Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2019;5(4): 226-234

Methodology

The Conundrum of *P*-Values: Statistical Significance is Unavoidable but Need Medical Significance Too

Abhaya Indrayan

Department of Clinical Research, Max Healthcare Institute, New Delhi, India.

ARTICLE INFO ABSTRACT

Received 05.05.2019
Revised 16.06.2019
Accepted 25.08.2019
Published 03.12.2019

Key words:
Empirical studies~~;~~
P-values ;
Sampling fluctuation;
Type-I error

Background: Small *P*-values have been conventionally considered as evidence to reject a null hypothesis in empirical studies. However, there is widespread criticism of *P*-values now and the threshold we use for statistical significance is questioned.

Methods: This communication is on contrarian view and explains why *P*-value and its threshold are still useful for ruling out sampling fluctuation as a source of the findings.

Results: The problem is not with *P*-values themselves but it is with their misuse, abuse, and over-use, including the dominant role they have assumed in empirical results. False results may be mostly because of errors in design, invalid data, inadequate analysis, inappropriate interpretation, accumulation of Type-I error, and selective reporting, and not because of *P*-values per se.

Conclusion: A threshold of *P*-values such as 0.05 for statistical significance is helpful in making a binary inference for practical application of the result. However, a lower threshold can be suggested to reduce the chance of false results. Also, the emphasis should be on detecting a medically significant effect and not zero effect.

Introduction

There is a raging controversy around the world on the use of *P*-values arising in null hypothesis significance testing (NHST) and consequent statistical significance that helps in obtaining empirical results. Although a discussion on their relevance in medical research was going on for long,(1,2) it precipitated by the decision of the editors of *Basic and Applied Social Psychology* in 2015 to ban *P*-values and consequent statistical significance in the articles published in their journal. In their opinion, *P*-values can become an excuse for lower quality research.(3) They found similar fault with the confidence intervals also. The American Statistical Association (ASA) subsequently set up a committee to examine this issue and recommended in 2016 that no decision should be based solely on *P*-value crossing a particular threshold such as 0.05.(4)This ASA statement attracted widespread media coverage, and reports of non-reproducibility of some statistically validated research findings, particularly in medical and health sciences(5,6) provided credence to these allegations. In the year 2019, *The American Statistician* brought out a Supplement with 43 articles on this topic, including the Editorial that concluded on the basis of consensus in these articles that the term ‘statistical significance’ be dropped entirely.(7) Also, an article in *Nature* cited a note with 800 signatures calling for a “stop to the use of *P*-values in the conventional dichotomous way to decide whether a result refutes and supports a scientific hypothesis”.(8)However, they stopped short of calling a ban on *P*-values. Thus, there are two distinct issues regarding the validity of *P*-values in reaching a result – first is regarding their any role whatsoever and second, if they have any role, regarding their threshold. We discuss both in this communication and present a contrarian view with reasons for our assertion that appropriately reported *P*-values and their threshold such as 0.05 are not only useful but are required for reaching an empirical result, particularly in medical research, at least for the time being.

Statisticians and researchers are aware of the limitation of the decisions based exclusively on *P*-values but the statements implying that *P*-values by themselves are of little value[7]and that no cutoff is appropriate[8] need a relook. Could this be a mistaken way of thinking? There is a need to find how many of statistically validated findings based on correct data and appropriate analysis have turned out false entirely due to *P*-values. Ioannidis(9) ascribes high false discovery rate to selection bias in reporting of results rather than to *P*-values themselves. Non-reproducibility of results could also be because of a variety of other reasons such as epistemic gaps, not able to capture all the known antecedents for an outcome, unstandardized instruments, errors in the data, and so forth, but the attention is focused on *P*-values. The problem possibly is not with *P*-values themselves but with their misuse, abuse, and over-use because of which these values are actually different from what are reported.(10) This can happen to increase the chance of publicationof the results(11) and because some researchers fall prey to incentives to make unsubstantiated claims.(12)

P-values are Unavoidable in Empirical Research

Perhaps nobody will disagree that the management of omnipresent medical uncertainties requires a powerful tool that can quantify them and thus helps in controlling their impact on our decisions. Probabilities quantify the uncertainties (rather certainties) and there is hardly any other tool to measure them. They do not necessarily lead us to the truth but make it likely to reach there. However, we must remind ourselves that probabilities work for groups and discount individuals. We all know that it is not necessary that a treatment found sufficiently effective in a perfectly executed clinical trial would work in each case.

