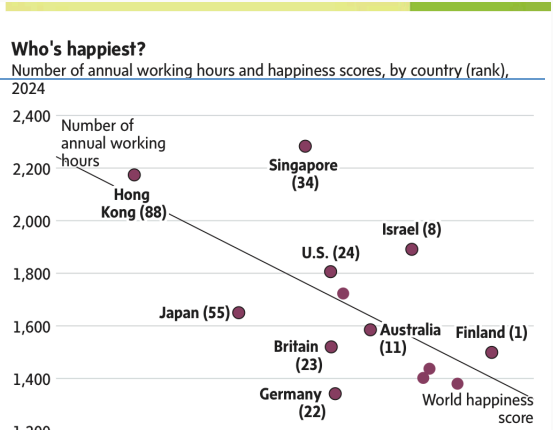
Statistical Theory for Data Science

STA2212H S LEC9101

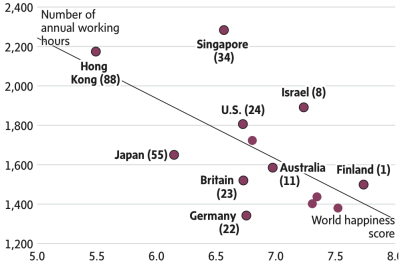
Week 6

February 10 2026



Who's happiest?

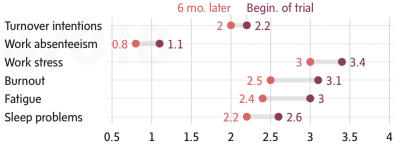
Number of annual working hours and happiness scores, by country (rank), 2024



THE GLOBE AND MAIL, SOURCE: THE CONFERENCE BOARD TOTAL ECONOMY DATABASE, MAY 2024; THE WORLD HAPPINESS REPORT, 2025

Effects of a four-day work week on health and well-being

Labour supply and work-related strain, scale 1-5



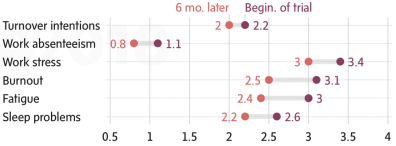
Work-related metrics based on employee surveys and interviews

THE GLOBE AND MAIL, SOURCE: "FOUR DAYS A WEEK," JULIET B. SCHOR, HARPER BUSINESS, 2025

THE GLOBE AND MAIL, SOURCE: THE CONFERENCE BOARD TOTAL ECONOMY DATABASE, MAY 2024; THE WORLD HAPPINESS REPORT, 2025

Effects of a four-day work week on health and well-being

Labour supply and work-related strain, scale 1-5

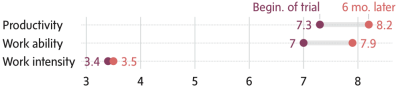


Work-related metrics based on employee surveys and interviews

THE GLOBE AND MAIL, SOURCE: "FOUR DAYS A WEEK," JULIET B. SCHOR, HARPER BUSINESS, 2025

Effects of a four-day work week on work

Work time, performance, and work design, scale 1-10



Work-related metrics based on employee surveys and interviews

THE GLOBE AND MAIL, SOURCE: "FOUR DAYS A WEEK," JULIET B. SCHOR, HARPER BUSINESS, 2025

Today

1. Midterm 2 March 10
2. Recap: Goodness-of-fit; diagnostic testing
3. Multiple testing
4. Reproducibility and replicability
5. Papers re project Google sheet

Recap: Goodness-of-fit

- multinomial tests:

$$W = 2\sum_{j=1}^{k} \log\left(\frac{Y_j}{np_j(\widetilde{\theta})}\right), \quad Q = \sum_{j=1}^{k} \frac{\{y_j - np_j(\widetilde{\theta})\}^2}{np_j(\widetilde{\theta})}$$

- $\widetilde{\theta}$ estimated from the binned data, not the MLE

- multinomial tests:

$$W = 2\sum_{j=1}^{k} \log\left(\frac{Y_j}{np_j(\widetilde{\theta})}\right), \quad Q = \sum_{j=1}^{k} \frac{\{y_j - np_j(\widetilde{\theta})\}^2}{np_j(\widetilde{\theta})}$$

