Discussion of a Special Issue of The American Statistician

Technical overview of some of the alternative procedures

Daniel R. Jeske

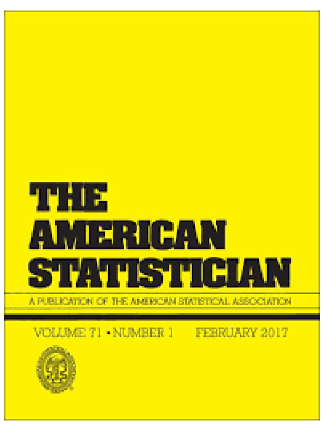
Professor

Department of Statistics

University of California – Riverside

Editor-in-Chief, *The American Statistician*

May 23, 2019



NISS National Institute of Statistical Sciences

UC RIVERSIDE UNIVERSITY OF CALIFORNIA

ASA AMERICAN STATISTICAL ASSOCIATION *Promoting the Practice and Profession of Statistics*

Wendy's television commercial (1984)

Where's the Beef?

Wendy's television commercial (1984)

Where's the Beef?

In the TAS Special Issue!

Setting the Scene

- **43 papers** in the special issue can be grouped into 5 categories:
    1. Getting to post $p<0.05$ era
    2. Interpreting and using p
    3. **Supplementing** or replacing p
    4. Adopting more **holistic** approaches
    5. Reforming institutions: changing publication policies and statistical education
- I will **discuss 7 papers** in the special issue that I found interesting.
- **Not all of the ideas in the papers are new**. Some of the papers highlight and/or add emphasis to previous work published elsewhere
- My specific aims are:
    1. Share some of the **techniques you might use.**
    2. Provide you with the **main ideas** and a **glimpse of some technical ideas**.
    3. Tweak your curiosity enough that you **look further at the special issue.**
- The P-value topic brings both **Bayesian and Frequentist** thinking into the conversation.

One Last Thing Before We Really Start

A nice paper in the special issue that reviews the *long history* of p-values, including their origins, the controversies, and many of the principal characters involved in these facets:

"**Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing**,"

by Lee Kennedy-Shaffer

Abandon Statistical Significance

McShane, Gal, Gelman, Robert and Tackett

Terminology

- No bright line threshold for reporting p-value results. Report **continuous p-values** (i.e., not p<.05 or p<.01). It does **NOT** mean we no longer should use p-values.
- **Avoid using the term "statistically significant"** to avoid confusion with scientifically important.
- It should be recognized that a small p-value is a poor measure of evidence against a null because it **only signals that there is a problem with at least one assumption** behind it, without saying which one.
- **Sharp null hypotheses are poorly suited** for statistical inference.
- Authors (and editors) should place **more emphasis on what motivates the research** questions by discussing 'currently subordinate factors,' such as prior evidence and possible mechanisms for real effects.

Terminology

TAS Topic-Contributed Session at JSM 2019
Monday, July 29th, CC-110

Editor's Choice: Papers Published in The American Statistician During 2018

- 10:35 AM — Abandon Statistical Significance
  Blakeley McShane,; Andrew Gelman, Christian Robert, David Gal, Jennifer Tackett
- 10:55 AM — On Mixture Alternatives and Wilcoxon's Signed-Rank Test
  Jonathan Rosenblatt, Yoav Benjamini
- 11:15 AM — A Bayesian Survival Analysis of a Historical Dataset: How Long Do Popes Live?
  Luciana Dalla Valle, Julian Stander, Mario Cortina-Borja
- 11:35 AM — Guns and Suicides
  Danilo Santa Cruz Coelho, Daniel Cerqueira, Marcelo Fernandes, Jony Pinto Junior
- 11:55 AM — Forecasting at Scale
  Sean Taylor
- 12:15 PM — Floor Discussion

Many thanks to *Biometrics Section*, *Section on Bayesian Statistical Science*, and *Section on Statistical Learning and Data Science*

Daniel R. Jeske, Department of Statistics, University of California, Riverside

Context

Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values

Sander Greenland

s - values

Let *p* denote the probability of an event *E*. Suppose we find *s* such that $p = \left(\frac{1}{2}\right)^s$.

This expresses *p* as the probability of getting *s* consecutive heads in tosses of a fair coin.

The *s*-value, defined as $s = -\log_2(p)$, is a translation of how likely or unlikely *E* was.

Context

Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values

Sander Greenland

s - values

Let p denote the probability of an event E. Suppose we find s such that $p = \left(\frac{1}{2}\right)^s$.

