Examples

1-sample t-test (Perez-Ortiz et al., *Ann. Stats. 2024*)



- Putting Jeffreys' improper $\frac{1}{\sigma}$ -prior on variance in null and alternative gives (optimal!) e-variable *…standard Bayes factor is e-process*

2x2 contingency tables (Turner et al., *JSPI 2024)*

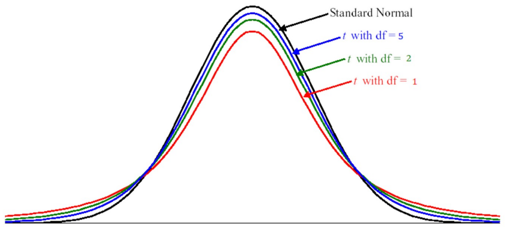


- optimal e-variable uses prior very different from Jeffreys'
- …standard Bayes factors: not at all e-processes

Examples

1-sample t-test (Perez-Ortiz et al., *Ann. Stats. 2024*)

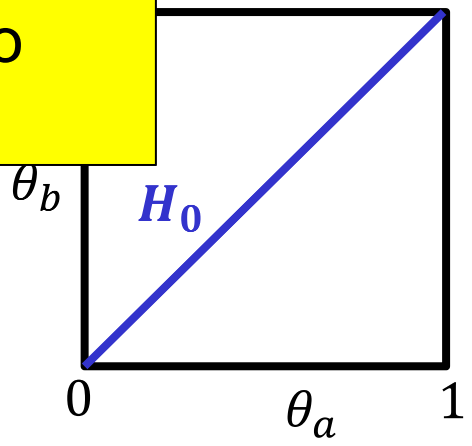
- Putting Jeffreys' improper $\frac{1}{\sigma}$ -prior on variance in null and alternative gives (optimal!) e-variable *…standard Bayes factor is e-process*



Myth Nr 3:

"Taking prior-weighted averages over parameters makes no sense in frequentist/non-Bayesian inference "

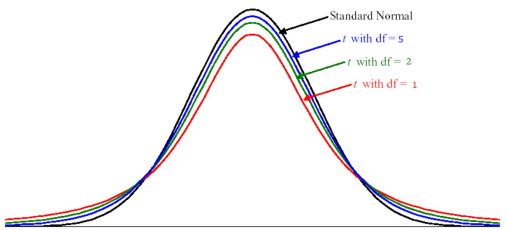
- optimal e-variable uses prior very different from Jeffreys
- …standard Bayes factors: not at all e-processes



Examples

1-sample t-test (Perez-Ortiz et al., *Ann. Stats. 2024*)

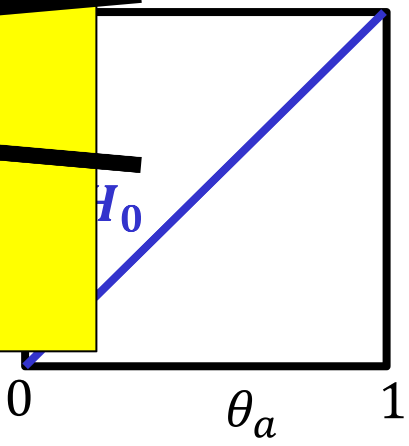
- Putting Jeffreys' improper $\frac{1}{\sigma}$-prior on variance in null and alternative gives (optimal!) e-variable …*standard Bayes factor is e-process*



Myth Nr 3:

~~"Taking prior-weighted averages over parameters makes no sense in frequentist/non-Bayesian inference"~~

The truth: sometimes it does, sometimes it doesn't



Nonparametric H_0

…will just give an extremely simple example here:

Testing the **Mean** of a Bounded Random Variable

(Waudby-Smith and Ramdas, *JRSS B*, 2024)

$X_1, X_2, \ldots \text{ iid} \sim P\,,\ X_i \in [-1,1]$

$$H_0\!: \mathbf{E}_P[X_i] = \mu$$

i.e. H_0 consists of all P with mean $\mathbf{E}_P[X_i] = \mu$.

We assume nothing further about *P*

Nonparametric $\boldsymbol{H_0}$

set $\boldsymbol{s_\lambda(x) := 1 + \lambda(x - \mu)}$

defined for fixed $\mu \in [-1,1]$ and all $\lambda \in \left[-\frac{1}{2}, \frac{1}{2}\right]$

$s_\lambda\,(X)$ is e-variable for H_0': $\mathbf{E}[X\,] = \mu$

…since under any $P \in H_0'$: $\mathbf{E}_P[s_\lambda(X\,)] = 1 + \lambda\,(\mu\ - \mu) = 1$

$S_{\lambda,1}, S_{\lambda,,2}$, ... with $S_{\lambda,n} = \prod_{i=1..n}\ s_\lambda\,(X_i)$ is an e-process for H_0

- follows easily from i.i.d. assumption

Nonparametric $\boldsymbol{H_0}$

set $\boldsymbol{s_\lambda(x) := 1 + \lambda(x - \mu)}$

$S_{\lambda,1}, S_{\lambda,2}$, ... with $S_{\lambda,n} = \prod_{i=1..n} s_\lambda(X_i)$ is an e-process for H_0

We can "learn" λ from the data – without compromising "e-processness"/α-validity:

e-process

- set $\hat{\lambda}_n$ to be the λ maximizing $S_{\lambda,n}$ ("maximize ~~likelihood~~")
- Then $S^*_{\lambda,n} := \prod_{i=1..n} s_{\hat{\lambda}_{i-1}}(X_i)$ is still an e-process

Nonparametric $\boldsymbol{H_0}$

set $\boldsymbol{s_\lambda(x) := 1 + \lambda(x - \mu)}$

$S_{\lambda,1}, S_{\lambda,2}$, ... with $S_{\lambda,n} = \prod_{i=1..n} s_\lambda(X_i)$ is an e-process for H_0

We can "learn" λ from the data – without compromising "e-processness"/α-validity

…can also put prior on λ and learn it in **pseudo-Bayesian manner**, by an analogue of Bayes' rule in which likelihoods are replaced by e-processes

Nonparametric E vs Bayes

nonparametric Bayes: need to put prior on the full, infinite-dimensional set of distributions

e-variables: it suffices to put a “prior” on single nuisance parameter!

Q: **Peter**, **why aren’t you a Bayesian?**

A: Because I don’t see why I would need to design a prior over an incredibly large set of distributions if I am only interested in learning a single, simple parameter (argument made before by Rob(b)ins, others)

Menu

1. e-values and corresponding tests
   - definition, likelihood ratio/Bayes factor interpretation (simple H_0)
   - solve optional continuation (OC) problem

2. from e-values to e-processes, from OC to OS

3. composite H_0

4. **e-based confidence intervals**

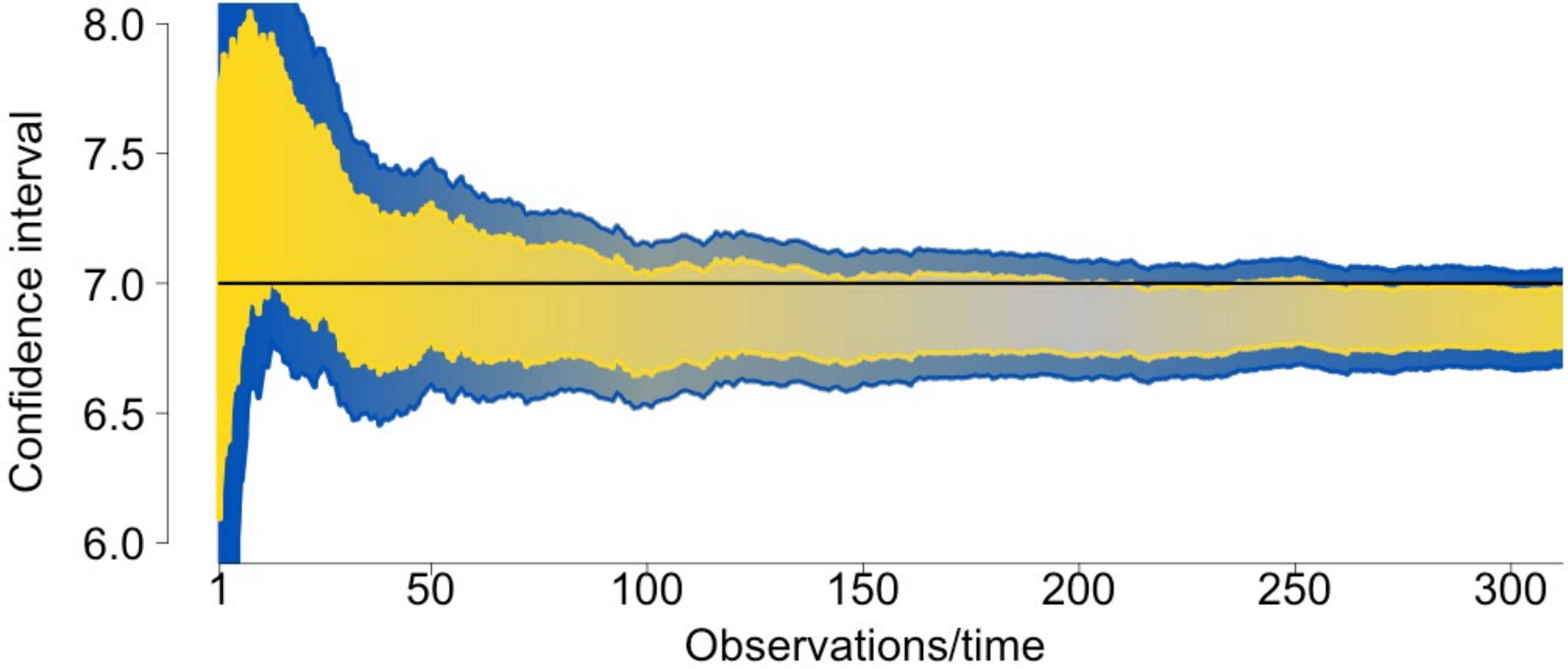
E-Based vs. Standard Confidence Intervals

