Indian J Med Res 151, April 2020, pp 275-278
DOI: 10.4103/ijmr.IJMR_980_19

Quick Response Code:

Perspective

Attack on statistical significance: A balanced approach for medical research

Most medical research around the world is empirical and uses data to derive a result. Many researchers substantially depend on statistical evidence such as *P* values to decide that an effect of a specific factor is present or not. Now, there is a storm around the world, and the *P* value, particularly the resulting statistical significance, has been not just questioned but also sought to be abolished altogether. Abandoning statistical significance has the potential to change research in empirical sciences such as medicine forever. This article discusses the arguments in favour and against this contention and pleads that medical scientists present a balanced picture in their articles where *P* values have a role but not as dominant as is currently seen in most publications. The following discussion would also make medical researchers aware of this raging controversy, help them to understand the involved nuances and equip them to prepare a better report of their research.

Attack on the *P* values

Though concerns have been expressed in the past regarding the validity of *P* values[1,2], the threshold such as 0.05 and the resulting statistical significance are now under attack. The onslaught began in 2015 from the editors of Basic and Applied Social Psychology who banned the use of these concepts for the articles published in their journal. These concepts, according to them, are often used to support low-quality research[3]. This ban created a furore and galvanized statistics professionals to sit back and re-think. The American Statistical Association (ASA) formed a committee to examine the issues and to make a recommendation. Consequently, the ASA issued a statement in 2016 saying, "Scientific conclusions and business or policy decisions should not be based only on whether *P*-value passes a specific threshold"[4]. As a follow up in 2019, The American Statistician brought out a special issue with 43 articles on this topic. Based on a review of these articles and other literature, the editorial of this issue concluded, "it is time to stop using the term statistically significant entirely"[5]. Two of the three authors of this editorial were the same who earlier framed the 2016 ASA statement but now suggested to abolish the term altogether[5].

Amrhein *et al*[6] called for "a stop to the use of *P*-values in the conventional dichotomous way - to decide whether a result refutes or supports a scientific hypothesis", implying to discard the label of statistical significance for $P<0.05$ or any other threshold. They prepared a comment which was signed by more than 800 scientists, including statisticians, clinical and medical researchers, biologists and psychologists from more than 50 countries[6]. However, they did not plead to ban *P* values altogether. Although current annoyance is mostly with the term 'statistically significant' and not so much with *P* values, yet this has tremendous implications for medical research that has shown so much dependence on *P* values.

What actually is a *P* value?

It seems that much of misgivings arise from the way the *P* values are explained and understood. For example, *P* value is sometimes misinterpreted as the probability of the null hypothesis being true given the sample. Woolston[7] stated that "The closer to zero the *P* value gets, the greater the chance that the null hypothesis is false". This is a simplistic explanation and possibly a root cause of its misinterpretation. The author quickly clarified that *P* value was the probability of obtaining the data at least as extreme as those observed if the null hypothesis was true[8]. This indeed is the correct meaning. Because of this complex meaning of *P* value, simplistic but erroneous explanations often emerge and are possibly responsible for statements such as *P* values confuse more than clarify and "they are misused, misunderstood and misrepresented"[9].

Considering *P* value as a wrong measure of evidence[10] is like saying 'guns kill people'. The problem is not so much with *P* values but is with their abuse, misuse and overuse.

In simple terms, *P* value is the measure of the consistency of sample values with the null hypothesis. If the null is that the efficacy of the new regimen is the same as of the existing regimen, *P* value with the trial data may show that it is inconsistent with this null. While it is true that *P* value does not reveal the 'plausibility, presence, truth or importance of an association or effect'[5], it does reflect the chance that the observed values have come from a population with the hypothesized values of the parameter. It certainly is a probability statement and, as any other probability, is valid in the long run under perfect conditions - in this case, for repeated large number of identical trials. If we toss a coin to find if it is unbiased, the null hypothesis is H_0: π=½. If out of three tosses, all show up head, the probability of this occurrence under the null is 1/8. This is not particularly small to conclude that the coin is biased. However, if out of 10 tosses, all 10 show up head, the probability under the null is 0.00097. This is the *P* value. But is this enough evidence against the null or not? Can we conclude with these 10 tosses that the coin is biased and π is not ½ but something else? Most will agree that we can. For this conclusion, we need to ensure that the tossing was fair and no other factor interfered because then only the result can be believed.