- $\widetilde{\theta}$ estimated from the binned data, not the MLE

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 727–749
https://doi.org/10.1214/18-AOAS1155SF

HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL MULTINOMIALS: A SELECTIVE REVIEW[1]

BY SIVARAMAN BALAKRISHNAN AND LARRY WASSERMAN

Carnegie Mellon University

In memory of Stephen E. Fienberg

The statistical analysis of discrete data has been the subject of extensive statistical research dating back to the work of Pearson. In this survey we review some recently developed methods for testing hypotheses about high-dimensional multinomials. Traditional tests like the χ^2-test and the likelihood ratio test can have poor power in the high-dimensional setting. Much of the research in this area has focused on finding tests with asymptotically normal limits and developing (stringent) conditions under which tests have normal limits. We argue that this perspective suffers from a significant deficiency: it can exclude many high-dimensional cases when—despite having non-normal null distributions—carefully designed tests can have high power. Finally, we illustrate that taking a minimax perspective and considering refinements of this perspective can lead naturally to powerful and practical tests.

1. Introduction. Steve Fienberg was a pioneer in the development of theory and methods for discrete data. His textbook [Bishop, Fienberg and Holland (1977)] remains one of the main references for the topic. Our focus in this review is on high-dimensional multinomial models where the number of categories d can grow with, and possibly exceed the sample size n. Steve's paper [Fienberg and Holland

Recap: Goodness-of-fit

- smooth GoF tests: some measure of the distance between empirical cdf $\widehat{F}_n(\cdot)$ and the hypothesized cdf $F_0(\cdot)$

$$K_n = \sup_t |\widehat{F}_n(t) - F_0(t)|$$

$$W_n = \int \{\widehat{F}_n(t) - F_0(t)\}^2 dF_0(t)$$

$$A_n = \int \frac{\{\widehat{F}_n(t) - F_0(t)\}^2}{F_0(t)\{1 - F_0(t)\}} dF_0(t)$$

  if $F_0(\cdot)$ is known, then $Y_i = F_0(X_i) \sim U(0,1)$ so we only need to consider testing uniform distribution

- see MS 9.3 for hints about limiting distribution s of K_n, W_n, A_n Brownian bridge

Recap: Goodness-of-fit

- definitive reference DasGupta, A. (2008)[1]
- typically $F_0(\cdot;\theta)$ has some unknown parameters

Theorem 28.1 Suppose $X_1, \ldots, X_n$ are iid F_{θ_0} for some $\theta_0 \in \Theta$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be $\sqrt{n}$-consistent for θ_0 and be asymptotically linear. Then each of $n\widetilde{C}_n$ and $n\widetilde{A}_n$ converges in law to the distribution of $\sum_{i=1}^{\infty} \lambda_i W_i$ for suitable $\{\lambda_i\}$, where $W_1, W_2, \ldots$ are iid χ^2_1. The coefficients $\{\lambda_i\}$ are the eigenvalues corresponding to the eigenfunctions $\{f_i\}$ of the covariance kernel $\rho(s,t)$ of a suitable zero-mean Gaussian process $X(t)$; i.e., λ_i satisfies

$$\lambda_i f_i(s) = \int \rho(s,t) f_i(t) dt.$$

Remark. See del Bario, Deheuvels, and van de Geer (2007) for a proof. The $\{\lambda_i\}$ can be found numerically by solving certain partial differential equations.

[1] *Asymptotic Theory of Statistics and Probability*, Springer. Chapter 26 Chapter 28

Distribution-free two-sample testing with blurred total variation distance

Rohan Hore*[1] and Rina Foygel Barber[2]

[1]Department of Statistics and Data Science, Carnegie Mellon University
[2]Department of Statistics, University of Chicago

February 6, 2026

Abstract

Two-sample testing, where we aim to determine whether two distributions are equal or not equal based on samples from each one, is challenging if we cannot place assumptions on the properties of the two distributions. In particular, certifying equality of distributions, or even providing a tight upper bound on the total variation (TV) distance between the distributions, is impossible to achieve in a distribution-free regime. In this work, we examine the blurred TV distance, a relaxation of TV distance that enables us to perform inference without assumptions on the distributions. We provide theoretical guarantees for distribution-free upper and lower bounds on the blurred TV distance, and examine its properties in high dimensions.

