Human Reproduction, Vol.31, No.11 pp. 2406–2410, 2016
Advanced Access publication on September 22, 2016 doi:10.1093/humrep/dew192

human reproduction

EDITORIAL COMMENTARY

P-values and reproductive health: what can clinical researchers learn from the American Statistical Association?

L. V. Farland[1,2,*], K. F. Correia[3], L. A. Wise[4,5], P. L. Williams[1,3], E. S. Ginsburg[2], and S. A. Missmer[1,2,6]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA [2]Department of Obstetrics, Gynecology, and Reproductive Biology, Brigham and Women's Hospital and Harvard Medical School, 221 Longwood Avenue, Boston, MA 02115, USA [3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA [4]Department of Epidemiology Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, USA [5]Slone Epidemiology Center, Boston University School of Public Health, 1010 Commonwealth Avenue, Boston, MA 02215, USA [6]Channing Division of Network Medicine, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA

*Correspondence address. OB/GYN Epidemiology Center, Brigham and Women's Hospital, 221 Longwood Ave, Boston, MA 02115, USA. E-mail: lfarland@mail.harvard.edu

Key words: *P*-values / confidence interval / statistical significance / data interpretation / biostatistics

Introduction

The use and misuse of the *P*-value has been under discussion for many years in epidemiology, biostatistics and psychology journals (Nuzzo, 2014; Rothman, 2014; Schmidt and Rothman, 2014; Greenland *et al.*, 2016), but recently *P*-values gained national attention after the American Statistical Association (ASA) released a statement on statistical significance and the use of *P*-values for data interpretation (Wasserstein and Lazar, 2016). The ASA statement provided six principles for using and interpreting *P*-values (Table I). The purpose of this commentary is to help elucidate the appropriate use of *P*-values for the readers of Human Reproduction. Through exercises and practical examples, we hope to encourage reflection and discussion regarding the use of *P*-values in designing our own reproductive health research, interpreting the scientific literature in our field and reviewing the work of our peers.

What is a *P*-value?

Although most of us were taught the definition of a *P*-value at some point in our careers, this definition may not always align with how we think about or interpret *P*-values on a daily basis. For many readers, a *P*-value is conceptualized as a strict cut-point, with a $P > 0.05$ interpreted as a lack of association. However, the definition of a *P*-value is substantially more complex and is often misunderstood.

As informally defined in the ASA's statement, a *P*-value is 'the probability under a specified statistical model that a statistical summary of the data (e.g. the sample mean difference between two compared groups) would be equal to or more extreme than its observed value'. To illustrate, consider an example where we are examining sex differences in birthweight. Suppose we observe in our sample a mean difference in birthweight of 0.25 kg between girl and boy infants. One assumption we must make in calculating the *P*-value is that there is no difference in mean birthweight in the underlying population (commonly referred to as the 'null hypothesis'). Another assumption we must confirm in calculating the *P*-value is that birthweight is normally distributed. Based on these assumptions, we can calculate validly a two sided test statistic for the probability of observing a mean difference at least as extreme as our observed statistic, i.e. further away in either direction from the hypothesized difference of zero (≤ -0.25 or ≥ 0.25 kg). This probability, calculated under a specific set of assumptions, is the *P*-value. The *P*-value does not tell us about the truth of the null hypothesis or the probability of random chance, as the ASA's statement reminds us. However, 'the *P*-value can indicate how incompatible the data are with a specified statistical model' (ASA Point 1) (Wasserstein and Lazar, 2016). In our example, the *P*-value is telling us the probability of finding an average difference in birthweight between boys and girls that is at least as extreme or more extreme than 0.25 kg under the null hypothesis and assuming birthweight is normally distributed.

The ASA statement emphasizes that 'scientific conclusions should not be based only on whether a *P*-value passes a specific threshold'. In other words, scientists must consider a broader range of information to reach scientific conclusions, such as how the study was designed, what methods were used to minimize chance, bias and confounding, and what previous studies on the topic have shown. The *P*-value threshold of 0.05 was first suggested by Fisher (1926) in the context of experimental agricultural research. This value was chosen because Fisher believed that 1/20 was a reasonable threshold to use, but there

Table I The ASA statement's six principles on statistical signficance and *P*-values.

1	*P*-values can indicate how incompatible the data are with a specific 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 the result.
6	By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis.

is no mathematical reasoning behind this arbitrary dichotomy that has come to dominate medical decision-making.