The purpose of *P*-values is to measure the uncertainties generated by the sampling of subjects for a generalized result for the concerned population. All empirical studies are based on samples and samples include only a fraction of the existing cases and none of the future subjects. The implication of results, particularly in medical research, is not just for the existing target population but generally also for future cases of similar nature. Thus, a probabilistic statement of results is imperative. There is no way to avoid probabilities in a sample-based study setup, and there is no way to avoid samples because of likely implication for future subjects. Science demands careful presentation of evidence to challenge the existing status and *P*-values help to decide in favor or against a hypothesis as explained next. They are not proof, though.

P-value is the probability of obtaining sample values as extreme as observed from the population for which the null hypothesis is true. Thus, this measures the consistency of the sample values with the null where the null generally is of no effect. A small *P*-value (the threshold for this is discussed later) is considered evidence that the sample is most likely inconsistent with the null hypothesis and thus the null is rejected. One of the main criticisms of the *P*-values is that it provides evidence against the null without telling us what they support.(13,14) They admittedly are for falsifying the null based on available data and not for validating the null. It is "neither sole nor dominant criterion to measure scientific value of a result (15) it never was – but is for assessing the role of sampling fluctuations. Empirical result for sample studies without *P*-value is possibly less scientific because the role of sampling fluctuation cannot be adequately studied with any other existing method. This system of inference is followed in several other setups without raising any question. For example, the same system is followed in court decisions in criminal cases where the *P*-value is compared with the probability of convicting an innocent.(16) The courts may not compute the probability of the evidence under the initial assumptions of no crime but strong evidence against this null helps to decide the case. The crime by the person must be established 'beyond reasonable doubt' for conviction and the accused is discharged in the case of insufficient evidence. The court decides that the evidence placed before it is sufficient or not to convict the accused, and any other evidence lying elsewhere is not considered. In the case of insufficient evidence, the discharge of the accused does not always mean that the person is innocent only that there is no sufficient evidence to punish the accused.

The nature of statistical decision based on *P*-values is similar but more exact because the likelihood is measured by a quantity. If the judicial system cannot be discredited, why discredit *P*-values? Both are based on the available evidence and both seek to examine the sufficiency of this evidence against the null. In the case of court judgment, there is a social convention to consider the person not convicted as innocent in the sense that s/he has not committed the crime but in scientific pursuits, we refrain from making such a sweeping statement and only say that the null is not rejected it is never accepted. Thus, abundant precaution is already taken in scientific inference based on *P*-values.

The court example convincingly establishes that the basic procedure of NHST is not invalid and may have to be accepted for empirical decisions at least till such time that a credible alternative emerges. As discussed later, there is none now. Thus, NHST will continue to stay for the time being despite a large number of authors expressing their strong reservation.(5,6,17) As mentioned earlier, a journal has rejected it outright.(3)However, the latest (2019) guidelines of New England Journal of Medicine say, “Despite the difficulties they pose, P values continue to have an important role in medical research and we do not believe that P values and significant tests should be eliminated altogether.”(18)An in-depth look at the articles expressing reservations indicates that the opposition is not so much against NHST and the resulting *P*-values but is against the dominant role they have assumed in much of empirical research reporting. For example, Szucz and Ioannidis(19)state, “NHST should no longer be a default, dominant statistical practice of all biomedical and psychological research”. The important role of other factors such as biological plausibility of the results, previous literature, corroborative evidence, and adequacy of the data has been rightly emphasized in reaching a conclusion,(7) but *P*-values help reach a data-based result. (A result in empirical research is mostly based on data whereas a conclusion is based on several other considerations as just mentioned.) The advice is to accept uncertainty and be modest in claims.(7)The call is to use *P*-values as one of the considerations and not a dominant consideration,(4) meaning thereby that *P*-values will stay, although with a diminished role.