A note on two-sided p-values

- a large goodness-of-fit statistic suggests proposed model fits poorly
- a very small value could suggest model fits too well
- null distribution is not symmetric

so using $|T|$ or T^2 not recommended

- $p^+_{obs} = \mathrm{pr}(T \geq t_{obs}), \quad p^-_{obs} = \mathrm{pr}(T \leq t_{obs})$

$p^+_{obs} + p^-_{obs} = 1 + \mathrm{pr}(T = t_{obs})$

- Define $P = min(P^+, P^-)$

- p-value for a two-sided test is $2\min(p^+_{obs}, p^-_{obs})$

$$\mathrm{pr}_0\{P \leq \min(p^+_{obs}, p^-_{obs})\} \sim$$

test is exact, dist'n of T is conts

Diagnostic testing

Wikipedia

1. Hypothesis testing

AoS Table 10.1

truth	H_0 not rejected	H_0 rejected
H_0 true		type 1 error
H_1 true	type 2 error	

2. Diagnostic testing

link

truth	test negative	test positive	
C19 neg	TN	FP	N
C19 pos	FN	TP	P

1. Hypothesis testing

AoS Table 10.1

		H_0 not rejected	H_0 rejected
truth	H_0 true		type 1 error
	H_1 true	type 2 error	

3. Multiple testing

AoS Table 10.2

		H_0 not rejected	H_0 rejected	
truth	H_0 true	U	V	m_0
	H_1 true	T	S	m_1
		$m - R$	R	m

FDP, FDR

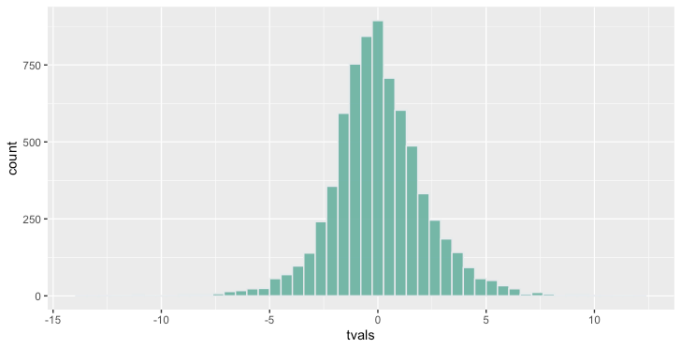
```
leukemia_big <- read.csv
  ("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
  dim(leukemia_big)
  [1] 7128   72
```

each row is a different gene; 47 AML responses and 25 ALL responses
we could compute 7128 t-statistics for the mean difference between AML and ALL

```
tvals <- rep(0,7128)
for (i in 1:7128){
  leukemia_big[i,] %>% select(starts_with("ALL")) %>% as.numeric() -> x
  leukemia_big[i,] %>% select(starts_with("AML")) %>% as.numeric() -> y
  tvals[i] <- t.test(x,y,var.equal=T)$statistic
  }
```

Multiple testing

EH 1.2, 15.2




```
summary(tvals)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-13.52611  -1.20672  -0.08406   0.02308   1.20886  12.26065
```

Multiple testing

- H_{0i} versus H_{1i}, $\quad i = 1, \ldots, m$
- p-values $p_1, \ldots, p_m$
- Bonferroni method: reject H_{0i} if $p_i < \alpha/m$ — FWER
- $\mathrm{pr}(\text{any } H_0 \text{ falsely rejected }) \leq \alpha$ — very conservative

Multiple testing

- H_{0i} versus H_{1i}, $\quad i = 1, \ldots, m$
- p-values $p_1, \ldots, p_m$
- Bonferroni method: reject H_{0i} if $p_i < \alpha/m$ — FWER
- $\mathrm{pr}(\text{any } H_0 \text{ falsely rejected }) \leq \alpha$ — very conservative

- FDR method controls the number of rejections that are false — FDP = V/R; $0/0 \equiv 0$