This expresses p as the probability of getting s consecutive heads in tosses of a fair coin.

The s-value, defined as $s = -\log_2(p)$, is a translation of how likely or unlikely E was.

The s-value *contextualizes* the p-value by representing the evidence it conveys against the null as the same evidence that seeing all heads in s tosses of a coin would convey against a hypothesis thay the coin is fair.

Context

Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values

Sander Greenland

For example, $s = -\log_2(.05) = 4.3$. Considering the .05 threshold as evidence against a null is no different than doubting a coin is fair because 4 tosses in a row came up heads. Is that really strong evidence?

On the other hand, $s = -\log_2(.005) = 7.6$.

Improving the Use of P-Values

Daniel Benjamin and James O. Berger

- If using the current language of ‘statistical significance’ for a novel discovery, **replace the .05 threshold with .005**. Refer to discoveries with a p-value between .005 and .05 as ‘suggestive,’ rather than ‘significant.’

Thresholds

Improving the Use of P-Values

Daniel Benjamin and James O. Berger

- If using the current language of 'statistical significance' for a novel discovery, **replace the .05 threshold with .005**. Refer to discoveries with a p-value between .005 and .05 as 'suggestive,' rather than 'significant.'

- When reporting a p-value in a test of a hypothesis $\mathrm{H}_0 : \theta = \theta_0$ vs. $\mathrm{H}_1 : \theta \neq \theta_0$ **also report**

$$P(H_0 \mid X = x) = \left[1 + \frac{1-\pi_0}{\pi_0} \times \frac{1}{\mathrm{BF}_{0:1}}\right]^{-1} \geq \left[1 + \frac{1-\pi_0}{\pi_0} \times \frac{1}{-e\,p\log(p)}\right]^{-1}$$

$\pi_0 = .5$

p	.10	.05	.01	.005	.001
$P(H_0 \mid X = x)$	.385	.289	.111	.067	.018

Thresholds

Improving the Use of P-Values

Daniel Benjamin and James O. Berger

$H_0 : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

Marginal density of X : $m(x) = \pi_0 f(x \mid \theta_0) + (1 - \pi_0) \int f(x \mid \theta) g(\theta) d\theta$

$$P(H_0 \mid X = x) = \frac{f(x \mid \theta_0)\pi_0}{m(x)}$$

$$= \left[1 + \frac{1 - \pi_0}{\pi_0} \times \frac{1}{\mathrm{BF}_{0:1}} \right]^{-1} ,$$

where $\mathrm{BF}_{0:1} = \dfrac{f(x \mid \theta_0)}{\int f(x \mid \theta) g(\theta) d\theta}$.

Improving the Use of P-Values

Daniel Benjamin and James O. Berger

From here there are a variety of ways to establish bounds for BF_{01} with one such way derived in Berger and Selke (JASA, 1987) and the later Selke (*The American Statistician*, 2001):

p is the observed data and p, given a, is distributed as beta$(a,1)$

$$H_0 : a = 1$$
$$H_1 : a \neq 1$$

$$\mathrm{BF}_{0:1} = \frac{1}{\int_0^1 a p^{a-1} g(a) da} \geq \frac{1}{\max_a \left(a p^{a-1} \right)} = -e\, p \log(p), \text{ and so}$$

$$P(H_0 \mid X = x) = \left[1 + \frac{1-\pi_0}{\pi_0} \times \frac{1}{\mathrm{BF}_{0:1}} \right]^{-1} \geq \left[1 + \frac{1-\pi_0}{\pi_0} \times \frac{1}{-e\, p \log(p)} \right]^{-1}$$

Thresholds

Aligning P-values and Bayes Factors

Jonathan Rougier

$X_1, \ldots, X_n$ *iid* $N(\theta, \sigma^2)$, $\Theta = \{\theta : \theta \geq 0\}$, σ^2 known

$H_0 : \theta = 0$ vs. $H_1 : \theta > 0$, $p = 1 - \Phi\left(\sqrt{n}\,\bar{x} / \sigma\right)$

Thresholds

Aligning P-values and Bayes Factors

Jonathan Rougier

$X_1, \ldots, X_n$ *iid* $N(\theta, \sigma^2)$, $\Theta = \{\theta : \theta \geq 0\}$, σ^2 known