With the growing popularity of 'big data', questions about correction for multiple comparisons continue to arise. Methods that mathematically account for multiple comparisons, such as a Bonferroni correction or control of the False Discovery Rate, reinforce the concept of a *P*-value as a dichotomous decision-making point which, as discussed in this commentary, should be avoided (Rothman, 1990). In addition, a single manuscript's incorporation of these adjustments ignores the total sum of comparisons that are quantified across many studies that arise from one 'big data' source. Epidemiologic and clinical researchers should approach research questions with strong *a priori* hypotheses informed by biologic mechanism and previous research. '*P*-value fishing expeditions' should be avoided. In addition, studies with multiple exposures or multiple outcomes—such as exploratory chemical safety studies evaluating toxicity and 'agnostic' genetic approaches like genome wide association studies (GWAS)—must balance the issue of inflating the risk of falsely concluding an association when none exists (i.e. increasing the Type I error rate) with the risk of failing to detect true associations (i.e. increasing the Type II error rate).

How should (and should not) a *P*-value be interpreted?

Below we have provided four examples of how *P*-values are often used in reproductive health research and how they should be interpreted in each context.

Example 1: *P*-values in demographic tables

In the context of a demographic table, is there utility in presenting a *P*-value and what does it tell us? A demographic table, usually 'Table 1' in a article, typically describes the population under study and gives the reader a sense of differences in demographic characteristics in the population according to exposure (or outcome). This is often used to guide the researcher as to potential confounders that merit adjustment in multivariable analysis. In our example, in which population below are the differences in age meaningful?

Table II Example demographic characteristics for Population 1.

	Mean (SD)	
	Not obese	**Obese**
Age (years)	38.1 (6.3)	40.0 (4.0)

Table III Example demographic characteristics for Population 2.

	Mean (SD)	
	Not obese	**Obese**
Age (years)	37.6 (6.3)	37.9 (4.0)

Looking at Tables II and III we may be more concerned about the difference in average age in Population 1 given the large absolute difference. Does adding a P-value to the table alter our thinking? What if the tables looked like this?

Table IV Example demographic characteristics for Population 1 with the addition of a *P*-value.

	Mean (SD)		*P*-value
	Not obese ($n = 20$)	**Obese ($n = 20$)**	
Age (years)	38.1 (6.3)	40.0 (4.0)	0.26

Table V Example demographic characteristics for Population 2 with the addition of a *P*-value.

	Mean (SD)		*P*-value
	Not obese ($n = 7000$)	**Obese ($n = 7000$)**	
Age (years)	37.6 (6.3)	37.9 (4.0)	<0.001

When we incorporate additional information in Tables IV and V, we see that there is a statistically significant difference in age in Population 2 (Table V) but not in Population 1 (Table IV). That is because *P*-values combine two important pieces of information: the magnitude of the effect size (in this example the average difference between two continuous variables) and the sample size. In this example, the large sample size of Population 2 is driving the conclusion that age is significantly different.

A naïve researcher observing Population 1 may construe the null *P*-value as justification for not accounting for age in subsequent analyses. Conversely, a naïve researcher observing Population 2 may declare differences in age between the obese and non-obese groups based on the highly significant *P*-value. However, a statistically

significant *P*-value does not necessarily indicate biological significance. Regardless of the *P*-values, there is a clinically meaningful difference in mean age among Population 1 and not a clinically meaningful difference in mean age among Population 2. Relying too heavily on *P*-values in this example may cause poor scientific decision-making.

Example 2

Table VI further illustrates how sample size can influence *P*-values. You can see that for Population 2, the *P*-values become smaller as the population increases but the effect size remains the same. The standard deviation (SD), which is a measure of the variability around the sample mean (i.e. how far does the age of a typical individual in the sample fall from the mean age?) does not vary by sample size, and is thus kept constant.

The standard error (SE) is a measure of how precise an estimated parameter (e.g. sample mean) is of the true underlying parameter (e.g. population mean). As the sample size increases, the SE decreases, the *P*-value becomes smaller and the difference between the two groups becomes statistically significant. However, as scientists, we must keep the effect size in mind. In this example, the small absolute difference in age between the groups remains unchanged across all sample sizes and thus the age difference still lacks clinical significance.

Table VI Example demographic characteristics shown for a range of sample sizes.

