Statistical ecotoxicology: A crisis of confidence, or, why does it hurt when I P (<0·005)?

David R. Fox
Environmetrics Australia, University of Melbourne
Ross Smith
Hydrobiology Australia
Wayne G. Landis
Western Washington University

Or given recent events …

Statistics in ecotoxicology: A crisis of confidence, or, why does it hurt when I P (<0·005)?

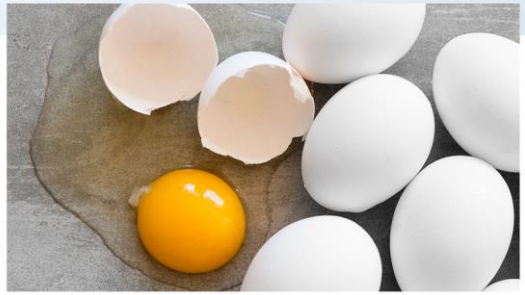
David R. Fox
Environmetrics Australia, University of Melbourne
Ross Smith
Hydrobiology Australia
Wayne G. Landis
Western Washington University

Drivers of change

The 'reproducibility crisis' and scientific flip-flopping

BLOG | Mar 16, 2019

New Study Finds Eggs Will Break Your Heart



Americans are eating 279 eggs per person a year, and a new study finds it's **killing** them.

Will eating eggs break your heart?

Eggs Not Harmful for Heart Health

A new study supports research that suggests eggs are not linked to cardiovascular disease.

By **Tala Salem**, Staff Writer May 7, 2018, at 4:44 p.m.

Not only are eggs "not bad" for cholesterol, but they are also good for weight loss, a new study finds. (GETTY IMAGES/ISTOCKPHOTO)

EATING UP TO A DOZEN eggs a week does not increase the risk of heart disease, according to a new study.

MORE HEALTH CARE NEWS

CIVIC
Healthcare of Tomorrow

NATIONAL NEWS
New Health Care Index Shows Increased Costs

nature > comment > article

a natureresearch journal

MENU

nature
International Journal of science

Subscribe Search Login

COMMENT • 20 MARCH 2019

Scientists rise up against 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.

Valentin Amrhein, Sander Greenland & Blake McShane

Statisticians Weigh into the debate (albeit in a timid way)

nature International weekly journal of science

Search Go Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 531 > Issue 7593 > News > Article

NATURE | NEWS

E-alert RSS Facebook Twitter

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

07 March 2016

PDF Rights & Permissions

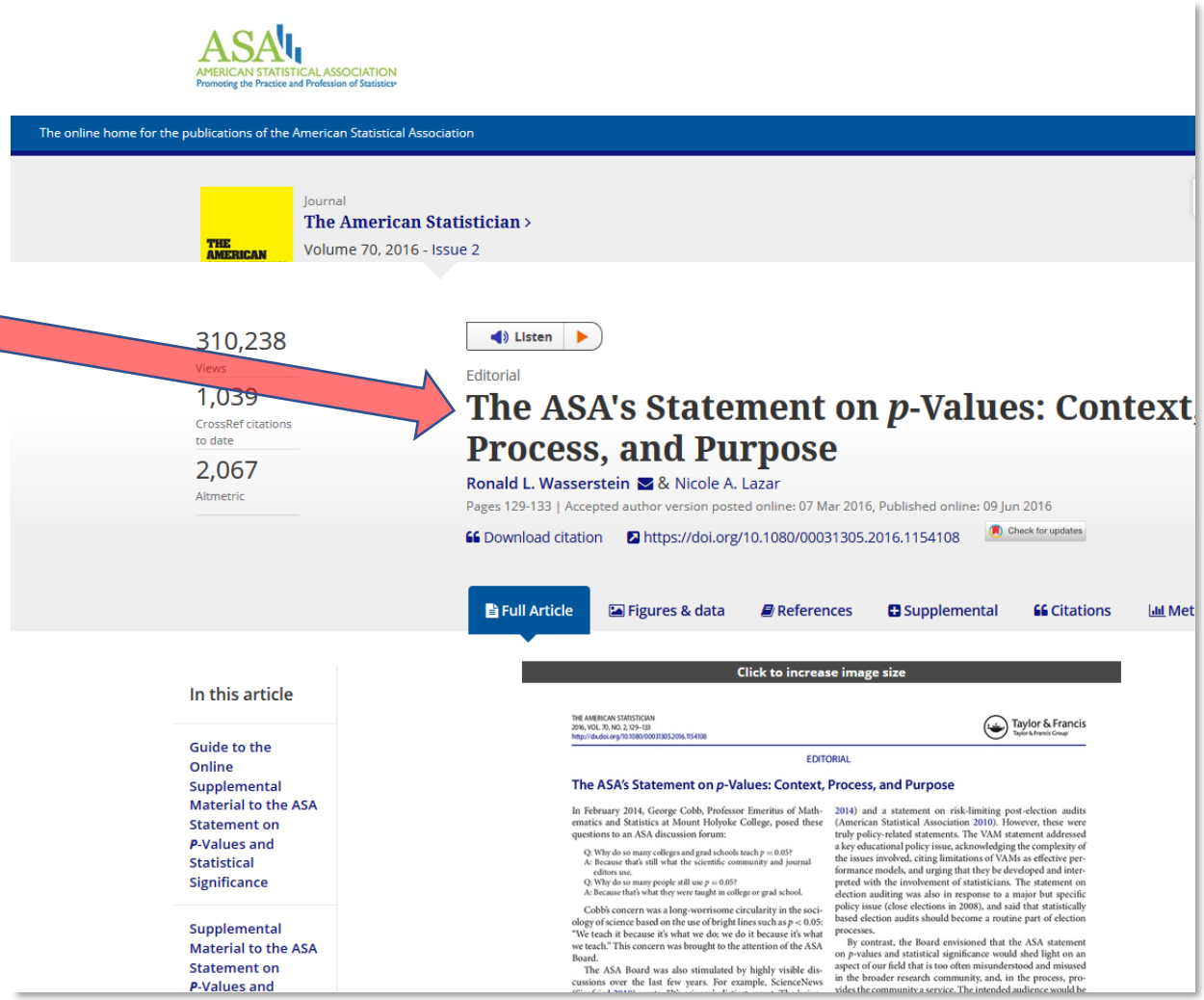
Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today[1]. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied in ways that cast doubt on statistics generally, he adds.