		H_0 not rejected	H_0 rejected	
truth	H_0 true	U	V	m_0
	H_1 true	T	S	m_1
		$m - R$	R	m

$FDR = \mathrm{E}(FDP)$

- order the p-values $p_{(1)}, \ldots, p_{(m)}$
- find i_{max}, the largest index for which

$$p_{(i)} \leq \frac{i}{m} q$$

- Let BH_q be the rule that rejects H_{oi} for $i \leq i_{max}$, not rejecting otherwise

- order the p-values $p_{(1)}, \ldots, p_{(m)}$
- find i_{max}, the largest index for which

$$p_{(i)} \leq \frac{i}{m} q$$

- Let BH_q be the rule that rejects H_{oi} for $i \leq i_{max}$, not rejecting otherwise

- Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_o q \leq q, \qquad \text{where } \pi_o = m_o/m$$

π_o unknown but close to 1

- change the bound under dependence

$$p_{(i)} \leq \frac{i}{m C_m} q \qquad C_m = \sum_{i=1}^{m} \frac{1}{i}$$

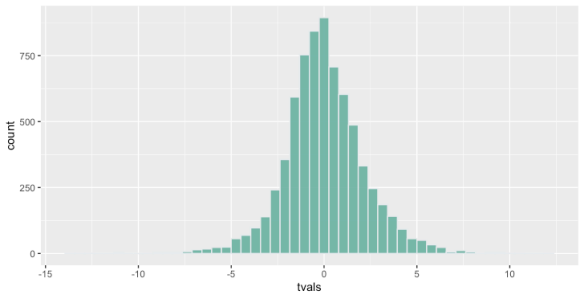
Example

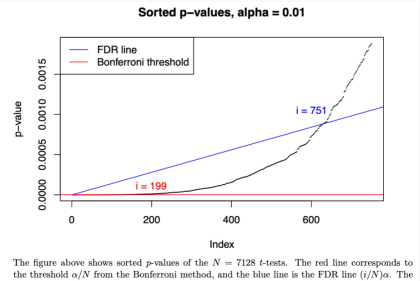
AoS Ex.10.28

index	1	2	3	4	5	6	7	8	9	10
pval	0.00017	0.00448	0.00671	0.00907	0.01220	0.33626	0.3934	0.5388	0.5813	0.9862
cut1	0.00500	0.01000	0.01500	0.02000	0.02500	0.03000	0.0350	0.0400	0.0450	0.0500
cut2	0.00171	0.00341	0.00512	0.00683	0.00854	0.01024	0.0119	0.0137	0.0154	0.0171

Leukemia example

EH 1.2, 15.2



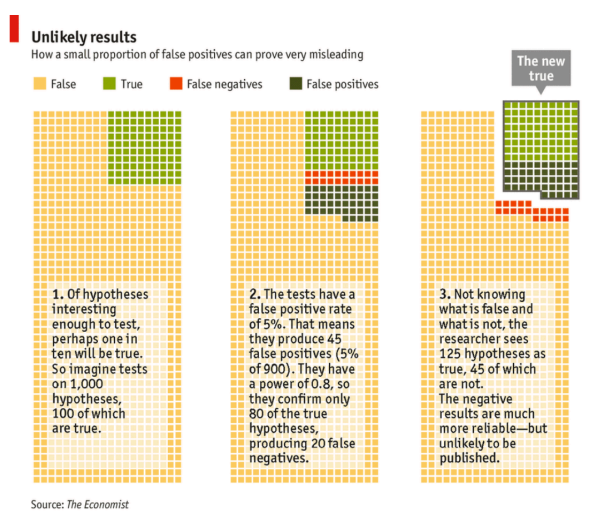


The figure above shows sorted p-values of the $N = 7128$ t-tests. The red line corresponds to the threshold α/N from the Bonferroni method, and the blue line is the FDR line $(i/N)\alpha$. The

```
> summary(ttest)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-13.52611  -1.20672  -0.08406   0.02308   1.20886  12.26065
```

prostate example: Efron LSI p.44

Multiple testing



Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

The Annals of Statistics
2006, Vol. 34, No. 4, 1827–1849
DOI: 10.1214/009053606000000425

