doi:10.1111/imj.13723

EDITORIAL

P-value. What value?

Long-standing concerns over the erroneous use of *P*-values and null hypothesis significance testing (NHST) in statistical inference prompted the American Statistical Society (ASA) to publish its *Statement on P-Values and Statistical Significance*.[1] Further, the ASA convened a symposium on statistical inference titled *Scientific Method for the 21st century: A World beyond P < 0.05*.[2] Reflecting the ubiquitous nature of the problem and a sense of frustration, the conference organisers aimed for broad participation, including research funders, journal editors and reviewers, media representatives, consumers, educators and professional practitioners from diverse fields.

> Although the problems identified in the statement have been known for several decades, previous expressions of concern and calls for action have not fostered broad improvements in practice.[2]

The malalignment between statistical and scientific reasoning has been cited as a major reason for replication failure.[3,4] Medical literature provides particular examples of the erroneous teaching of statistical inference as illustrated by the following statement which appeared in an educational review article from a high-profile medical journal.

> A P value of 0.05 carries a 5% risk of a false positive result (i.e. there is no true difference between treatments). If a trial is meant to provide proof of a genuine treatment difference beyond reasonable doubt, a much smaller P value – say p < 0. 001 – is required.[5]

A subsequent letter to the editor pointing out the error and referencing the ASA[1] statement (Table 1) was met with re-affirmation while obfuscating with a true statement:

> We disagree ….that our statement… is erroneous. According to the null hypothesis, P < 0.05 will occur 5% of the time.[6]

No editorial corrigendum has appeared.

What is a *P*-value and what does it measure?

A *P*-value is the area under the curve of a probability distribution defined by a mathematical model. The model, usually presented graphically, describes the expected distribution of a sample statistic around a central measure, the parameter or theoretical 'true' value, for example the population mean, μ. Under the central limit theorem, this would be the standard normal distribution of sample means generated by repeat sampling of a population variable of interest.

The mean of the sample means would equal the 'true' population mean, μ. In medicine, it is rare for us ever to know the true value of the variable of interest. However, we can usefully assign a value in the special case of a difference statistic, for example the difference in mean outcome variables in a placebo-controlled drug trial. In this case, the sampling distribution would represent that of the difference statistic. In this case, if the value we assign μ is zero then the mathematical model becomes the null hypothesis used in NHST. By way of contrast, non-inferiority drug trials require a non-zero value to be assigned.

The cumulative AUC of the sampling distribution of a continuous variable is represented by a mathematical function called the cumulative density function. In medical science, most study variables are continuous or, if categorical, are transformed using the logit model. As the *P*-value is a mathematical integral, that is the cumulative AUC, it cannot take on a precise value as there is no AUC defined by a single point on the curve, for example the *P*-value ≤ 0.05, but not $P = 0.05$. While this may seem pedantic, the semantics of statistical inference are influential in thinking and decision-making yet misinterpretation and misuse of terminology are commonplace.

Under the null hypothesis, one sample mean that happens to fall within an extreme region of the standard normal distribution may be expected to occur with a low frequency, say $P \leq 0.05$ meaning such a sample mean or one more extreme would be expected to occur with a frequency of 5% or less. To be valid, the assumptions of independence and random selection of each sample mean selected from the normal distribution of sample means must be assumed. Another way of stating this is as a conditional probability:

probability data (e.g. mean difference) $P \leq 0.05$ | null hypothesis ($\mu = 0$).

Note: | means 'given'.

It is important to understand that the *P*-value is a measure conditional on the assumption that the mathematical model describes the distribution of sample means and

Table 1 2016 Statement by the American Statistical Association on statistical significance and *P*-values[1]

1	*P*-values can indicate how incompatible the data are with a specified statistical model
2	*P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3	Scientific conclusions and business or policy decisions should not be based only on whether a *P*-value passes a specific threshold
4	Proper inference requires full reporting and transparency
5	A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result
6	By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis

is not a measure of the probability of the 'truth' of the mathematical model. To make this claim would invert the conditional probability statement and commit an error of reasoning called transposing the conditional[7] aka the prosecutor's fallacy:

probability null hypothesis, $P \leq 0.05$ | the data (mean difference).