How scientists fool themselves – and how they can stop

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions based on *P* values alone. Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

Véronique Kiermer, executive editor of the Public Library of Science journals, says that the ASA's statement lends weight and visibility to longstanding concerns over undue reliance on the *P* value. "It is also very important in that it shows statisticians, as a profession, engaging with the problems in the literature outside of their field," she adds.



ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics

The online home for the publications of the American Statistical Association

THE AMERICAN

Journal
The American Statistician ›
Volume 70, 2016 - Issue 2

310,238 Views
1,039 CrossRef citations to date
2,067 Altmetric

Listen

Editorial

The ASA's Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online: 09 Jun 2016

Download citation https://doi.org/10.1080/00031305.2016.1154108 Check for updates

Full Article | Figures & data | References | Supplemental | Citations | Met

In this article

Guide to the Online Supplemental Material to the ASA Statement on *P*-Values and Statistical Significance

Supplemental Material to the ASA Statement on *P*-Values and

Click to increase image size

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

Q: Why do so many colleges and grad schools teach $p = 0.05$?
A: Because that's still what the scientific community and journal editors use.
Q: Why do so many people still use $p = 0.05$?
A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provides the community a service. The intended audience would be

854 Scientists sign letter to Nature

Now, academics are striking back against the tyranny of this threshold. More than 850 have signed a letter to the journal Nature arguing for "the entire concept of statistical significance to be abandoned". Whether a result refutes or supports a scientific hypothesis, they say, goes beyond an arbitrary cut-off.

854 Scientists sign letter to Nature

Now, academics are striking back against the tyranny of this threshold. More than 850 have signed a letter to the journal Nature arguing for "the entire concept of statistical significance to be abandoned". Whether a result refutes or supports a scientific hypothesis, they say, goes beyond an arbitrary cut-off.

... including one prominent statistician - David Speigelhalter

Signatories of the Nature letter include Cambridge university statistician David Spiegelhalter. The problem is not the p-values themselves, he says, but the "nonsensical reduction of science to the simplistic labelling of pass or fail".

there are not plenty of powerful approaches to improved inference.

To address the statement's shortcomings, the ASA is convening a Symposium on Statistical Inference this October. The tagline for the symposium is "Scientific Method for the 21st Century: A World Beyond $p < 0.05$". Discussions will centre on specific approaches for improving statistical practice as it intersects with three broad components of research activities: conducting research; using research; and sponsoring, disseminating, and replicating research. The vision of the symposium is to push change forward, change that leads to lasting improvements in research, in communicating and understanding uncertainty, and ultimately in decision-making.

We cannot accomplish this simply by having presentations at a conference. Instead, we envision teams of symposium delegates developing papers, briefs, practice guides, and statements on a wide variety of topics to help researchers, research sponsors, journal editors and referees, regulators, educators, the media, and policy- and other decision-makers.

If the symposium is successful in doing *some* of this, research will benefit. Yet if the symposium is successful at *all* of this, we will not really have achieved success until we have not only identified for researchers a rich variety of inferential methods and the situations in which they should be applied, but also ensured that these methods are being taught wherever researchers are being trained. ■

■ **Ron Wasserstein** is executive director of the American Statistical Association

"Too familiar to ditch"

By David Spiegelhalter

I have a confession to make. I like *p*-values. I like skimming regression output or large tables for those twinkling stars (and mentally checking if the proportion is any more than I would expect from chance alone). And I also like a single carefully adjusted *p*-value that helps summarise an entire experimental programme, such as the "five sigma" ($p = 1$ in 3.5 million) attached to the Higgs boson. As the first point of the 2016 ASA statement says, *p*-values can be useful summaries of the compatibility between data and hypotheses.

Concern about *p*-values is being driven by claims of a "reproducibility crisis". But how much are *p*-values to blame for this situation? Among the fine commentaries accompanying the ASA statement, many point out that the problem lies not so much with *p*-values in themselves as with the willingness of researchers to lurch casually from descriptions of data taken from poorly designed studies, to confident generalisable inferences. The ASA critique is great, but what is to be done about this issue that any half-decent statistician knows so well?

It should be possible to establish firm general principles which focus on what is right rather than what is wrong

Robert Matthews appropriately calls for "authoritative guidance on dealing with standard inferential problems encountered in each discipline", although I do wonder how this guidance is to be produced when there are so many different opinions among "authorities". He then argues that "significance testing has no place in such guidance, except to illustrate its pitfalls", and if by this he means all use of *p*-values, then I am afraid I must disagree. *p*-values are just too familiar and useful to ditch (even if it were possible).

But we can agree on scepticism about formal or informal rules that mechanically dichotomise findings into "significant" and "non-significant", and which can apply equally to rigid interpretation of intervals. Fortunately, Neyman and Pearson's decision-theoretic idea of "accepting the null" has just about been consigned to the overflowing dustbin of inappropriate scientific ideas, even if it lingers on in the misinterpretation of a "non-significant" result. Could we ditch "significant" as a similar anachronism? Sadly I think not, due to the habit of use and the lack of an alternative (apart from anything else, it would mean renaming this magazine). So, what are we left with? I have some personal opinions.

While there is not one universal solution, I believe it should be possible to establish firm general principles which focus on what is right rather than what is wrong. Then more specific guidance for different disciplines, to be enshrined in revised statistical education and statistical guidelines for journals and other outlets.

The crucial issue, identified by Berry, Gelman, Few and other commentators on the ASA statement, is to try and clearly separate (a) data description, (b) what it might be reasonable to believe in the light of this new evidence, and (c) categorical decisions and recommendations. *p*-values can have a role, although not be the sole determinant, at all stages. In particular, when describing data at stage (a) it may be fine to litter a results section with exploratory *p*-values, but these should not appear in the conclusions or abstract unless clearly labelled as such – perhaps by a specific notation p_{exp}.