ON THE BENJAMINI–HOCHBERG METHOD

BY J. A. FERREIRA[1] AND A. H. ZWINDERMAN

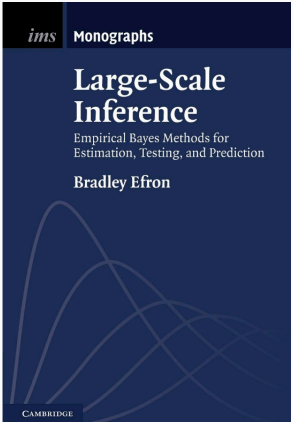
University of Amsterdam

We investigate the properties of the Benjamini–Hochberg method for multiple testing and of a variant of Storey's generalization of it, extending and complementing the asymptotic and exact results available in the literature. Results are obtained under two different sets of assumptions and include asymptotic and exact expressions and bounds for the proportion of rejections, the proportion of incorrect rejections out of all rejections and two other proportions used to quantify the efficacy of the method.

1. Introduction. Let $X = \{X_1, X_2, \ldots, X_m\}$ be a set of m random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ such that, for some positive integer $m_0 \leq m$, each of $X_1, X_2, \ldots, X_{m_0}$ has distribution function (d.f.) F and $X_{m_0+1}, \ldots, X_m$ all have d.f.'s different from F, and consider the problem of choosing a set $\mathcal{R} \subseteq X$ in such a way that the random variable (r.v.)

$$\Pi_{1,m} = \frac{S_m}{R_m \vee 1},$$

where $R_m = \#\mathcal{R}$ and $S_m = \#(\mathcal{R} \cap \{X_1, \ldots, X_{m_0}\})$, is guaranteed to be small in some probabilistic sense. In more ordinary language, the problem is that of discovering observations in X which do not have d.f. F without incurring a high



Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

		H_0 not rejected	H_0 rejected	
truth	H_0 true	U	V	m_0
	H_1 true	T	S	m_1
		$m - R$	R	m

- $p_1, \ldots, p_m$ p-values in $[0, 1]$, $\quad p_{(1)} < p_{(2)} < \cdots < p_{(m)}$ create a process in t
- Define $R(t) \equiv \#\{p_i \leq t\}, 0 < t \leq 1$ and $a(t) \equiv \#\{p_i \leq t \text{ and } H_0\}$

"number of null cases with $p_i \leq t$"

Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

truth	H_0 not rejected	H_0 rejected	
H_0 true	U	V	m_0
H_1 true	T	S	m_1
	$m - R$	R	m

- $p_1, \ldots, p_m$ p-values in $[0,1]$, $\quad p_{(1)} < p_{(2)} < \cdots < p_{(m)}$ — create a process in t
- Define $R(t) \equiv \#\{p_i \leq t\}, 0 < t \leq 1$ and $a(t) \equiv \#\{p_i \leq t \text{ and } H_0\}$

"number of null cases with $p_i \leq t$"

- false discovery proportion $FDP(t) = a(t)/\max\{R(t), 1\}$
- define $Q(t) = mt/\max\{R(t), 1\}, \quad t_q = \sup_t\{Q(t) \leq q\}$

Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

		H_0 not rejected	H_0 rejected	
truth	H_0 true	U	V	m_0
	H_1 true	T	S	m_1
		$m - R$	R	m

- $p_1, \ldots, p_m$ p-values in $[0, 1]$, $\quad p_{(1)} < p_{(2)} < \cdots < p_{(m)}$ — create a process in t
- Define $R(t) \equiv \#\{p_i \leq t\}, 0 < t \leq 1$ and $a(t) \equiv \#\{p_i \leq t \text{ and } H_0\}$

"number of null cases with $p_i \leq t$"