- Suppose $X_1, X_2, \ldots$ i.i.d. $\sim \mathrm{N}(\mu, 1)$ (**z-test**)
- Standard CI = Bayesian 95% credible interval (noninformative prior)

$$\left[\widehat{\mu}_n - \frac{1.96}{\sqrt{n}}, \widehat{\mu}_n + \frac{1.96}{\sqrt{n}}\right]$$

- e-process based CI based on Bayes factor with same prior:

$$\left[\widehat{\mu}_n - \sqrt{\frac{6 + \log(n)}{n}}, \widehat{\mu}_n + \sqrt{\frac{6 + \log(n)}{n}}\right]$$



Yellow: Bayes 95% credible interval based on noninformative prior = standard confidence interval = $\bar{X} \pm 1.96/\sqrt{n}$

Blue: 95% e-based interval based on same prior

e-based intervals are anytime-valid

- The e-process based intervals are **anytime-valid**

- The e-process property ensures that the probability that the “true” parameter will **ever** fall out of the interval is bounded by α
  - it does not matter how often you look !

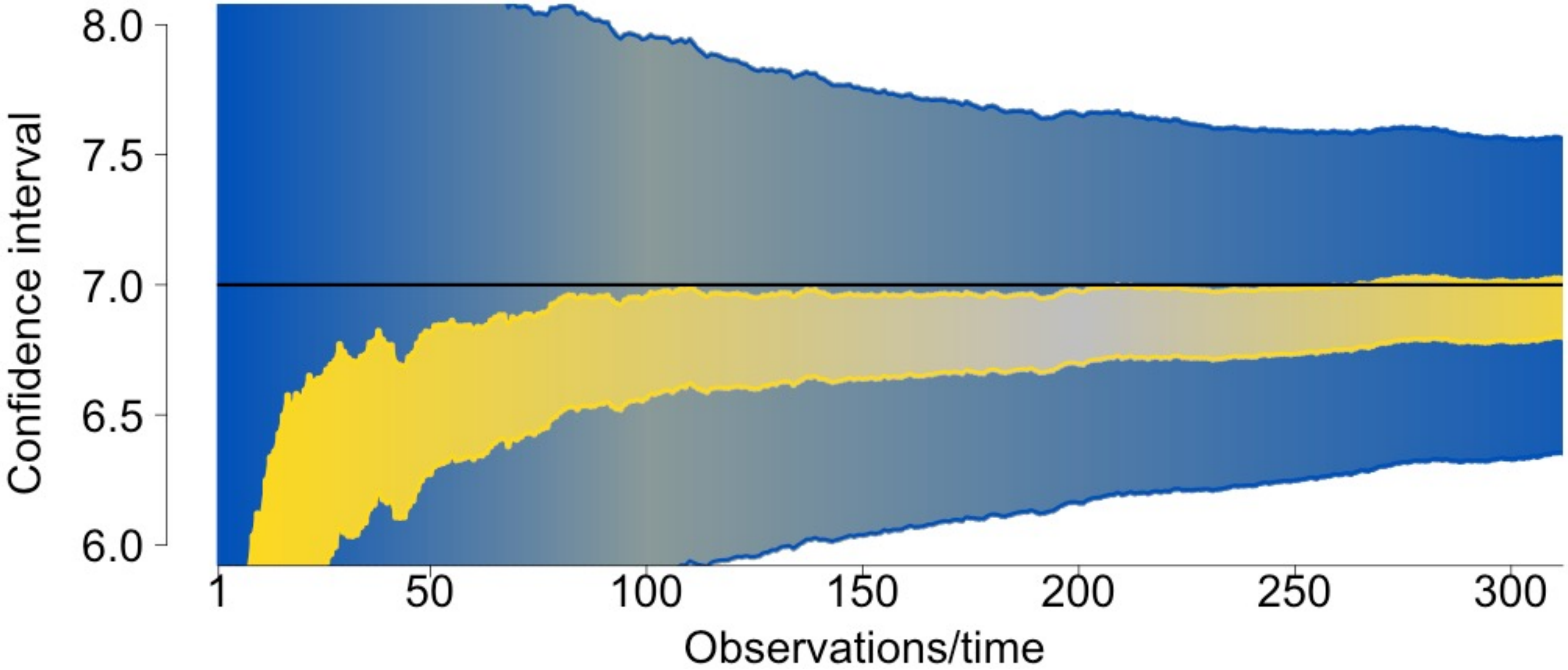
- In contrast, standard CIs every now and then exclude the true parameter value. In fact (Pace & Salvan, 2020), with standard CIs, **probability that at some point in future you will get interval which does not overlap with your current interval is 1**