Medical literature is full of warnings that *P*-values by themselves should not be used for decisions unless they are supported by corroborative evidence.(20,21) While this certainly is sane advice and should be adhered to as much as possible but real progress in science occurs in areas where previous knowledge is meager. Thus, unexpected but ‘statistically significant’ finding that does not have any justification at present need not be ignored. If the same finding is repeatedly seen in different settings, it seems prudent to believe it in the hope that biological explanation may emerge later. For example, the analysis of data of a large number of healthy persons may reveal that the ABO blood group distribution in a population is different ($P<0.01$) in males than in females. There is no prior reason to believe that this could be so in the concerned population, yet the finding provides a hypothesis for future investigation. Manning et al. (22) forwarded a hypothesis of skewed child sex ratio due to parental age gap, and Direful (23) observed that the age gap between spouses can affect their survival. An adequate biological explanation may not be immediately available also with a newly emerging disease where a particular signs-symptoms syndrome is observed to occur more than expected by chance and the causative agent is identified later. This happened with HIV/AIDS although no *P*-value was calculated in this case. Thus, it is not correct to say that *P*-values by themselves are of little value. They can be valuable in some cases for setting up the direction of research. However, they must be interpreted with abundant caution.

Cautions Required in Interpreting *P*-values

The principled use of statistical methods, particularly of *P*-values, is crucial. *P*-values quantify only the uncertainties generated by sampling and nothing else. They are calculated for random samples, mostly simple random samples, but many studies use convenience sampling (24) and come up with statistically significant results that have limited applicability, if at all. Informed consent and inclusion/exclusion criteria in medical research further restrict the applicability. The question one must ask, but possibly never asked, is how close we meet the requirements of computing *P*-values. Closer we are, better is the chance of obtaining a valid *P*-value. *P*-values are also mostly based on specified distribution such as Gaussian, which is simply assumed in many cases (25) without realizing the repercussion of its violation on the results. The results also assume that the data obtained are correct with no error and are valid for the stipulated results. This is too much to expect in some setups such as where the responses are based on an interview even the laboratory investigation findings are sometimes questionable.(26) In all these instances, the actual *P*-values are not what are reported. Minor violations are accepted without much concern, but they can have a butterfly effect in some situations. (27) The most serious challenge, however, is the epistemic uncertainties[16] since all studies are based on existing knowledge. For example, a study on risk factors of an outcome will have to be necessarily based on what can be conjectured. Our knowledge is far too inadequate in most scientific endeavors and contributes to chance. Randomization in clinical trials and random selection in all setups are supposed to take care of unknown factors but these methods work in the long run and may fail in individual studies. How does this affect the *P*-value is seldom discussed. On top of this is the selective reporting as highlighted by Ioannidis(9) results with higher *P*-values tend to be suppressed causing skewed reporting toward ‘positive’ findings.*P*-hacking is another malpractice that ails the current research.(28)

Bias, either due to unaccounted confounders or because of intentional and unintentional prejudice in the collection, recording, analysis, and interpretation of data, is another source that tends to make the *P*-values unrealistic. It can irreparably inflict the results but non-reproducibility is unnecessarily ascribed to *P*-values. The result is sometimes not properly adjusted for known confounders too due to intricacies involved in their elicitation and difficulties in assessment, and simplistic study is done instead that fails to provide the correct results. For example, the role of low-to-moderate prenatal alcohol exposure in child academic achievement and behavior is difficult to assess. (29) Missing values and errors in eliciting the information and recording sometimes go unnoticed in the best of setups. Instruments generally used for obtaining the data are sometimes not sufficiently equipped to provide valid measurements. The care required to avoid such errors may not have been used in an investigation and the burden is unnecessarily placed on our dear *P*-values. Unusually large sample size, as in data mining, can cause low *P*-values (30) when the null is no difference. Such over-powered studies bring in the question of medically significant effect as discussed later in this communication.