- false discovery proportion $FDP(t) = a(t)/\max\{R(t), 1\}$
- define $Q(t) = mt/\max\{R(t), 1\}, \quad t_q = \sup_t\{Q(t) \leq q\}$
- $R(p_i) = i \implies Q(p_{(i)}) = mp_{(i)}/i \implies$ BH-rule is

$$\text{Reject } H_{0(i)} \text{ for } p_{(i)} \leq t_q$$

Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

- $R(p_i) = i \implies Q(p_{(i)}) = mp_{(i)}/i \implies$ BH-rule is

$$\text{Reject } H_{0(i)} \text{ for } p_{(i)} \leq t_q$$

- With $A(t) = a(t)/t$, $\quad \mathrm{E}\{A(s) \mid A(t)\} = A(t), s \leq t$ martingale as $t \downarrow 0$

$$\mathrm{E}\{A(t_q)\} = E\{A(1)\} = m_0$$

Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

- $R(p_i) = i \implies Q(p_{(i)}) = mp_{(i)}/i \implies$ BH-rule is

$$\text{Reject } H_{0(i)} \text{ for } p_{(i)} \leq t_q$$

- With $A(t) = a(t)/t$, $\mathrm{E}\{A(s) \mid A(t)\} = A(t), s \leq t$ martingale as $t \downarrow 0$

$$\mathrm{E}\{A(t_q)\} = E\{A(1)\} = m_0$$

- $FDP(t_q) = \dfrac{q}{m}\dfrac{a(t_q)}{t_q}, \qquad \mathrm{E}\{FDP(t_q)\} = \pi_0 q \equiv \dfrac{m_0}{m} q$

Theorem: If the p-values corresponding to valid null hypotheses are independent of each other, then

$$FDR(BH_q) = \pi_0 q \leq q, \qquad \text{where } \pi_0 = m_0/m$$

1. For $p_1, \ldots, p_n$ assumed i.i.d. $U(0,1)$, define $U = \sum_{i=1}^{n} I\{p_i \leq s\}, \quad V = \sum_{i=1}^{n} I\{s < p_i < t\}$
2. Show that $\mathrm{pr}(U = j \mid U + V = k) = \binom{k}{j}(s/t)^k(1 - s/t)^{k-j}$
3. Conclude

$$\mathrm{E}\{a(s) \mid a(t)\} = \frac{s}{t}a(t), \quad s \leq t$$

Distributions for Parameters

Nancy Reid
University of Toronto

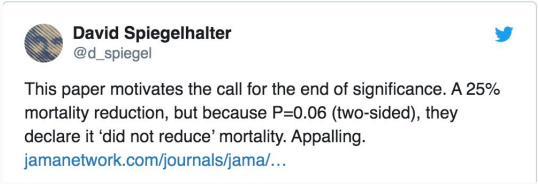
Monash University

Feb 21 2020

Reproducibility and Statistical theory

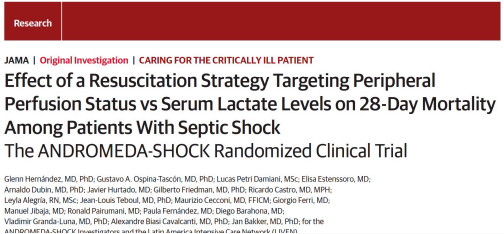
Those pesky *p*-values

medium.com



… those pesky *p*-values

JAMA Feb 2019



Research

JAMA | Original Investigation | CARING FOR THE CRITICALLY ILL PATIENT

Effect of a Resuscitation Strategy Targeting Peripheral Perfusion Status vs Serum Lactate Levels on 28-Day Mortality Among Patients With Septic Shock

The ANDROMEDA-SHOCK Randomized Clinical Trial

Glenn Hernández, MD, PhD; Gustavo A. Ospina-Tascón, MD, PhD; Lucas Petri Damiani, MSc; Elisa Estenssoro, MD; Arnaldo Dubin, MD, PhD; Javier Hurtado, MD; Gilberto Friedman, MD, PhD; Ricardo Castro, MD, MPH; Leyla Alegría, RN, MSc; Jean-Louis Teboul, MD, PhD; Maurizio Cecconi, MD, FFICM; Giorgio Ferri, MD; Manuel Jibaja, MD; Ronald Pairumani, MD; Paula Fernández, MD; Diego Barahona, MD; Vladimir Granda-Luna, MD, PhD; Alexandre Biasi Cavalcanti, MD, PhD; Jan Bakker, MD, PhD; for the ANDROMEDA-SHOCK Investigators and the Latin America Intensive Care Network (LIVEN)