Subjective and Objective, at same time

- e-process based CIs rely on a prior, just like Bayesian posteriors…

…but they **remain valid** irrespective of prior you use

…suppose for example you have a **pretty mistaken prior belief** that $\theta = 0$, with variance 0.5 …



Confidence interval
8.0
7.5
7.0
6.5
6.0
1
50
100
150
200
250
300
Observations/time

Subjective and Objective, at same time

- "E-Posteriors" and the CIs they induce rely on a prior, just like Bayesian posteriors…

  …but they **remain valid** irrespective of prior you use

with a bad prior, "e-posterior" gets wide rather than wrong

G. The E-Posterior, *Phil. Trans. Royal Soc. London A*, 2023

Subjective and Objective, at same time

- “E-Posteriors” and the CIs they induce rely on a prior, just like Bayesian posteriors…

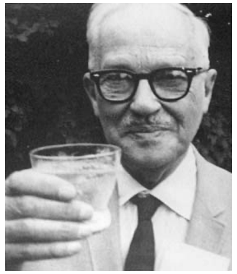
  …but they **remain valid** irrespective of prior you use

with a bad prior, “e-posterior” gets wide rather than wrong

Note: nonstandard Bayes intervals advocated by Pawel and Wagenmakers (*American Statistician ‘23),* coincide, for 1-d models, with e-based intervals

Paradigms

Orthodox/Classical//Frequentist methods

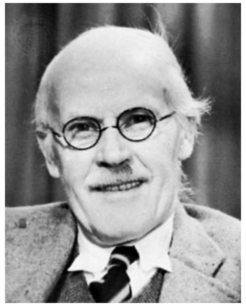


- **Neyman**-Pearson/ "α-validity"

  Type-I **error guarantees/confidence intervals**

- **Fisher**ian/ "evidential"

  focus on evidence (as measured by p-values)



- Bayesian methods – use of priors
  - **Jeffreys**ian
  - General

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"
  Type-I **error guarantees/confidence intervals**
- **Fisher**ian/ "evidential"
  focus on evidence (as measured by p-values)
- Bayesian methods – use of priors
  - **Jeffreys**ian
  - General

E-values:

generalization of α-validity methods to cases with OS, OC, roving α

replace p-value by e-value and interpret as "evidence"

can also use prior distributions, but these have different interpreations

Paradigms

Orthodox/Classical//Frequentist methods

- **Neyman**-Pearson/ "α-validity"

  Typ

- **Fis**

  foc

- Ba
  -
  - General

E-values:

generalization of α-validity

e

different interpreations

Jim Berger (IMS Neyman Lecture, 2003):

Could Fisher, Jeffreys and Neyman have agreed on testing?