(Years)	Sample size per group	Not obese		Obese		*P*-value
		Mean (SD)	SE	Mean (SD)	SE	
Age	20	37.6 (6.3)	1.41	37.9 (4.0)	0.89	0.86
Age	100	37.6 (6.3)	0.63	37.9 (4.0)	0.40	0.69
Age	1000	37.6 (6.3)	0.20	37.9 (4.0)	0.13	0.20
Age	2500	37.6 (6.3)	0.13	37.9 (4.0)	0.08	0.04
Age	5000	37.6 (6.3)	0.09	37.9 (4.0)	0.06	0.004
Age	10 000	37.6 (6.3)	0.06	37.9 (4.0)	0.04	<0.001

The effect of sample size on *P*-values is important when we consider typical sample sizes available in reproductive health studies. An individual fertility clinic may have more uncertainty (reduced statistical precision) in their ability to detect an association because of their smaller sample size compared with national level data, such as that collected by the Society for Assisted Reproductive Technology (SART), however, both studies may find the same patterns of association. A large sample size, which can contribute to a statistically significant *P*-value, does not guarantee lack of bias, misclassification or confounding in the effect estimate. Conversely, a study may lack precision because of a small sample size, but may have been well designed and analyzed in a way that is free from bias and confounding. While not the main focus of this commentary, taking into account study design and methods used to minimize bias and confounding should contribute more to interpreting and contextualizing study findings than statistical significance alone.

Example 3

Now let us consider the role of *P*-values in interpreting the main findings of a study, assuming the study results are free from bias and confounding. Given only the information provided in Table VII, how would one interpret the findings?

Table VII Example table showing main effect estimates and *P*-values with descriptive information covered.

Exposure [redacted]	[redacted] mean (SD)	Mean difference in [redacted] (95% CI)	*P*-value
No ($n = 601$)	[redacted]	(Referent group)	
Yes ($n = 399$)	[redacted]	93.0 (67.7–118.4)	<0.0001

CI: Confidence interval.

We may conclude that this Table VII shows a 'highly significant' relationship, a $P < 0.0001$. All too often people imagine such a highly significant result to be synonymous with a very important result, a very strong relationship and/or a definitive association. However, a *P*-value alone does not speak to any of those points.

Does our interpretation of the relationship change when additional information on the exposure (ICSI) and the outcome (peak Estradiol (E_2)) level is provided in Table VIII?

Table VIII Example table showing main effect estimates and *P*-values with descriptive information revealed.

Exposure ICSI used	Peak E_2 mean (SD)	Mean difference in peak E_2 (95% CI)	*P*-value
No ($n = 601$)	2399.9 (209.3)	(Referent group)	
Yes ($n = 399$)	2492.9 (189.6)	93.0 (67.7–118.4)	<0.0001

E_2, estradiol.

As we learn more information about the question of interest and examine the distribution of the outcome our opinion may change. First, we may see that the question investigated—the effect of ICSI on peak E_2 levels—lacks biologic plausibility because ICSI is not temporally or biologically related to peak E_2 levels. We may also see that the mean difference between ICSI and non-ICSI users (93.0 units), while highly statistically significant, does not represent a meaningful absolute difference in peak E_2 levels. A *P*-value alone does not inform the reader about the importance, magnitude, or the direction of the effect. Nor does the *P*-value provide insight into how robust the results are with regard to study design, bias and confounding. Relying on a statistically significant *P*-value to interpret our findings can lead us to formulate inaccurate conclusions that could stall or misdirect the progression of science, and in worse case scenarios, lead to inappropriate use of medical treatment (Rothman, 2016). In fact, the inclusion of the *P*-value in this

Table IX Example table showing relationship between four exposures (A, B, C and D) and live birth.

	OR (95% CI)	*P*-value
Exposure A		
No	1.00 (ref)	
Yes	1.34 [redacted]	<0.0001*
Exposure B		
No	1.00 (ref)	
Yes	1.34 [redacted]	0.042*
Exposure C		
No	1.00 (ref)	
Yes	1.34 [redacted]	NS
Exposure D		
No	1.00 (ref)	
Yes	1.34 [redacted]	NS

*$P < 0.05$.
OR: odds ratio; NS represents non-significance.

Table X Example table showing relationship between four exposures (A, B, C and D) and live birth with 95% CIs revealed.

	OR (95% CI)	*P*-value
Exposure A		
No	1.00 (ref)	
Yes	1.34 (1.19–1.51)	<0.0001*
Exposure B		
No	1.00 (ref)	
Yes	1.34 (1.01–1.78)	0.042*
Exposure C		
No	1.00 (ref)	
Yes	1.34 (0.99–1.81)	0.058
Exposure D		
No	1.00 (ref)	
Yes	1.34 (0.67–2.70)	0.41

*$P < 0.05$.
NS represents non-significance.

context does not add any additional useful information beyond what the confidence interval (CI) provides (which will be discussed in more detail in Example 4). For this reason, some associate editors of this journal (L. A.W., S.A.M.) encourage omitting *P*-values altogether.