Andromeda Trial

Hernández et al. 2019

- comparing two treatments for septic shock
- randomized clinical trial

- estimated hazard ratio 0.75 [0.55, 1.02] — after adjusting for confounders
- 2-sided p-value 0.06 — 34.9% vs 43.4% unadjusted

- Discussion: " a peripheral perfusion-targeted resuscitation strategy did not result in a significantly lower 28-day mortality when compared with a lactate level-targeted strategy"
- Abstract: "Among patients with septic shock, a resuscitation strategy targeting normalization of capillary refill time, compared with a strategy targeting serum lactate levels, did not reduce all-cause 28-day mortality."

A recent timeline

- 2014: *Basic and Applied Social Psychology* published an editorial banning *p*-values

  actually "null hypothesis significance testing"

- "prior to publication, authors will need to remove all vestiges of the NHSTP ... *p*-values, ... , statements about 'significant differences' or lack thereof, and so on"

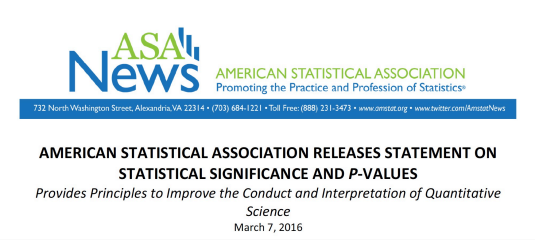
  "confidence intervals are also banned"

- 2014: *Nature* published a News Feature by R. Nuzzo: "*p*-values, the gold standard of statistical validity, are not as reliable as many scientists assume"

- 2016: American Statistical Association released a public statement on statistical significance and *p*-values

… a recent timeline

2016



- 2017: Another *Nature* article $p < 0.005$

comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

- Articles solicited for special issue of *American Statistician*

- 2019: *American Statistician* publishes special issue 43 articles; 400 pages
- Editorial introduction advises "abandon 'statistical significance' "

- *Nature* publishes a letter agreeing with this
- "we are not advocating a ban on P values, confidence intervals or other statistical measures – only that we should not treat them categorically
- "This includes dichotomization as statistically significant or not, as well as categorization based on other statistical measures such as Bayes factors."

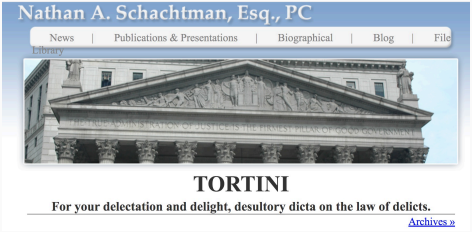

COMMENT

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Later that year

2019



"Lawyers and judges pay close attention to standards, guidances, and consenus statements from respected and recognized professional organizations."

"Despite the fairly clear and careful statement of principles, legal actors did not take long to misrepresent the ASA principles."

2016

"distorted into strident assertions that statistical significance was unnecessary for scientific conclusions."

outlines a 2018 Supreme Court case appealing a conviction for wire fraud, based on misleading investors

Harkonen v. United States 13-180

the fraud centered on *p*-hacking the results of a Phase III trial of a drug marketed by Harkonen

in the appeal "his defenders argued that the ASA guide provides compelling new evidence that the scientific theory upon which petitioner's conviction was based [that of statistical significance testing] is demonstrably false"

What to do?

- report actual p-value, not "*", $p < 0.05$, etc. — to sensible number of decimal points
- supplement p-value with sample size, estimated power, etc.
- clarify 'exploratory' and 'confirmatory' p-values — Spiegelhalter 2017

- report effect sizes and estimated standard errors
- report confidence intervals

- pre-register trials, specifying primary and secondary outcomes
- pre-specify data analysis — NEJM

- provide a p-value function — significance function
- or some analogous distribution — Bayes posterior

Also recommended

Harvard Data Science Review • Issue 2.4, Fall 2020

Selective Inference: The Silent Killer of Replicability

Yoav Benjamini[1]

[1]**Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel**

The MIT Press

Published on: Dec 16, 2020

DOI: https://doi.org/10.1162/99608f92.fc62b261