$H_0 : \theta = 0$ vs. $H_1 : \theta > 0$, $p = 1 - \Phi\left(\sqrt{n}\,\bar{x} / \sigma\right)$

$$\mathrm{BF}_{0:1} = \frac{f(x \mid \theta = 0)}{\int_0^\infty f(x \mid \theta) g(\theta) d\theta} \geq \frac{f(x \mid \theta = 0)}{f(x \mid \theta = \hat{\theta})} \quad , \quad \hat{\theta} = \max(0, \bar{x})$$

$$= \begin{cases} \exp(-n\bar{x}^2 / \sigma^2) & , \text{if } \bar{x} \geq 0 \\ 1 & , \text{if } \bar{x} < 0 \end{cases}$$

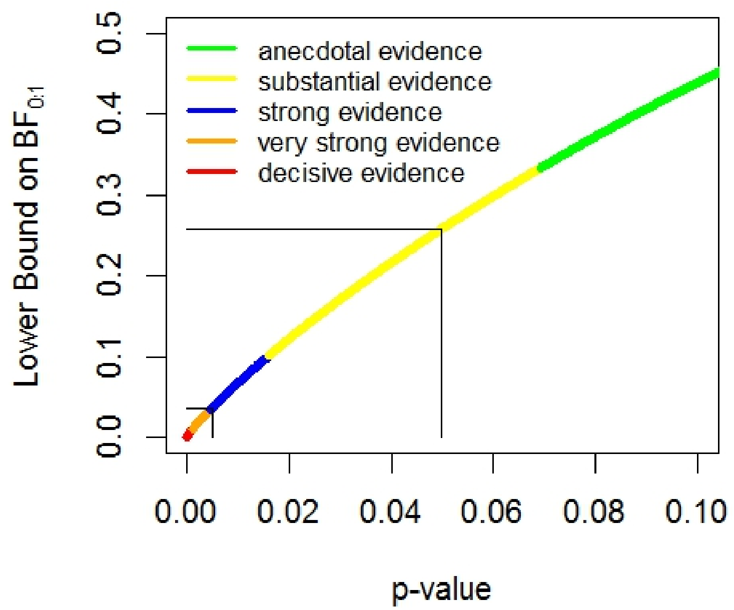
$$= \begin{cases} \exp\left\{-(1/2)[\Phi^{-1}(1-p)]^2\right\} & , \text{if } \bar{x} \geq 0 \\ 1 & , \text{if } \bar{x} < 0 \end{cases}$$

(see also, Edwards, *Psychological Review*, 1963)

Thresholds

Aligning P-values and Bayes Factors

Jonathan Rougier



p	Lower Bound on $BF_{0:1}$	"Jeffrey's Evidence" Against H_0	Lower Bound on $P(H_0\|x)$
.05	.259	(at best) there is substantial evidence	.205
.01	.067	(at best) there is strong evidence	.063
.005	.036	(at best) there is very strong evidence	.035

Interval Null

Second-Generation P-Values

Blume, Greevy, Welty, Smith, Dupont

Basic Idea

- Switch from point null to **interval null**
- A **descriptive statistic** that conveys the fraction of data-supported hypotheses that are null hypotheses
- Retain old characteristics of p-values (e.g., $0<p<1$) but add new characteristics such as an **ability to indicate when data supports the null**

Interval Null

Second-Generation P-Values

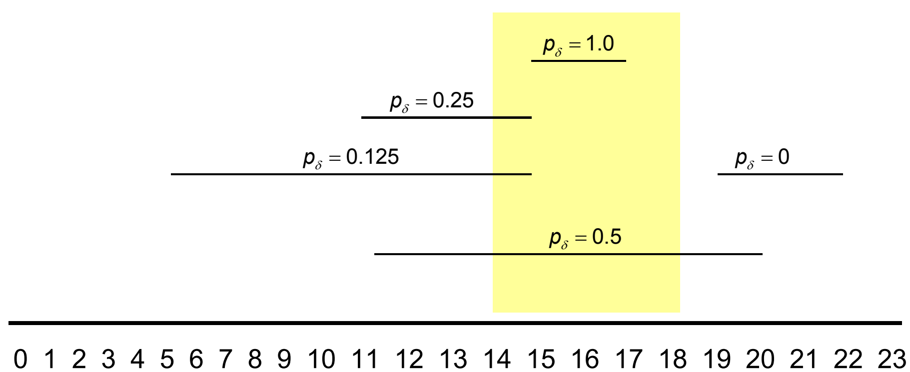
Blume, Greevy, Welty, Smith, Dupont

<u>Basic Idea</u>

- Switch from point null to **interval null**
- A **descriptive statistic** that conveys the fraction of data-supported hypotheses that are null hypotheses
- Retain old characteristics of p-values (e.g., 0<p<1) but add new characteristics such as an **ability to indicate when data supports the null**