In this example we have assumed a well-designed study which is free of bias and confounding to focus our discussion on the interpretation of *P*-values. However, if the study was not well designed, had residual confounding or residual bias neither the *P*-value nor CI would be accurate. A *P*-value cannot be understood in a vacuum. Information regarding the design of the study, possibility of bias, biologic relevance of the question, and magnitude of the effect should all contribute to scientific decision-making.

Example 4

In our last example, we will compare the information provided by *P*-values and by CI. Consider Table IX in which we are comparing the effect of four different exposures on odds of live birth assuming a well designed study with minimal bias and confounding. How would you interpret the relationship between our exposures of interest and live birth? Based on the information provided in Table IX, which exposure do you think is more important for influencing live birth?

Exposures A and B are associated with statistically significant increased odds of live birth. Exposures C and D are not associated with live birth at the $P < 0.05$ level. The effect size of all exposures is the same. How does the addition of CIs in Table X expand our understanding of the relationship?

By including the information provided by the CIs in Table X, we have a better sense of the precision or uncertainty of the effect estimate, which can help us in interpreting scientific findings. For example, despite the statistical significance of both Exposures A and B, the CI is much wider for Exposure B than for Exposure A, indicating less precision in the effect estimate for B. When we compare Exposures B and C we see that the difference in magnitude of the CIs between the two exposures is very minimal; however, Exposure B is statistically significant at a $P < 0.05$ level and Exposure C is not. Additionally it is important to note here, that we should refrain from using CIs as surrogate significance tests, as this would present the same problems with interpretation and dichotomization as discussed previously (Cummings, 2012; Rothman, 2016). When we compare Exposures C and D, we see that both exposures are not statistically significant; however the precision in Exposure C is much greater than that for Exposure D. Data provided by the CIs in Table X clearly show the dissimilarities between exposures A and B and exposures C and D, despite the fact that exposures A and B, but not C and D, are statistically significant.

CIs show information on the size and precision of the effect. This information also contributes to the calculation of the *P*-value, but is not clearly expressed in the *P*-value itself. Again, including the *P*-value in the table does not provide any additional useful information from what is provided in the CI alone. It is also important to note that 'NS' notation is entirely uninformative and is never an acceptable notation in scientific publication.

Summary

So what should the reproductive clinical research community learn from the ASA's statement? We are encouraging readers and authors of Human Reproduction to take the following steps going forward to more thoughtfully incorporate and utilize *P*-values in their research.

- Researchers should carefully consider whether the *P*-value adds anything meaningful to descriptive population statistics, as is often presented in demographic tables and Table I.
- Findings with strong biologic plausibility from carefully conducted studies should be welcomed and discussed by the scientific community, even if they lack statistical significance. Conversely, statistically significant findings that lack biological significance should be viewed with greater skepticism.

– As researchers and consumers of the scientific literature, effort should be made to understand CIs when presenting and interpreting study findings. When CIs are presented, *P*-values do not offer additional information and are the less informative of the two measures.

Authors' roles

S.A.M. conceived of the commentary. L.V.F. and K.F.C. created examples. L.V.F. drafted the article. L.V.F., S.A.M., K.F.C., P.L.W., E.S.G., L.A.W. critically reviewed the article and approved its final version.

Funding

L.V.F. was supported by a T32 grant (#HD060454) in reproductive, perinatal and pediatric epidemiology from the Eunice Kennedy Shriver National Institute of Child Health and Human Development and by the National Cancer Institute (3R25CA057711), National Institutes of Health. K.F.C. was supported by the National Institute of Allergy and Infectious Diseases (T32AI007358), National Institutes of Health. P.L.W. was supported by grants ES009718 and ES022955 from the National Institute of Environmental Health Sciences (NIEHS). L.A.W. was supported by NIEHS R01-ES024749, National Institutes of Health.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Cummings G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge, 2012.

Fisher R. The arrangement of field experiments. *J Ministry Agric* 1926;**33**: 503–513.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;**31**:337–350.

Nuzzo R. Scientific method: statistical errors. *Nature* 2014;**506**:150–152.

Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;**1**:43–46.

Rothman KJ. Six persistent research misconceptions. *J Gen Intern Med* 2014;**29**:1060–1064.

Rothman KJ. Disengaging from statistical significance. *Eur J Epidemiol* 2016; **31**:443–444.

Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. *Int J Cardiol* 2014;**177**:1089–1090.

Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* 2016;**70**:129–133.