A *p*-value should only be considered part of a confirmatory analysis at stage (b), and perhaps given the notation p_{con}, if the analysis has been pre-specified, all results reported, and *p*-values adjusted for multiple comparisons, and so on. Any p_{con}-values should be supplemented by informal, and even formal, Bayesian analysis that takes into account what else is known, the context, and in particular whether the null hypothesis or values close to it has any particular salience or plausibility, in which case Bayes factor arguments can be used to show the weakness of $p_{con} < 0.05$ and the need for higher thresholds.

But even if some agreement could be reached on a "positive" statement, then there is the problem of promulgating and enforcing it. At this point I get rather authoritarian. I believe that drawing unjustified conclusions based on selected exploratory *p*-values should be considered as scientific misconduct and lead to retraction or correction of papers. This requires both encouragement and training, but also publicly calling out journals, press offices and authors.

A colleague once told me of being confronted by a doctor at 4 pm on a Friday with "Could you just '*t* and *p*' this data for Monday?" While it would be wonderful if every analysis was going to be informed by someone skilled in statistical methodology, whether a nominal "statistician" or not, the rise of data science means that even more practitioners will be without a formal professional training, and continue to do their thing anyway. We must do our best to help them. ■

■ **David Spiegelhalter** is chair of the Winton Centre for Risk and Evidence Communication at the University of Cambridge. He is currently president of the Royal Statistical Society, although this article is written in his personal capacity

"Concern about *p*-values is being driven by claims of a 'reproducibility crisis'. But how much are *p*-values to blame for this situation?"

"I am afraid I must disagree. *P*-values are just too familiar and useful to ditch (even if it were possible)

"I have a confession to make. I like *p*-values"

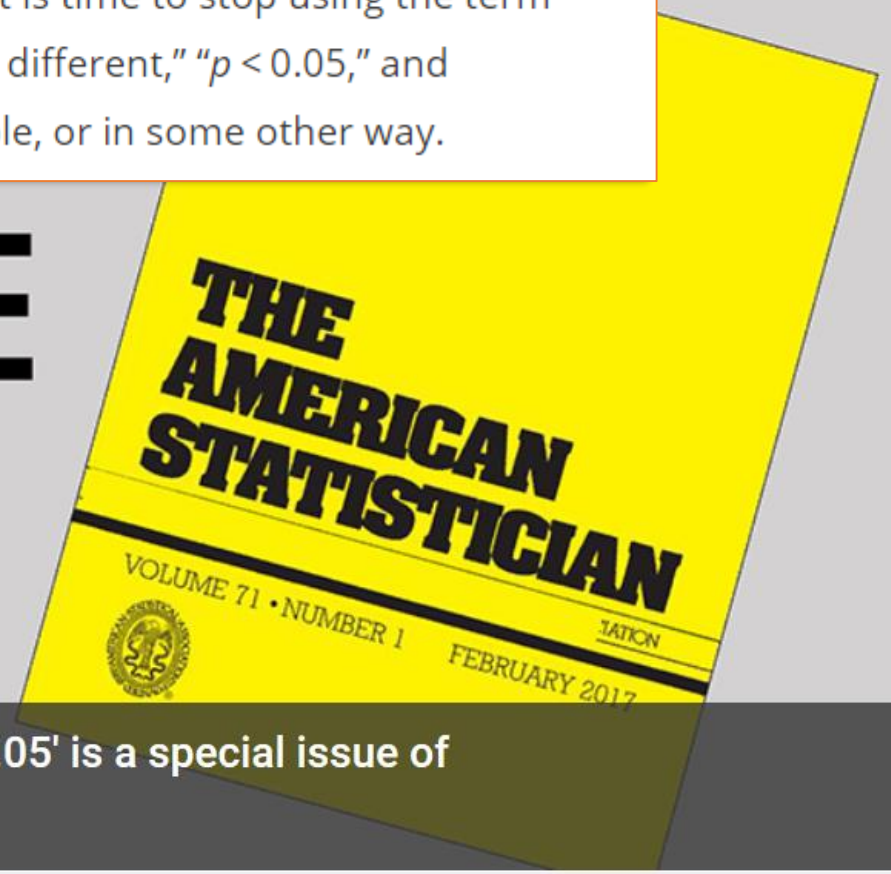
The Spiegelhalter backflip

"Ban Statistical Significance"

2 Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

SPECIAL ISSUE ON P-VALUE



Statistical Inference in the 21st Century: A World Beyond 'p<0.05' is a special issue of *The American Statistician*.

A good tradesman never blames his tools

The majority of documented ‘problems’ with NHST arise from

- Inappropriate use
- Lack of understanding
- Poor training
- Deliberate manipulation
- Confusion
- Misunderstanding
- Incorrect interpretation

These are <u>all</u> shortcomings of the end-user and NOT NHST. However these human failures have been used to malign a statistical methodology “that is now purported to suffer from ‘problems’ and ‘fatal flaws’ and criticised for not allowing the type of inferences that researchers seek” Garcia-Perez (2017)

A good tradesman never blames his tools

The majority of documented 'problems' with NHST arise from

- Inappropriate use
- Lack of understanding
- Poor training
- Deliberate manipulation
- Confusion
- Misunderstanding
- Incorrect interpretation

"After reading a year's worth of BASP [Basic and Applied Social Psychology] articles, you'd almost start to suspect p-values are not the real problem. Instead, it looks like researchers find making statistical inferences pretty difficult, and forcing them to ignore p-values didn't magically make things better".

Lakens, D. (2016) http://daniellakens.blogspot.com/2016/02/so-you-banned-p-values-hows-that.html

The dilemma of the dichotomy

Criticism: The $p < \alpha$ | $p \geq \alpha$ dichotomy results in binary decision-making and, according to Hurlbert, Levine, and Utts (2019):

> *"Situations requiring binary decisions solely on the basis of individual p-values are vanishingly rare in both basic and applied research".*

Response:

- We believe this is not only false, but importantly, is being used as a reason to avoid the stark reality and inconvenient truth that *when* a binary decision must be made, there is no alternative to weighing up the evidence (by whatever means, processes, and metrics) and making a choice.