Extension of this to medical research is immediate. If we know that particular surgery is successful in 70 per cent cases of a specific type, and a new procedure turns out to be successful in 9 out of 10 cases, can it be claimed that the success rate of the new procedure is higher than 70 per cent? One method is to accept it on face value, and the second is to express doubt because the sample size is too small to arrive at a firm result. If the sample size is 100 with 90 successes, would that remove the doubt? In my opinion, *P* value for H_0: π=0.70 is the answer. Again, it is to be ensured that the patients come from the same population as the one undergoing the previous procedure with success rate 70 per cent and nothing else has changed. The merits of the new procedure may have to be enumerated, which would raise the expectation that the success rate would be higher. However, that is only the expectation and it is to be supported by evidence. This evidence in this case comes from 90 successful surgeries out of 100.

What is statistical significance and what are its implications?

There has been a convention to consider small *P* value as an indication that the data contradict the null hypothesis and call the result statistically significant. The ASA statement says, "the smaller the *P* value, the greater the statistical incompatibility of the data with the null hypothesis"[4]. How small is small? The convention is to use 0.05 as the threshold for most investigations. This threshold is arbitrary but has been accepted without much concern so far.

Although the initiator of the present debate banned *P* value itself[3], the recent debate is not so critical of the *P* values as of the term statistical significance[5]. The dichotomy between significance and non-significance arises using a threshold such as 0.05. Cautions are advised for interpreting the borderline values of *P*[11,12], but the cut-off is so ingrained in medical research that statistical significance many times transcends to 'worthy' results and unfairly used to decide what results to report and publish. The problem is not so much with *P* values, nor possibly with statistical significance, it is with the dominant role that such significance has started playing in decision-making - not just in what is to be published but also, more importantly, in reaching to a decision about the usefulness of a result in prevention and treatment of a health condition. This certainly needs to be stopped, as any clinical conclusion cannot be based entirely on *P* value. *P* value also incorporates obscure uncertainty such as due to sample not really random, values not independent and distribution different from the one postulated[11]. Other considerations such as the clinical significance of the effect, biological plausibility and previous findings have to be considered before reaching a conclusion.

Though the considerations mentioned above are important, the role of *P* values cannot be completely disregarded. Even though *P* value is inversely calculated - the probability of the sample given the null hypothesis instead of the probability of the null given the sample - it remains the only prominent and easily understood objective criterion to measure the role of sampling fluctuation when the sampling is random. It automatically incorporates the contribution of the small or large standard error of the estimate of the parameter under consideration. However, caution is needed in interpretation as this probability is for repeated samples and does not apply to an individual study. Clinical decisions are made by individuals for individuals. When a new patient comes, the

uncertainties (or rather certainties) are measured in terms of probability just as the probability of head is ½ in a single toss of a coin although this is actually applicable to a large number of tosses. Furthermore, the alternatives to *P* values discussed so far do not inspire much confidence yet.

Alternatives to *P* values

No widely acceptable alternative to the *P* value is available as of now, but many proposals have come up. The most popular of these, now being pushed for the past several decades, is estimation of the effect size and its confidence interval (CI) that can also be used to test a null hypothesis[13]. However, this requires an independent assessment of the medical significance of the estimated effect size in the sense of being capable of changing the present practice. The primary problem with this proposal is that a 95% CI is as arbitrary as the five per cent level of significance. The second method is Bayesian hypothesis testing[14] that allows researchers to quantify the evidence and monitor its progression as further data become available. However, the Bayes factor used in this case can be hacked (just like *P* values)[14]. The most promising proposal is second-generation *P* values[15]. This requires setting up a composite null hypothesis containing the range of trivial effects. This range must be specified at the planning stage. The difficulty with this is that the definition and meaning of trivial effect may differ from physician to physician thus needs to be fully justified.

All these alternatives are still evolving. Not much evidence is available yet that any one of these will work better than the other or whether any one of these will perform better in the long run than the existing *P* values. It has taken decades to find flaws with the threshold of *P* values, and a similar time may be needed to establish one of these as a better method. Till such time that a credible alternative establishes itself as a better method, *P* value may continue to guide us to arrive at a result one way or the other. However, as mentioned earlier, *P* value should be only one of the several considerations under the balanced approach.

Balanced approach

Ideally, *P* value should be free of obscure uncertainty due to unknown or unaccounted factors. The quality of data generated by the study, including the appropriate design, the right methods used for analysis and correct interpretation are important considerations to reach a valid result.

A threshold such as 0.05 for *P* value has been a great asset to be objective and uniform in our approach, and such a cut-off has also been helpful to convert it to a binary decision such as a particular diagnostic method is better than the other, or one treatment regimen is more efficacious than the other. These dichotomies will require a threshold just as is required for many medical parameters for diagnosis. This can be reduced to 0.01 to minimize non-reproducibility. The other commonly advocated approach of estimating the effect size, and interpreting it in the context of its sampling variability as measured by its standard error, is helpful only when the minimum medically significant effect is defined. This definition is not easy because physicians differ from one another for what minimum effect is to be considered medically significant. An agreed *P* value threshold, such as 0.01, has no such problem in most cases. The threshold can also be dispensed with, and the emphasis may be on reporting the exact *P* value as is being advocated[16]. In addition, it is worthwhile to adhere to the guiding principles emerged from the ongoing debate. These principles mainly require consideration of (*i*) "related prior evidence, plausibility of mechanism, study design and data quality, real-world costs and benefits, novelty of finding and other factors that vary by research domain"[17]; (*ii*) "countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualize and interpret the potential for error"[18]; and (*iii*) "make all judgments as carefully and rigorously as possible and document each decision and rationale for transparency and reproducibility"[19]. These considerations will strengthen the balanced approach to come to a valid and robust conclusion in medical research.

In the absence of a threshold such as 0.05 (or 0.01), the assessment that a *P* value is small or large may depend on considerations such as estimate of the magnitude of effect and its precision. If the *P* value is large, conclude that the study could not gather sufficient evidence to change the *status quo*. The study results will still provide lessons for future studies of that kind. Interpretation of *P* values assumes that the design, data and analysis are correct. If the *P* value is small, further assessment is needed before reaching a conclusion. The biological plausibility of the result is desirable but, in the case of some new findings, this can emerge in the future. Most importantly, there is a need to distinguish between the statistical significance of a result and its clinical significance[11,12]. The clinical significance of a result in most cases would depend

on the effect size and the precision of its estimate. High precision may require a large sample that most studies cannot afford. Thus, results from small-scale studies should be considered as indicative and not conclusive. On the other hand, a large-scale study can give a small *P* value for a medically trivial effect. Statistical significance in terms of a low *P* value is required before assessing clinical significance because statistical significance helps to rule out the role of sampling fluctuations in exhibiting the non-null effect. Besides biological plausibility, cost-benefit, merits of the design in controlling biases and confounders, validity of the data, correct analysis and appropriate inference are all important considerations to reach to a valid conclusion. The conclusion also depends on the results of previous studies and other concomitant information available on the topic. All these should be considered for presenting a convincing argument regarding the validity of the conclusion. In all cases, accept uncertainty as an integral part of research endeavours and be modest in making a claim[5]. A similar result in replications in other settings strengthens the conclusion. Strategies such as meta-analysis can combine varying results to come up with a more reliable conclusion.

***Conflicts of Interest*:** None.

Abhaya Indrayan
Biostatistics Consultant, Max Healthcare,
Saket, New Delhi 110 017, India
a.indrayan@gmail.com

Received June 1, 2019

References

1. Cohen J. The earth is round ($p < 0.05$). *Am Psychol* 1994; *49* : 997-8.
2. Pharoah P. How not to interpret a *P* value? *J Natl Cancer Inst* 2007; *99* : 332-3.
3. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psychol* 2015; *37* : 1-2.
4. Wasserstein RL, Lazar NA. The ASA statement on *p*-values: Context, process and purpose. *Am Stat* 2016; *70* : 129-33.
5. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p<0.05$". *Am Stat* 2019; *73* : 1-19.
6. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019; *567* : 305-7.
7. Woolston C. Psychology journal bans *P* values. *Nature* 2015; *519* : 9.
8. Woolston C. Clarification. *Nature* 2017; *550* : 549-52.
9. Siegfried T. P-value ban: Small step for a journal, giant leap for science. *Sci News* 2015.
10. Berger VW. In defense of hypothesis testing. *BMJ* 2001; *322* : 1184.
11. Indrayan A, Malhotra RK. Confidence intervals, principles of tests of significance, and sample size. In: *Medical biostatistics*, 4th ed. UK/USA: Chapman & Hall/CRC Press; 2018. p. 353-421.
12. Indrayan A, Malhotra RK. Statistical fallacies. In: *Medical biostatistics*, 4th ed. UK/USA: Chapman & Hall/CRC Press; 2018. p. 811-49.
13. Dekkers OM, Groenwold RHH. [Significance of p-values: misinterpreted and overrated]. *Ned Tijdschr Geneeskd* 2018; *162* : D2161.
14. Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, *et al*. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev* 2018; *25* : 35-57.
15. Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An introduction to second-generation *p*-values. *Am Stat* 2019; *73* (Supp 1) : 157-67.
16. Domenech RJ. The uncertainties of statistical "significance". *Rev Med Chil* 2018; *146* : 1184-9.
17. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. *Am Stat* 2019; *73* (Suppl 1) : 235-45.
18. Calin-Jageman RJ, Cumming G. The new statistics for better science: Ask how much, how uncertain, and what else is known. *Am Stat* 2019; *73* : 271-80.
19. Brownstein NC, Louis TA, O'Hagan A, Pendergast J. The role of expert judgment in statistical inference and evidence-based decision-making. *Am Stat* 2019; *73* : 56-68.