The Importance of a Threshold

There is a strong plea that the *P*-values should be reported on a continuous scale and not within or beyond a threshold such as 0.05. Almost all reputed journals now insist on reporting exact *P*-value and many follow it religiously. The ASA also seems to have buckled and suggested to abolish any threshold and corresponding statistical significance.(7) Whereas reporting of exact *P*-values is welcome, a threshold seems unavoidable if a result is to be converted to a binary decision of yes or no – whether to switch to a new treatment modality or not, a particular factor is to be considered a risk or not for a defined outcome, and such other binary decisions. It does transform probability to an inferential statement as required for practical application of the results without making it alchemy to transform it to deterministic result – the result remains probabilistic. It is right to insist that such a binary decision should depend on multiple factors such as biological plausibility, cost and convenience, possibility of side-effects, and alternatives available to alleviate the suffering, and most importantly the effect size and the error in its estimate, but *P*-values also are needed to rule out the role of sampling fluctuations in reaching to a result. This is especially so for a medical research setup where variations and uncertainties are predominant. A threshold is needed for this paradigm although it may not have to be 0.05 all the time. It could vary from study to study and within a study from one measurement to another depending on the seriousness of the consequences. It can be argued that when all other considerations are favorable to a result, a small *P*-value is a good support without worrying about its specific threshold. But that reduces objectivity and compromises this essential feature of science. (30) Dispensing with a cut-off altogether will increase subjectivity because of varying interpretations and, thus, may have a deleterious effect on science. A cut-off of 0.05 has helped to be uniform in our approach for comparability across studies and in reducing subjectivity. A more stringent cut-off such as 0.01 can be suggested to reduce false positivity but a cut-off 0.005 as suggested by Benjamin et al. (32) may be too stringent considering that it is only for sampling fluctuations.

A researcher can be advised to be flexible in the case of *P*-values but there is no denying the role of a threshold in the practice of medicine. For example, the thresholds of blood pressure ≥140/90 mmHg for diagnosing hypertension and fasting plasma glucose level of ≥126mg/dl for diabetes have a defining role. Agreed that these thresholds are based on consensus for the prognostic implications while 0.01 is arbitrary and any other threshold will also be equally arbitrary, but consensus can be developed in this case also based on consequence in different setups. Many medical thresholds too are arbitrary. For example, the hemoglobin threshold for anemia varies from researcher to researcher and they are rarely questioned. The normal values of all medical parameters provide a reference interval beyond which the values are considered unhealthy. These also are thresholds and one can argue that these too are arbitrary. No threshold is absolute, but it is useful in evaluating the condition of a patient and in deciding the course of treatment. Just as the borderline values of medical parameters need a cautious approach and additional attention, so do the borderline values of *P*.