- That choice will no doubt utilise the concept of 'beyond reasonable doubt' (or some variant) and most likely be based on a metric (*p*-value; Bayes Factor; some other information theoretic measure).

- The destination is the same, irrespective of the path taken – the researcher concludes yes/no; accept/reject; same/better; toxic/not toxic; exists/doesn't exist; extinct/not-extinct; impacted/not impacted; complies/doesn't comply …

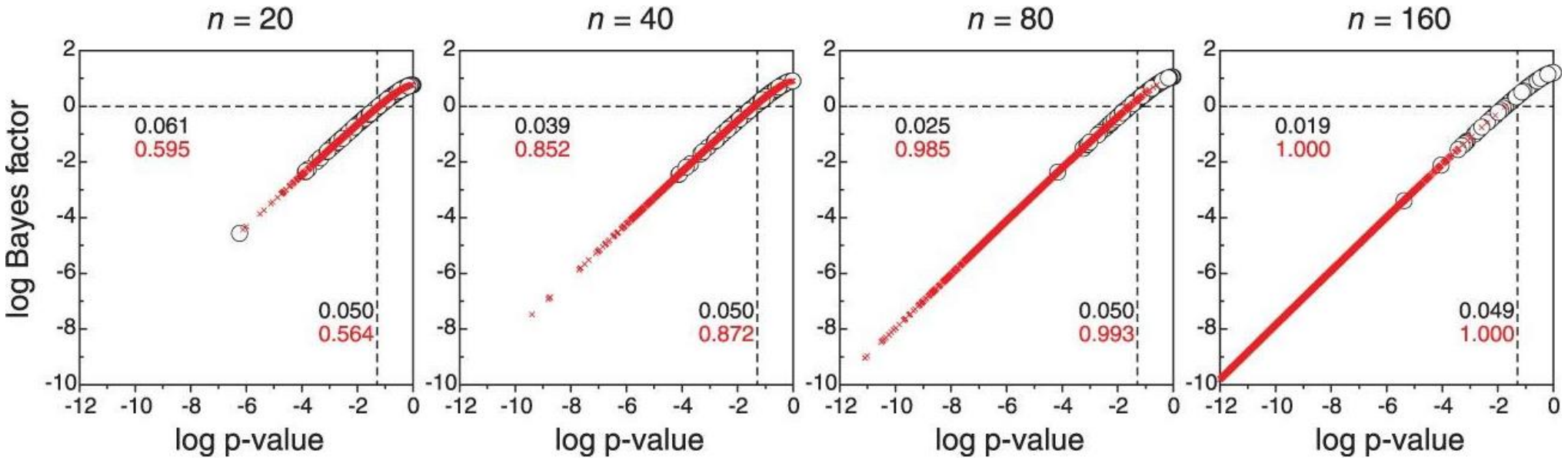
The Bayesian smoke-screen

- Bayes is often touted as a better alternative because dichotomous decisions are not made.
- Bayes factors are interpreted as strength of evidence resulting in "nuanced" proclamations such as:

"the data are x times more likely under the null (alternative) than under the alternative (null)"; or "the data display weak/strong/very strong evidence in favour of the null (alternative) hypothesis".

The Bayesian smoke-screen

- Garcia-Perez (2017) aptly demonstrated the duality between a *p*-value and Bayes factor.
- His figure below illustrates the one-to-one relation between a *p*-value a Bayes factor.
- Garcia-Perez (2017) concluded the *"Bayes factor does not carry any information that is not also in the p-value for given n … the Bayes factor is only a transformation of the p-value".*



Scatterplots of log Bayes factor against log *p* value for true (open circles) and false (red crosses) null hypotheses at four different sample sizes (panels) in a paired-samples (or one-sample) test for the mean.

DoE an unintended consequence?

The concept of 'statistical significance' is now banned.

As a result:

- α (level of significance) ceases to exist;
- Computation of power is no longer possible (since that requires specification of α);
- Sample size determination now impossible because that requires specification of power;

And therefore:

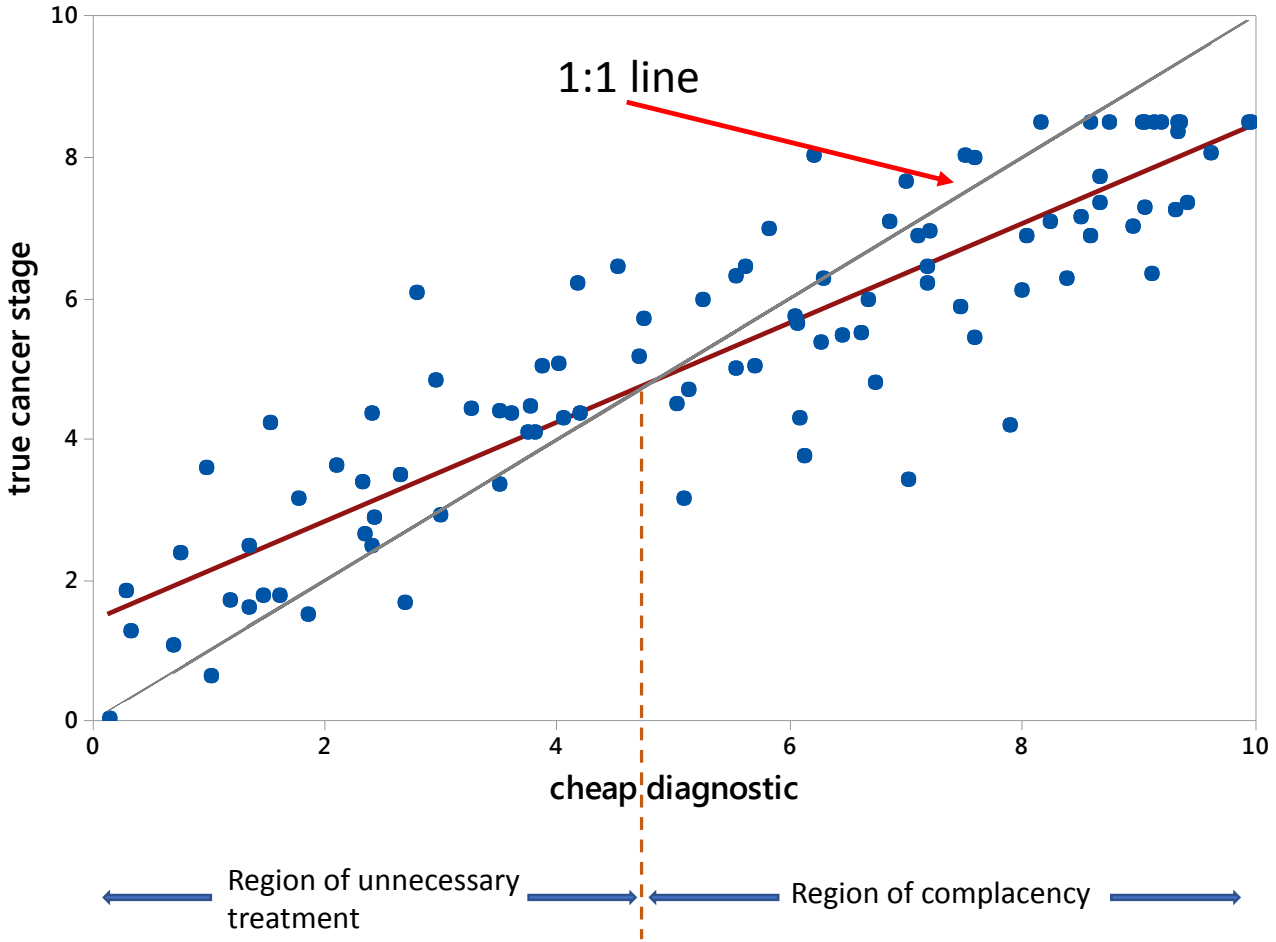
Experimental design as we currently know it, ceases to exist!

Nuanced thinking and interpretation

Hurlbert and Lombardi (2009) want to replace the use of ‘statistically significant’ with “nuanced thinking and nuanced language”

Let’s see how that might work with a contrived, but nonetheless realistic example.

Nuanced thinking and interpretation



The now 'discredited' interpretation:

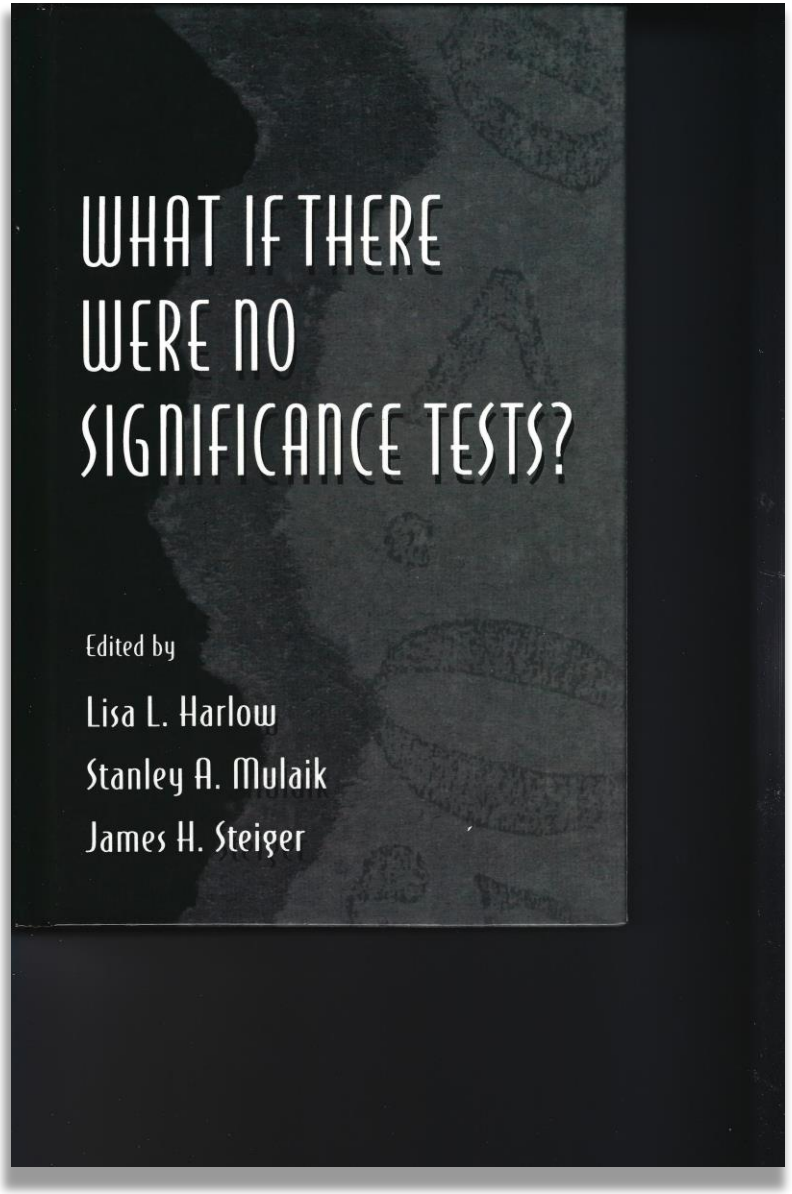
The slope of the regression line is significantly different from unity ($p<0.0000$) and therefore the cheap diagnostic procedure should not be used to predict the true cancer stage.

The 'nuanced' interpretation:

The difference between the true cancer stage and the estimated cancer stage depends on the value predicted by the cheap diagnostic. For predicted cancer stages of 8 and above or 2 and below, the differences are quite large.