In reasoning from NHST, the commonly used definition of the *P*-value as 'a measure of evidence against the null hypothesis' is potentially misleading in that it seems to legitimise transposing the conditional as if it were a mathematically valid function rather than a matter of intuition. It was the intuitive interpretation that Fisher used in his *a posteriori* model of NHST.[8,9] His aim was to use the *P*-value as an aid in deciding which experiments to repeat. If on several repetitions, a consistent extreme *P*-value for the sample statistic was obtained then that would accumulate evidence for a true experimental effect. If no such effect was present, regression to the mean parameter (μ) would be expected ($P \geq 0.05$).

In real-life scenarios, many factors inhibit repetition and replication of experiments; however, modelling can give us insight into the precision and reproducibility of extreme *P*-values[10,11] and hence the intuitive weight we place on the *P*-value 'as a measure of evidence against the null hypothesis'.

Table 2 is a reproduction.[10] It describes the results of simulating repeat experimentation and the probability of producing a *P*-value ≤ 0.05 under the prescribed conditions of the simulated experiment. It may be surprising to many how poorly reproducible the *P*-value is as a bright line test (a bright line test is a clearly defined rule or standard, the purpose of which is to produce consistent and predictable results). For example, if in the first experiment $P \leq 0.05$ was produced there would be a 50% probability of reproducing $P \leq 0.05$ in a repeat experiment; if $P \leq 0.01$was produced in the first experiment the probability of producing $P \leq 0.05$ in a repeat experiment, would be 73%; and if $P \leq 0.001$ was produced in the first experiment the probability of $P \leq 0.05$ in a repeat experiment would be 91%. The magnitudes of a number of these first experiment *P*-values are those commonly used in pharmaceutical trials and other medical analyses.

The *P*-value is also sensitive to sample size. Irrespective of the effect size, with increasing sample size (n) the *P*-value can be made as small as you wish[12] because the standard error is proportional to the inverse of n. If statistical significance is substituted for 'clinical significance' even small irrelevant differences may be regarded as worthy of investment. Large sample sizes are often a feature of pharmaceutical trials of secondary and primary prevention interventions such as preventive therapies in atherosclerotic diseases and osteoporosis.

What does a *P*-value not measure? Error rates

The quoted extract from the article on clinical trials mistakenly promotes the *P*-value as a measure of error and further states that the error rate can legitimately be adjusted depending on the magnitude of the *P*-value thus providing 'proof of a genuine treatment difference beyond reasonable doubt'.

This erroneous interpretation has arisen from the illusion of coherence resulting from the conflation of the dominant models of hypothesis testing.[8,9] The setting of theoretical type 1 (α) and type 2 (β) error rates in the Neyman and Pearson model envisions the frequency of error 'in the long run of experience' (experimental repetition) given randomness and independence of sample means from two juxtaposed probability distributions. *A priori* two

Table 2 Results of stimulated experiments illustrating the reproducibility probability of $P < 0.05$. (Reproduced from Boos and Stefanski[10] with permission. Reproducibility probability estimates from two-sided tests of a single mean, variance known, Equation (6) with $\alpha = 0.05$. For equation (6), readers are referred to the referenced paper)

P-value	0.10	0.05	0.03	0.01	0.005	0.001	0.0001	0.00001
Reproducibility probability	0.38	0.50	0.58	0.73	0.80	0.91	0.97	0.99

For a stated *P*-value from a first experiment (row 1, *P*-value), the probability of obtaining $P \leq 0.05$ for a repeat experiment is given as the reproducibility probability. For example, for $P \leq 0.05$ in the first experiment the probability of producing $P \leq 0.05$ in a repeat experiment is 50%, for $P \leq 0.001$ in the first experiment the probability of producing $P \leq 0.05$ in a repeat experiment is 91%.

identical populations are imagined except that they differ in mean parameters, null μ_0 and alternative μ_A. This model is valuable in providing a rationality to sample size selection.

However, the conflation has resulted in confusion between Fisher's *P*-value and Neyman's α giving the *P*-value an apparent legitimacy as an *a posteriori* 'sliding' type 1 error rate. Even if this were logical, decreasing α would increase β, resulting in a decrease in power (1-β). Also the dichotomous approach of pitting null hypothesis against alternative hypothesis carries the risk of blinding the researcher or the consumer to other explanatory hypotheses.

For those who think the use of confidence intervals (CI) overcomes the problems described, think again. Although it has greater intuitive value especially with respect to estimating effect size, the CI relies on the same premises as the *P*-value. For example the CI of juxtaposed probability distributions can be made as large or as small as can be paid for by increasing the sample size such that for any small difference the CI can be made not to overlap.