Consider an example of comparison of two regimens of nutritional supplementation for increasing vitamin D level in patients with this deficiency. A sample of 80 homogenous persons meeting the pre-set inclusion and exclusion criterion was randomly allocated to the two regimens with 40 in each group and the patients and assessors both were blinded. All precautions were taken to get the correct data. The trial revealed that the average increase with regimen A was 8.6 ng/ml (SD = 3.18) and with regimen B was 13.8 ng/ml (SD = 4.30). Apparently, the increase with regimen B was higher. One can subjectively say that the mean difference of 5.2 ng/ml in this study is substantial and there is no need for a statistical test. This is an estimate of the effect size in this case. Suppose the previous literature on this issue is conflicting and no biological reason is known for one regimen to work better than the other. Nonetheless, many would like to know that this difference is not a fluke in this study and is likely to be present across such other samples. Some others may like to know that an average difference of more than 3 ng/ml is most likely present or not. Let this be the medically important effect. Both require a statistical test. To validate the statistical requirement for this test, the increases in vitamin D level in subjects in each group were separately plotted and the distribution looked like a Gaussian. Anderson-Darling test for Gaussianity revealed $P = 0.72$ for regimen A and $P = 0.11$ for regimen B. This test, for that matter any statistical test, would only say that there is no evidence of a violation of the assumption of Gaussianity, but would not say that the distribution is indeed Gaussian. The same may happen with Leven test for equality of variances as required for a two-sample Student *t*-test. Suppose the Student *t*-test gives $P< 0.001$ for equality of change on average and 0.0055 for the post-hoc null of mean difference ≤ 3 ng/ml for one-tailed alternative: mean difference > 3 ng/ml.(Many would question such a post-hoc test. This test can be done by conducting another study.) Under the validity conditions, the former is the probability that the sample values are consistent with the null of no difference on average in the regimens and the latter that it is consistent with the null of mean difference = 3ng/dl. ASA statement does not forbid the use of *P*-value but forbids using any threshold such as 0.05. According to this statement, we can not say that the difference between the regimens on average is ‘statistically significant’, and, because of lack of corroborative evidence, we cannot reach the conclusion that one regimen is different from the other in effecting a change. Can we accept that the mean difference is likely to be more than 3 ng/ml in repeated samples. Where do we go from here?

There is another statistical need of a threshold in, say, selection of variables, as in the statistical stepwise procedure, when too many candidates are available with little background information on their relative importance. Consider an example of diagnosis of hypothyroidism based on clinical signs-symptoms alone in areas where laboratory facilities are not available. More than 100 clinical features can be considered in this setup and selection of a few important ones for a particular age-sex-ethnic group requires such selection. In the absence of prior information, *P*-values with a specific threshold are used to decide which variables can be useful for that group. Ability to detect a medically significant effect substantially depends on statistical power for that effect size but that too, in turn, depends to some extent on the pre-specified threshold of *P*-value.

It seems essential to have a threshold up to which a *P*-value will be tolerated for a binary decision. Those now preferring to abolish ‘statistical significance’ may have an answer of how to reach to a result without this threshold. Their response is briefly discussed in a later section of this communication. The objective in science is to reach the truth and there are competing views on how to get there. In any case, a threshold of *P*-value implies that a small percentage of results can have false significance, but a larger percentage of results may fail due to a variety of reasons as stated earlier. Another one is the accumulation of Type-I error.

Accumulation of Type-I Error

Amidst this conundrum regarding *P*-values and statistical significance, there is another serious problem that has not received much attention. The Type-I error accumulates and builds up due to repeated and multiple uses of *P*-values in the same research. Perhaps this is the root cause of the non-reproducibility of some statistically significant research findings despite being based on a perfectly executed study.

We all know that ‘multiple comparisons’ can have an enormous deleterious effect on the credibility of the results. (33, 34) Our example in the previous section used *P*-values at three places without adjustment for multiple uses. This is a regular practice and many articles can be cited (35, 36) that used multiple *P*-values. What seems to have escaped attention is that many investigations are based on previous results, which themselves are subject to such Type-I error. For example, see Patel et al. (30) and Alves and Yu (37). Generally no adjustment for such double counting is done and the actual probability of Type-I error many times becomes much beyond the threshold without us realizing that this has happened. Unsurprisingly the results fail to replicate. This amounts to building error on error and is an example of the over-use of *P*-values. Such accumulation of Type-I error due to multiple and repeated use of *P*-values is obviously a threat to the statistical validity of the results.