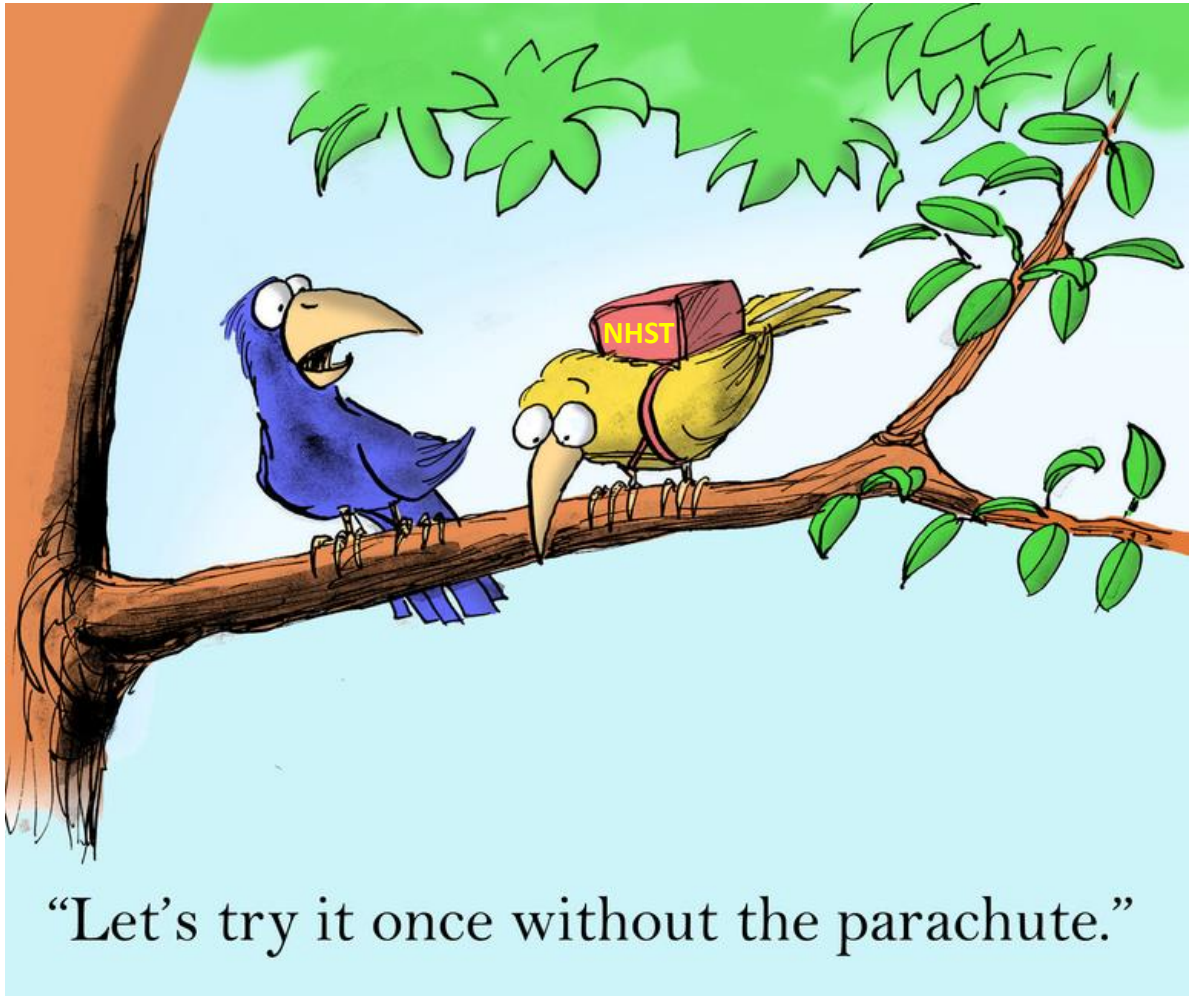
I know which one I prefer!

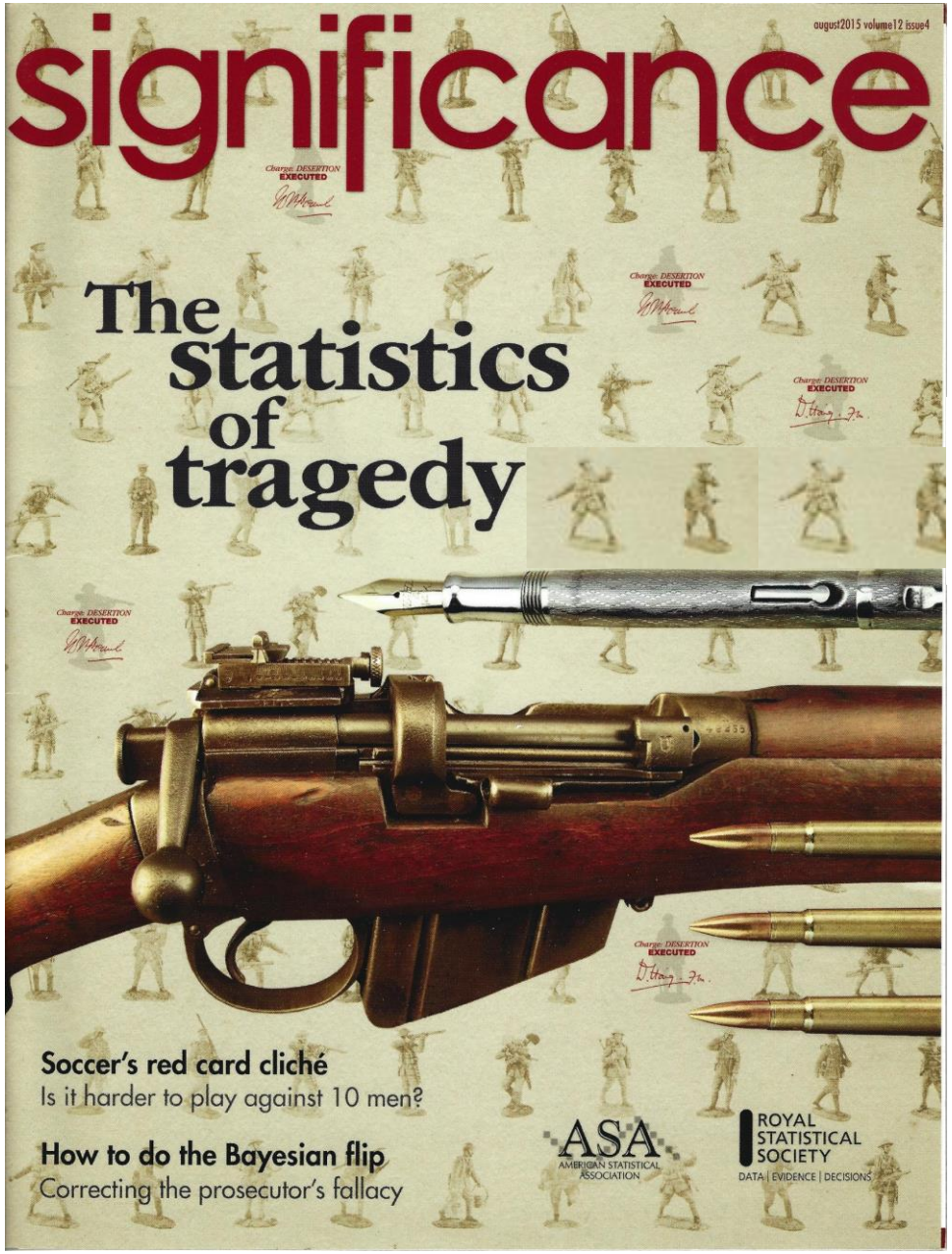
In 1997, the question was asked ...