Pose the null hypothesis as $\theta \in [a, b] \equiv H_0$. Let $I = [L, U]$ be a $100(1-\alpha)\%$ confidence interval for θ. Denote measure of overlap as $|I \cap H_0|$

$$p_\delta = \begin{cases} \dfrac{|I \cap H_0|}{|I|} & \text{, if } |I| < 2|H_0| \\[2ex] \dfrac{1}{2}\dfrac{|I \cap H_0|}{|H_0|} & \text{, if } |I| > 2|H_0| \end{cases}$$



$$P(p_\delta = 0 \mid H_0) \le \alpha$$

A proposed hybrid effect size plus p-value criterion

William Goodman, Susan Spruill, Eugene Komaroff

Interval Null

Basic Idea

- Switch from point null to **interval null**
- Decision criteria: **Reject null only if there is no overlap** between the interval null and a 95% confidence interval
- Another option: Reject null for cases where **p-value is smaller than .05 and the observed effect size is greater than a "minimum effect size** of interest"

A Close Relative

Dan's Comment

We know how to test an interval null hypothesis with a Union-Intersection Test

$X_1, \ldots, X_n$ *iid* $N(\mu, \sigma^2)$, σ^2 known. $H_0 : \mu \in [a, b]$ vs. $H_1 : \mu \notin [a, b]$

Reject if either $\dfrac{\bar{X} - b}{\sigma / \sqrt{n}} > z_{\alpha/2}$ or $\dfrac{\bar{X} - a}{\sigma / \sqrt{n}} < -z_{\alpha/2}$

There is even a UMPU test for this particular case (Schervish, TAS, 1996)

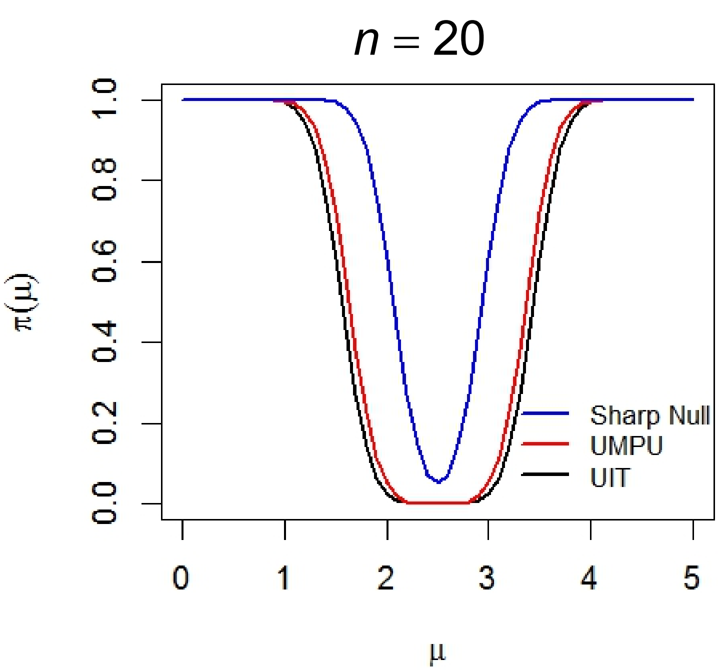
Reject if either $\left| \bar{X} - (a + b)/2 \right| > c$, choosing c to satisfy

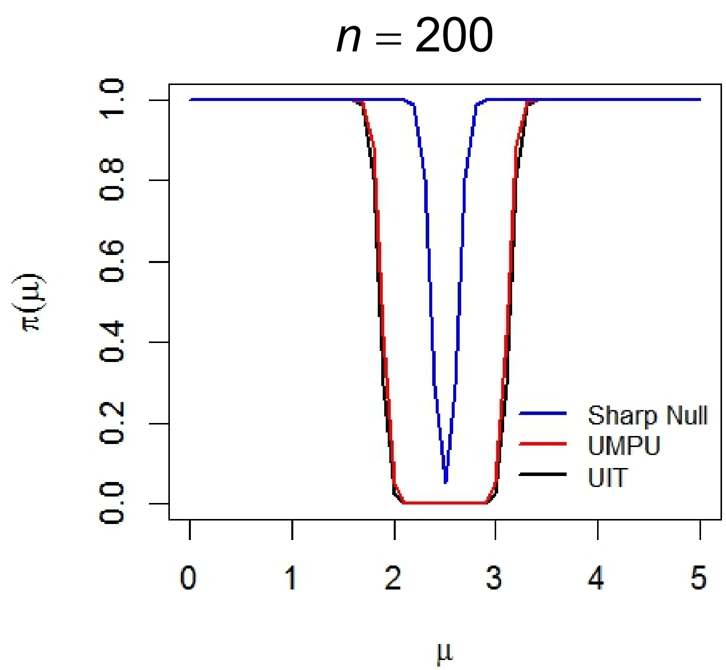
$$\Phi\left[\frac{(a-b)/2 - c}{\sigma / \sqrt{n}}\right] + \Phi\left[\frac{(b-a)/2 - c}{\sigma / \sqrt{n}}\right] = \alpha$$

Daniel R. Jeske, Department of Statistics, University of California, Riverside

A Close Relative

$$X_1, \ldots, X_n \text{ } iid \text{ } N(\mu, 1)$$
$$H_0: \mu \in [2,3] \text{ vs. } H_1: \mu \notin [2,3]$$
$$\alpha = .05$$





Daniel R. Jeske, Department of Statistics, University of California, Riverside

Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility

Robert Matthews

Reconcile

Basic Idea

- A data set is analyzed by a frequentist
- Find “the” priors that would lead to a Bayesian analysis that would support the frequentist analysis
- Assess the feasibility of those priors

Daniel R. Jeske, Department of Statistics, University of California, Riverside

Reconcile

Moving Towards the Post p < 0.05 Era via the Analysis of Credibility

Robert Matthews

Basic Idea

- A data set is analyzed by a frequentist
- Find "the" priors that would lead to a Bayesian analysis that would support the frequentist analysis
- Assess the feasibility of those priors

$X_1, \ldots, X_n$ *iid* $N(\mu, \phi)$, ϕ known

A 95% confidence interval is $\bar{X} \pm 1.96\sqrt{\phi / n} \equiv (L, U)$.

Suppose $0 \notin (L, U)$ so the frequentist declares a non-zero effect.

Daniel R. Jeske, Department of Statistics, University of California, Riverside

Reconcile

Moving Towards the Post p < 0.05 Era via the Analysis of Credibility

Robert Matthews

A skeptical Bayesian might use a prior like $\mu \sim N(0, \phi_0)$.

The resulting 95% credibility interval is

$$\left(\frac{n}{\phi} + \frac{1}{\phi_0}\right)^{-1} \frac{n\bar{X}}{\phi} \pm 1.96 \sqrt{\left(\frac{n}{\phi} + \frac{1}{\phi_0}\right)^{-1}}$$

Reconcile

Moving Towards the Post p < 0.05 Era via the Analysis of Credibility

Robert Matthews

A skeptical Bayesian might use a prior like $\mu \sim N(0, \phi_0)$.

The resulting 95% credibility interval is

$$\left(\frac{n}{\phi}+\frac{1}{\phi_0}\right)^{-1}\frac{n\bar{X}}{\phi} \pm 1.96\sqrt{\left(\frac{n}{\phi}+\frac{1}{\phi_0}\right)^{-1}}$$

Credibility interval **excludes 0** if and only if $\phi_0 \geq \dfrac{(U-L)^4}{1.96^2 \times LU}$.

If the skeptical Bayesian felt they had strong enough prior information that $\phi_0 < \dfrac{(U-L)^4}{1.96^2 \times LU}$ was reasonable they would dispute the frequentist finding.

Reconcile

Moving Towards the Post p < 0.05 Era via the Analysis of Credibility

Robert Matthews

Alternatively, suppose $0 \in (L, U)$ so that the frequentist declares no effect.

An advocating Bayesian might use a prior like $\mu \sim N(\mu_0, \phi_0)$ with the constraint $\mu_0 - 1.96\sqrt{\phi_0} = 0$.

Reconcile

Moving Towards the Post p < 0.05 Era via the Analysis of Credibility

Robert Matthews

Alternatively, suppose $0 \in (L, U)$ so that the frequentist declares no effect.

An advocating Bayesian might use a prior like $\mu \sim N(\mu_0, \phi_0)$ with the constraint $\mu_0 - 1.96\sqrt{\phi_0} = 0$.

Credibility interval **includes 0** if and only if $\phi_0 \geq \frac{(U+L)^2}{L^2U^2} \frac{(U-L)^2}{1.96^2 \times 16}$.

If the advocating Bayesian felt they had strong enough prior information that $\phi_0 < \frac{(U+L)^2}{L^2U^2} \frac{(U-L)^2}{1.96^2 \times 16}$ was reasonable they would dispute the frequentist finding.

Thank You!