Statistical Significance versus Medical Significance

A consensus seems to be developing a couple of decades ago that the researchers report the effect size in place of its statistical significance. This implies that we shift from “Is there an effect?” to “How much is the effect?” The latter assumes some effect will always be there – it could be tiny or enormous. This looks plausible because any two populations would be different although the difference can be negligible. The same can be stated for any other measure of effect. Thus, there are two aspects of this debate: first is substituting *P*-values with confidence intervals (CI), and second is the medical significance of the effect size. The CI with a 95% confidence level is essentially the same as the threshold 0.05 for *P*-value and suffers from the same criticism. Many statisticians and research workers preferred the CI approach because of its emphasis on the effect size instead of finding that any effect is present or not. However, the quantitative effect size too is an average obtained in a sample and the estimation is based on validity assumptions such as random sampling and symmetric unimodal distribution– mostly Gaussian. Sampling fluctuations would never allow us to be fully confident, and the possibility of not fully meeting the underlying assumptions adds to the spectrum of uncertainty. An effect size of clinically significant magnitude can still arise due to sampling fluctuation. Thus, this approach too is not infallible.

The second point regarding the medical significance of the effect size is more relevant. This also is mostly subjectively determined. A researcher may consider an effect size of 5 mmHg of reduction in average systolic blood pressure after medication as medically important but another may consider it trivial. Some researchers may prefer to use the percentage of cases that were benefitted instead of the average improvement in a quantitative medical parameter. The definition of 'benefit' in this case could vary from researcher to researcher. Thus, subjectivity is not ruled out in this case also although the result is much more definite regarding the presence of the specified medically significant effect.

Other Alternatives to *P*-Values

Besides the effect size and CI, Nuzzo(5) has proposed several alternatives to P-values. One is the use of Bayes' rule for assessing the plausibility of a hypothesis for the sample values in hand but that entails a certain amount of subjectivity for imputing unknown probabilities. Second is to try to analyze the data in multiple ways and hope that most lead to the same result. An Editorial in Nature also suggests that we should try to answer the question in many ways and develop consensus.(38)This can always be recommended. The third is the two-stage analysis that requires separation of confirmatory analysis from exploratory analysis and publish both together as complements. Fourth is a call to use scientific judgment about the plausibility of hypothesis and consider clinical knowledge, previous results, and the possible mechanism to reach to a conclusion. Various articles in the 2019 Supplement of The American Statistician also made similar suggestions. Most promising of these is second generation P-values (39) that requires setting up the null hypothesis in terms of the range of trivial effect, but the problem with this is that the definition of trivial effect can vary from specialist to specialist. The same is true for equivalence testing. (40) Nevertheless, all these have merit and deserve a fair trial. The basic problem is that these proposals are new and evolving. Different authors and journals may prefer different procedures as per their understanding. It will take a while before a consensus develops. Also, there is no evidence yet that such alternatives would substantially reduce the false results. (9) It has taken decades to realize the fault with P-values but the new consensus method may be able to establish its credentials soon due to a better understanding of issues. The time will tell.

Conclusion

Till the time a consensus on credible alternative emerges, the *P*-values must be allowed to stay albeit with less role than they have been playing in empirical results so far. A threshold is helpful if a result is to lead to a binary decision. This can, however, vary from problem to problem although a fixed cut-off is a big help in being uniform in our approach and not being subjective. A debate should be initiated to develop a consensus regarding appropriate cut-off to effectively rule out the sampling fluctuation as an explanation of the results. If not 0.05, can it be 0.01? The null hypothesis can be shifted from zero effect to a medically significant effect with an explicit specification of what is considered medically significant and why. Arguments to abrogate *P*-values seem away from reality emotions do alter our perception of reality. (41)Everybody seems to be riding on bandwagon, validating the paradigm that most of us are inclined to be in harmony with the current value system and to ignore the contrarian view.(42)The fault can be with data, sampling, design, analysis, interpretation, and selective reporting but rarely with *P*-values per se.