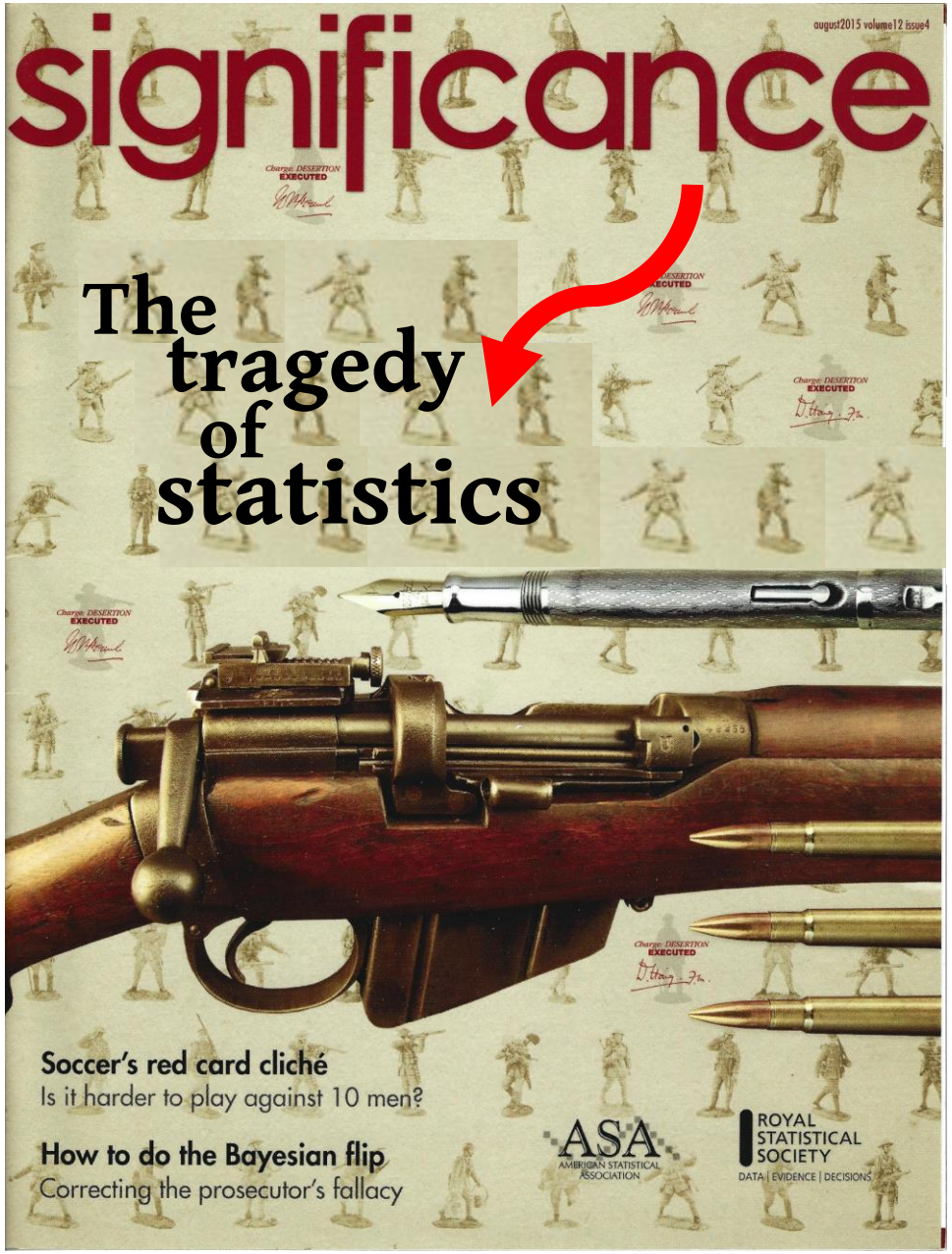
Wait no longer -
An Insignificant Future has arrived!

It remains to be seen if science flourishes or flounders.





Significance Magazine
April 2015



Significance Magazine
May 2019?

Thank you.

APRIL 2019 : STOP PRESS!

President of the American Statistical Association has doubts.

president's corner

P-Values: To Own or Not to Own?

The debate about the value of hypothesis testing and the over-reliance on *p*-values as a cornerstone of statistical methodology started well over a century ago, and it continues today. Many researchers, including statisticians, have commented about their use—and their abuse. Building on the presentations at the 2017 Symposium on Statistical Inference (*ww2.amstat.org/meetings/ssi/2017*), the ASA published the March 2019 issue of *The American Statistician* devoted entirely to this topic. (If you haven't done so already, I encourage you to read this issue. NPR, *Nature*, and many others commented on it the day the issue appeared.) The messages in the articles from that issue (all online) are not surprising to us: The "0.05 threshold" for *p*-values is arbitrary, and the notion of "$p < 0.05$" as "statistically significant" hardly makes sense in many (much less all) situations. Perhaps what is, or should be, surprising to us is where statisticians were when the "abuse" started to take hold.