What is the solution?

Statistical analyses are very valuable tools for extracting information from data. However, the reliability of the knowledge generated is dependent on many more important factors *inter alia*, evidential justification of the experimental hypothesis, study design, study conduct and data collection and cleansing, competence in choice of statistical model, valid reasoning, reviewer bias, publication bias and replication. Much of the criticism of medical science centres on its overemphasis on the importance of the *P*-value, NHST and statistically defined effect sizes.

A better understanding of how sound statistical inferences are made and how they influence decision making will be key elements to improving all aspects of healthcare. This is critically important in acknowledgement of individuals as complex adaptive systems with characteristics of emergence, adaptability, non-linearity and unpredictability[13] rather than as static population averages.

Surveys suggest statistical literacy amongst doctors is low.[14,15] Teaching and assessing knowledge and application of statistical inference, critical appraisal and decision-making skills should be a primary focus of medical schools and specialist colleges. Difficult concepts underpinning statistical inference may be more effectively and efficiently taught using computer simulation whereby the learner can manipulate effect sizes, sample sizes and other statistics in order to see how parameter estimates, *P*-values and CI change with reproduction and replication.[16] This will foster a more in-depth understanding of the limits of statistical inference, making clinicians better able to choose wisely amongst the myriad of investigations and treatment options on offer.

Addendum

Subsequent to article submission and review the author attended the referenced ASA conference.[2] A special issue of the ASA journal reporting the conference proceedings is planned for 2018. In the opening addresses, the 400 participants were encouraged to devote their energies to developing proposals and goals to address the long standing yet stubbornly persistent errors in statistical inference described in this article. While concrete proposals are yet to be endorsed by the ASA, many speakers emphasised the need to place greater emphasis on teaching the conceptual framework of the different philosophical approaches to science (mastering the concepts as a priority rather than the mechanics of statistical inference). The need for better understanding of statistical semantics on the part of non-statistician scientists was also highlighted. Further that the best way to achieve understanding would be to develop context-specific learning modules. An aspect of the conference that resonated with the author with respect to prediction in medical science was the idea that science defines degrees of uncertainty (not certainty) apropos caution must be applied to the use of prediction models in medical practice lest they be over-extended.

Received 23 August 2017; accepted 4 November 2017.

John L. O'Donnell [1,2]
[1]Department of Rheumatology Immunology and Allergy, Christchurch Hospital, and [2]Canterbury Health Laboratories, Canterbury District Health Board, Christchurch, New Zealand

References

1 Wasserstein RL, Lazar NA. The ASA's statement on *P*-values: context, process, and purpose. *Am Stat* 2016; **70**: 129–33.
2 American Statistical Association. ASA Symposium on Statistical Inference. 2017 [cited 2018 Jan 3]. Available from URL: ww2.amstat.org
3 Goodman SN. Aligning statistical and scientific reasoning. *Science* 2016; **352**: 1180–1.
4 Nuzzo RL. Statistical errors: *P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature* 2014; **506**: 150–2.
5 Pocock SJ, Stone GW. The primary outcome is positive – is that good enough? *N Engl J Med* 2016; **375**: 971–9.

6 Hu D. The nature of the *P* value. *N Engl J Med* 2016; **375**: 2205.
7 Evett IW. Avoiding the transposed conditional. *Sci Justice* 1995; **35**: 127–31.
8 Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front Psychol* 2015; **6**: 223.
9 Fienberg SE, Tanur JM. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *Int Stat Rev* 1996; **64**: 237–53.
10 Boos DD, Stefanski LA. *P*-value precision and reproducibility. *Am Stat* 2011; **65**: 213–21.
11 Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle *P* value generates irreproducible results. *Nat Methods* 2015; **12**: 179–85.
12 Demidenko E. The *P*-value you can't buy. *Am Stat* 2016; **70**: 33–8.
13 Sturmberg JP, Martin CM, eds. *Handbook of Systems and Complexity in Health*. New York: Springer Science Business Media; 2013.
14 Anderson BL, Williams S, Schulkin J. Statistical literacy of obstetrics-gynecology residents. *J Grad Med Educ* 2013; **5**: 272–5.
15 Martyn C. Risky business: doctors' understanding statistics. *BMJ* 2014; **349**: g5619.
16 Cumming G, Calum-Jageman R. *Introduction to the New Statistics: Estimation, Open Science, & Beyond*. New York: Routledge, Taylor and Francis Group; 2017.