References:

1.McGough JJ, Faraone SV. Estimating the size of treatment effects: Moving beyond p values. Psychiatry (Edgmont) 2009;6(10):21–29.https://www.ncbi.nlm.nih.gov/pubmed/20011465

2.Hubbard R, Lindsay RM. Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol 2008;18(1):69–88. https://doi.org/10.1177/0959354307086923

3.Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology 2015;37;1–2. https://www.researchgate.net/publication/304150529_Editorial

4.Wasserstein RL Lazar NA. The ASA statement on p-values: Context, process and purpose. Am Stat 2016;70:129–133.https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

5.Nuzzo R. Scientific method: Statistical errors. Nature 2014;506:152–156. https://www.nature.com/news/scientific-method-statistical-errors-1.14700

6.Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/

7.Wasserstein RL,Schirm AL, Lazar NA. Moving to a world beyond "p<0.05". The Am Stat 2019;73(Sup1):1–19. https://doi.org/10.1080/00031305.2019.1583913

8.Amrhein V, Greenland S, Mc Shane B. Scientistsrise up against statistical significance. Nature 2019;567:305–307. https://www.nature.com/articles/d41586-019-00857-9

9.Ioannidis JPA. What have we (not) learnt from millions of scientific papers with P-values? Am Stat 2019;73(Sup1):20–25. https://www.tandfonline.com/doi/pdf/10.1080/00031305.2018.1447512?needAccess=true

10.Gagnier J, Morgenstern H. Misconception, misuses and misinterpretation of P-value and significance testing. J Bone Joint Surg 2017;99(18):1598–1603.https://insights.ovid.com/pubmed?pmid=28926390

11.Young NS, Ioannidis JPA, Al-Ubaydli O. Why current publication practices may distort science. PLoS Med 2008;5(10):e201. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0050201

12.Gelman A. Ethics in statistical practice and communication: Five recommendations. Significance 2018 (October); 37:40-43.http://www.stat.columbia.edu/~gelman/research/published/SIGN_15(5)_09_InPractice_Gelman_EthicsAndComm.pdf

13.Nahm FS. What the P values really tell us. Korean J Pain. 2017;30(4):241–242. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5665734/

14.Cohen HW. P-Values: Use and misuse in medical literature. Am J Hypert. 2011;24:18–23. https://academic.oup.com/ajh/article/24/1/18/165807

15.Wei YY. Statistical P-values do not dominate scientific research.Europmc2019;53(5):441–444. https://europepmc.org/abstract/med/31091597

16.Indrayan A, Malhotra RK. Medical Biostatistics, Fourth Edition. CRC Press, 2018.

17.Sullivan LM, Weinberg J, Keaney JF Jr. Common statistical pitfalls in basic science research. J Am Heart Assoc. 2016;5(10):e004142. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5121512/

18.Harrington D, D'Agostino RB, Sr., Gatsonis C, Hogan JW, Hunter DJ, Normand ST, Drazen JM, M.D., Hamel MB.New guidelines for statistical reporting in the Journal (Editorial). N Engl J Med 2019; 381:285-286. DOI: 10.1056/NEJMe1906559.https://www.nejm.org/doi/full/10.1056/NEJMe1906559

19.Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: A reassessment. Front Hum Neurosci. 2017;11:390. https://www.ncbi.nlm.nih.gov/pubmed/28824397

20.Lytsy P. P in the right place: Revisiting the evidential value of P-values. J Evid Based Med 2018;11:288–291.https://onlinelibrary.wiley.com/doi/full/10.1111/jebm.12319

21.Concato J, Hartigan JA. P values: From suggestion to superstition. J Investig Med2016;64:1166–1171.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5099183/

22.Manning JT, Anderson RH, Shutt M. Parental age gap skews child sex ratio. Nature 1997;389:344. https://www.nature.com/articles/38647

23.Drefahl S. How does the age gap between partners affect their survival? Demography 2010;47:313–326.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000022/

24.Nahar VK. Using the multitheory model to predict initiation and sustenance of physical activity behavior among osteopathic medical students J Am Osteopath Assoc 2019;119:479–487.https://jaoa.org/article.aspx?articleid=2739371

25.Foster K, Younger N, Aiken W, Brady-West D, Delgoda R. Reliance on medicinal plant therapy among cancer patients in Jamaica. Cancer Causes & Control 2017;28:1349–1356. https://link.springer.com/article/10.1007%2Fs10552-017-0924-9

26.Plebani M. Errors in clinical laboratories or errors in laboratory medicine?Clin Chem and Lab Med 2006;44:750–759.https://www.degruyter.com/view/j/cclm.2006.44.issue-6/cclm.2006.123/cclm.2006.123.xml

27.Indrayan A, Holt MP. Concise Encyclopedia of Biostatistics for Medical Professionals. CRC Press, 2016.

28.Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol2015;13:e1002106. doi: 10.1371/journal.pbio.1002106. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4359000/

29.Jacobson SW, Jacobson JL. The risk of low-to-moderate prenatal alcohol exposure on child academic underachievement and behaviour may be difficult to measure and should not be underestimated. Evid Based Med 2014;19:e7. doi:10.1136/eb-2013-101535. https://ebm.bmj.com/content/19/2/e7.long

30.Patel CJ, Ji J, Sundquist J, Ioannidis JPA, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: A method to conduct a population-wide medication-wide longitudinal study. Scientific Reports 2016;6: 31308.https://www.nature.com/articles/srep31308

31.Padovani F, Richardson A, Tsou JY (Editors). Objectivity in Science: New Perspective from Science and Technological Studies, Springer, 2015.

32.Benjamin DJ. Redefine statistical significance. Nature Human Behaviour 2017;2: 6–10.https://www.nature.com/articles/s41562-017-0189-z

33.Victor A, Elsässer A, Hommel G, Blettner M. Judging a plethora of p-values: How to contend with the problem of multiple testing--part 10 of a series on evaluation of scientific publications. DtschArztebl Int 2010;107(4):50–56. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2822959/

34.Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002;2:8.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC117123/

35.Vega JF, Strnad GJ, BenaJ, Spindler KP. Predicting the need for surgical intervention prior to first encounter for individuals with knee complaints: A novel approach. Orthop J Sports Med 2019;7(7):2325967119859485. Published 2019 Jul 25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6659191/

.36.Koratala A, Dass B, Alquadan KF, Sharma S, Singhania G, Ejaz AA. Static pressures, intra-access blood flow and dynamic Kt/V profiles in the prediction of dialysis access function. World J Nephrol 2019;8(3):59-66. https://www.wjgnet.com/2220-6124/full/v8/i3/59.htm

37.Alves G, Yu YK. Accuracy evaluation of the unified P-value from combining correlated P-values. PLoS One. 2014;9(3):e91225.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3963868/

38.Editorial. Significant debate: Looking beyond statistical significance would make science harder, but might help to avoid false positives, overhyped claims and overlooked effects. Nature 2019;567:283. https://www.nature.com/magazine-assets/d41586-019-00874-8/d41586-019-00874-8.pdf

39.Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD (2019) An introduction to second-generation p-values, Am
Stat 2019;73(Sup1):157-
167. DOI: 10.1080/00031305.2018.1537893. https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1537893

40.Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science 2018;1: 259–269. https://doi.org/10.1177/2515245918770963.https://journals.sagepub.com/doi/pdf/10.1177/2515245918770963

41.Duffy B. The Perils of Perception: Why we are Wrong About Nearly Everything Atlantic Books, 2018.

42.Kahan DM, Wittlin M, Peters E et al. The Tragedy of the Risk-Perception Commons: Culture Conflict, Rationality Conflict, and Climate Change. Temple University Legal Studies Research Paper No. 2011-26; Cultural Cognition Project Working Paper No. 89; Yale Law & Economics Research Paper No. 435; Yale Law School, Public Law Working Paper No. 230. Disponibileall'indirizzo: https://ssrn.com/abstract=1871503