Stephen Stigler notes this connection between *p*-values of 0.05 and "statistical significance" started well before Fisher: "Even in the 19th century, we find people such as Francis Edgeworth taking values 'like' 5%—namely 1%, 3.25%, or 7%—as a criterion for how firm evidence should be before considering a matter seriously" (*CHANCE* 21:4, 2008: doi: 10.1007/s00144-008-0033-3).

This sentence raises the central issue. How firm should evidence be "before considering a matter seriously"? The answer is one we statisticians have given frequently to our clients: "It depends." (Statisticians can be accused of using that phrase excessively.) How big is the study, how many inquiries do you plan to make of the data, how many analyses do you plan to run, what other data might bear on this study, what are the risks of false claims, …? In short, the answer requires us to *think*. (What a concept.)

Many years ago, I met a wonderful lady named Edith Flaster, a biostatistician from Columbia University. Throughout her life, Edith approached problems—in statistics and elsewhere—with sensible and practical solutions. Professionally, Edith had learned much from giants like Cuthbert Daniel and Fred Wood, who came through Columbia on numerous occasions. On one evening, she recalled the old days of computing on main frames, when every department had a computer budget and analyses cost real money. "Consequently," she said, "you had to think very carefully before you burned your computer budget on an analysis; you wanted to be sure the analysis made sense before you ran it. Today, computing is cheap, so people run hundreds of analyses, without even thinking before they run them. I don't care if you think before you run the analysis or after—but somewhere along the line you have to think." Calculating *p*-values does not relieve us of our duty to remind our collaborators we still have to think. And the more *p*-values we calculate, the more we have to think.

Karen Kafadar

Many of us would agree that, if we were to remove all thresholds for deciding when to take a result seriously, we may find ourselves back in the days of the Wild West. (Some may fear we are already there, given the proliferation of journals and analyses they contain.) We, unlike a few journal editors, recognize that adherence to a fixed *p*-value in all situations is not the antidote. And it is not a substitute for thinking. How many times has your collaborator insisted you include "($p < 0.05$)" in the paper you are writing, "because the journal requires it"? Regrettably, stating the *p*-value (to several decimal places no less, as if anyone would believe them) has become a requirement for many journals.

On the other hand, we need some sort of structure. We agree that the fixed threshold of "$p < 0.05$," and its identification with the term "statistical significance", is not sensible. (Even Sir Ronald, who receives "credit" (or "blame") for popularizing the 5% threshold,

"Many of us would agree that, if we were able to remove all thresholds for deciding when to take a result seriously, we may find ourselves back in the days of the Wild West"

"We, unlike a few journal editors, recognize that adherence to a fixed *p*-value in all situations is not the antidote"

"On the other hand, we need some sort of structure. We agree that the fixed threshold of *"p<0.05"*, and its identification with the term "statistical significance", is not sensible".

APRIL 2019 : STOP PRESS!

President of the American Statistical Association has doubts.

president's corner

If anything, the continued controversy about p-values and statistical significance reminds us our job as statisticians is far from done.

was reported to have said he'd be more likely to trust a result where $p < 0.05$ in 10 experiments than a result where $p < 0.005$ in a single experiment.) But if we advise scientists to dismiss any notion of thinking in advance about a level beyond which we take a result seriously, our profession may run the risk of being dismissed altogether—especially when our clients can go to "data scientists," who won't bother them with p-values at all—or, in fact, with any firm statistical foundations for their "scientific findings."

The real question is, where were we statisticians, and where have we been, when our collaborators and journal editors set this limit as a criterion for publication? Many of us were conducting research that has allowed our profession to flourish. That's been terrific. But the well-intentioned editors of scientific journals either ignored any notion of "thresholds for evidence" or insisted on an "algorithm" or "golden rule"—like "$p < 0.05$." As we've reminded our colleagues in other professions, algorithms don't always lead us to "truth." Nonetheless, the structure of an algorithm can be useful in getting us to think.

Stigler ends his article in *CHANCE* with a thoughtful sentiment:

> One may look to Fisher's table for the F-distribution and his use of percentage points as leading to subsequent abuses by others. Or, one may consider the formatting of his tables as a brilliant stroke of simplification that opened the arcane domain of statistical calculation to a world of experimenters and research workers who would begin to bring a statistical measure to their data analyses. There is some truth in both views, but they are inextricably related, and I tend to give more attention to the latter, while blaming Fisher's descendants for the former.

Alas, we are the descendants. We must take responsibility for the situation in which we find ourselves today (and during the past decades) regarding the use—and abuse—of our well-researched statistical methodology. And we must also, therefore, take responsibility for trying to change it.

I fervently hope the articles in the special issue of *The American Statistician* will not be viewed as a call to dismiss an area of our profession that has served, and continues to serve, us and science so well. Rather, I hope the articles will inspire us to encourage our colleagues to think about the data analysis process and to speak up to editors who, in their desire to bring structure to the inference process, may have gone just a little overboard. If anything, the continued controversy about p-values and statistical significance reminds us that our job as statisticians is far from done and that we are needed more than ever in this era of "data science" that embraces algorithms (with appealing names) and shuns complicated statistical inference. As noted in the last two columns, the debate reminds us to do the following:

a. Showcase all our talents—logical thinking, identification of process steps, design of relevant data collection, analysis and inference, characterization of uncertainty, clear results

b. Seize opportunities to create the demands for our talents—and then meet the demands with hard thinking

c. Be prepared to use our skills to present reasonable approaches to solving problems and encourage hard thinking, rather than blind adherence to fixed thresholds.

Please share your experiences—and your successes—in our mission to bring "sound thinking" to your collaborators. I look forward to hearing about them!

"But if we advise scientists to dismiss any notion of thinking in advance about a level beyond which we take a result seriously, our profession may run the risk of being dismissed altogether – especially when our clients go to "data scientists", who don't bother them with p-values at all – or, in fact, with any firm statistical foundations for their 'scientific findings'".

A video of this presentation will be available online here:

All enquiries to:

Prof. David Fox
david.fox@environmetrics.net.au