I claim: using e-values, they could (or rather: **should**) have

see G., the E-Posterior, Proc. Roy. Soc. London A

Take Home

- E-values: more robust & flexible than p-value and Bayes factor

- There are many more cool things (roving α, universality) - one could view e-value theory as a basis for a **G**rand **U**nified theory of **S**tatistic**s** (**GUTS**)

- There are (of course) also issues though. No time to tell you about them…

- Read/Do More?
  - Ly et al., R Package *SafeStats* on CRAN*, 2020*
  - Ly et al. A Tutorial on Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments Based On E-Values. *PsyArXiv, 2025*

Extra Slides

Pseudo-Bayesian Learning λ

Set e-variables $\boldsymbol{s_\lambda(x) := 1 + \lambda(x - \mu)}$

Now put "**prior**" w on λ

Set $S_i^\circ = \int s_\lambda(x_i) w\left(\lambda | x^{i-1}\right) \mathrm{d}\lambda$

with "**posterior**" $w\left(\lambda | x^{i-1}\right) \propto w(\lambda) \prod_{j=1}^{i-1} s_\lambda\left(X_j\right)$

Note S_i is still e-variable for $H_0' : \mathbf{E}[X_i] = \mu$

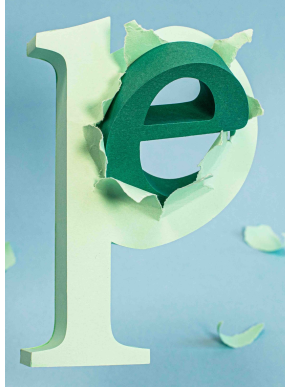
Set $S_i := \prod_{i=1..n} S_i^\circ = \int \prod_{i=1}^{n} s_\lambda(X_i) w(\lambda) \mathrm{d}\lambda$

All-or-Nothing E-Variable

There exists an e-variable $\boldsymbol{S}_{\mathrm{np}}$ s.t. S_{np}–based test when applied at sample size n rejects iff NP based on p-value p for sample size n rejects:

$S_n(X^n) = 0$ if $p > \alpha$; $S_n(X^n) = 1/\alpha$ if $p \leq \alpha$

Universality of E-Methods



- **All of Neyman-Pearson** ("α-validity") can be mimicked with e instead of p
  - everybody writes down p-values, but NP never asked us to!
  - S_{np} achieves optimal **1-shot** performance: statistical **power**.
  - Yet S_{np} hopeless when **optional continuation** comes into play **(why?).**
    …so e-lovers prefer GRO (growth-rate optimal – an analogue of power) e-variables instead

Take Home

E-values provide notion of evidence more robust & flexible than p-value

- Corresponding tests/CIs more robust/flexible than standard CIs, more robust than Bayes credible intervals
- Optional Continuation; **Data-Dependent α**

  G., *Beyond* Neyman-Pearson, *PNAS*, 2024, Hemerik and Koning, *Stat. Science* 2025; Koning "Continuous Testing", *arXiv* 2024

  "Quasi-Conditional Inference": Bridge between Bayes and Frequentist (G.2023)

Price to pay: need more data in single study (less power, wider CIs). Yet:

- this can often be mitigated by optional (earlier) stopping…

Main Future Challenges:

- design e-methods for complex statistical problems

competitiveness: the power of e-based tests

Compare standard NP (Neyman-Pearson) test with GRO e-value-based test as function of point alternative $\delta = \mu$ in z-test.

Sample size n_{np} defined so that NP test achieves required power 0.8

- Fixed sample size n_e to achieve power 0.8 with $S^{\langle n_e \rangle}$:

$n_e \approx 2.2\, n_{\mathrm{np}}$ at $\alpha = 0.05$, $n_e \approx 1.7\, n_{\mathrm{np}}$ at $\alpha = 0.01$

…if we use standard (GRO) e-values

with OS, very competitive!

Variable sample size τ_e defined so as to achieve power 0.8 with $S^{\langle \tau_e \rangle}$ based on **aggressive optional stopping** at $\alpha = 0.05$

- can only be done if decomposition property holds
- for all δ:

$$\mathbf{E}_{P_\delta}[\tau_e] < n_{\mathrm{np}} \; ; 1.4 \, n_{np} \leq \max \tau_e \leq \; 1.7 \, n_{\mathrm{np}}$$

- Proof by simulation, confirmed for small δ by Brownian motion analysis

The tragedy of the commons

Essentially same for any other model we tried:

*on **average** e-variable approach with OS is very competitive with a classical Neyman-Pearson (NP) approach in terms of sample size; and you get a much more robust notion of evidence from it!*

*but every individual research group that uses e-based tests and needs power guarantees has to prepare for using more data than NP in **worst-case. So they have incentive to be greedy and go for NP***