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 ...
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
Drivers of change

The ‘reproducibility crisis’ and scientific flip-flopping
Eggs Not Harmful for Heart Health

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

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hype claims and the dismissal of possibly crucial effects.
Statisticians Weigh into the debate (albeit in a timid way)
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.
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 Spiegelhalter

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”.
I have a confession to make. I like p-values.

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 already strongly convinced by the recent Spiegelhalter backflip and 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?

The Spiegelhalter backflip

"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"
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.
The majority of documented ‘problems’ with NHST arise from

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

These are all 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)
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”.

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):

“The 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.](image)
DoE an unintended consequence?

The concept of ‘statistical significance’ is now banned.

As a result:

• $\alpha$ (level of significance) ceases to exist;

• Computation of power is no longer possible (since that requires specification of $\alpha$);

• Sample size determination now impossible because that requires specification of power;

And therefore:

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

“Let’s try it once without the parachute.”
The statistics of tragedy

Soccer’s red card cliché
Is it harder to play against 10 men?

How to do the Bayesian flip
Correcting the prosecutor’s fallacy
The tragedy of statistics

Soccer’s red card cliché
Is it harder to play against 10 men?

How to do the Bayesian flip
Correcting the prosecutor’s fallacy

Significance Magazine
May 2019?

Thank you.
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.”

P-Values: To Own or Not to Own?

Many years ago, I met a wonderful lady named Edith Flowers, 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 Carlbert Daniel and Fred Woolf, who came through Columbia on numerous occasions. One evening, she recalled the old days of computing on mechanical frames, when every department had a compter 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 see the analysis made sense before you ran it. Today, computing is cheap, to people run hundreds of analyses, without even thinking before they run them. I don’t care if you think before you ran 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 colleagues we still have to think. And the more p-values we calculate, the more we have to think.

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 collaboration insisted you include “p<0.05” in the paper you are writing, “because the journal requires it”? Regrettably, stating the p-value (in several decimal places no less, as if anyone would believe them) has become a requirement for many journals.

“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”
President of the American Statistical Association has doubts.

“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’.”
All enquiries to:

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

A video of this presentation will be available online here: