

The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives

Chittaranjan Andrade

ABSTRACT

The calculation of a P value in research and especially the use of a threshold to declare the statistical significance of the P value have both been challenged in recent years. There are at least two important reasons for this challenge: research data contain much more meaning than is summarized in a P value and its statistical significance, and these two concepts are frequently misunderstood and consequently inappropriately interpreted. This article considers why 5% may be set as a reasonable cut-off for statistical significance, explains the correct interpretation of $P < 0.05$ and other values of P , examines arguments for and against the concept of statistical significance, and suggests other and better ways for analyzing data and for presenting, interpreting, and discussing the results.

Key words: *Compatibility interval, confidence interval, P value, statistical significance*

In empirical research, statistical procedures are applied to the data to identify a signal through the noise and to draw inferences from the data collected. Statistical procedures, therefore, steer us toward a better understanding of the data and toward drawing conclusions from the data. It is therefore important to fully understand what statistical procedures and their results mean when these procedures are applied in research.


All inferential statistical tests end with a test statistic and the associated P value. This P value has been accorded such an elevated status that, now, everybody

who performs or reads research is familiar with the expression " $P < 0.05$ " as a cut-off that indicates "statistical significance." In this context, most persons interpret $P < 0.05$ to mean that "the probability that chance is responsible for the finding is less than 5%" and that "the probability that the finding is a true finding is more than 95%." Both these interpretations are incorrect; unfortunately, they are widely prevalent because they are an easy way to explain and understand a slightly tricky concept.

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Andrade C. The P value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian J Psychol Med* 2019;41:210-5.

Access this article online	
Website: www.ijpm.info	Quick Response Code 
DOI: 10.4103/IJPSYM.IJPSYM_193_19	

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bengaluru, Karnataka, India

Address for correspondence: Dr. Chittaranjan Andrade

Department of Psychopharmacology, National Institute of Mental Health and Neurosciences, Bengaluru - 560 029, Karnataka, India.

E-mail: andrade@gmail.com

Received: 19th April, 2019, **Accepted:** 19th April, 2019

This article considers why 5% could be a reasonable cut-off for statistical significance, explains what $P < 0.05$ really means, discusses the concept of statistical significance and why it has been roundly criticized, and suggests other and perhaps better ways of interpreting the results of statistical testing.

WHY 5%?

Imagine that you toss a coin and it falls tails. Then you toss it again, and it falls tails again. Well, that can certainly happen. You toss it a third time, and it falls tails again. This, too, can sometimes happen; the same face shows thrice in a row. When you toss it a fourth time, and it falls tails, you sit up and take notice. And when you toss it a fifth time, and it falls tails yet again, you develop a strong suspicion that there is something wrong with the coin.^[1] Why? Theoretically, if you toss an unbiased coin in runs of five for several dozen trials, a run of five identical faces can certainly happen by chance. However, you did not toss the coin in dozens of trials. You tossed it in just one trial. You found that the coin showed the same face on all five occasions in that one trial. In other words, something that should have been a rather rare occurrence happened the very first time. This suggests that at least for that coin, it may not have been a rare occurrence, after all. In other words, you consider that your finding is significant. That is, you reject the null hypothesis that the coin is unbiased and accept an alternate hypothesis – that the coin is biased.

Simple mathematics tells us that the probability that a tossed coin will display the same face (heads or tails) five times in a row is $0.5 \times 0.5 \times 0.5 \times 0.5$; that is, 0.0625. This P value, 0.0625, is rather close to the value 0.05 that is by general convention set as the cut-off for “statistical significance.”

A slightly more scientific explanation for choosing 5% as the cut-off is that approximately 5% (4.5%, to be more precise) of the normal distribution comprises outlying or “significantly different” values, that is, values that are more than two standard deviations distant from the mean. Other explanations have also been offered.^[1]

WHAT DOES $P < 0.05$ REALLY MEAN?

Imagine that you conduct a randomized controlled trial (RCT) that compares a new antidepressant drug with placebo. At the 8-week study endpoint, you find that 60% of patients have responded to the drug and 40% have responded to placebo. The Chi-square test that you apply yields a P value of 0.04, a value that is less than 0.05. You conclude that significantly more patients responded to the antidepressant than to placebo. Your interpretation is that the new antidepressant drug truly

has an antidepressant effect. The conclusion is correct but iffy because the 5% cut-off and even the concept of statistical significance are being challenged. The interpretation is wrong because a P value, even one that is statistically significant, does not determine truth.

So, what are the right conclusion and the right interpretation? This requires an understanding of what statistical testing means.^[2] Imagine that the null hypothesis is true; that is, the new antidepressant is no different from placebo. Now, if you conduct a hundred RCTs that compare the drug with placebo, you would certainly not get an identical response rate for drug and placebo in each RCT. Rather, in some RCTs, the drug would outperform placebo, and in other RCTs, placebo would outperform the drug. Furthermore, the magnitude by which the drug and placebo outperformed each other would vary from trial to trial. In this context, what $P = 0.04$ (i.e., 4%) means is that *if the null hypothesis is true and if you perform the study a large number of times and in exactly the same manner, drawing random samples from the population on each occasion, then, on 4% of occasions, you would get the same or greater difference between groups than what you obtained on this one occasion.*

However, you did not perform the RCT a large number of times. You performed it just once. You found that on the *single* occasion that you performed the RCT, the result that you obtained was something that would be considered rare. So, perhaps the finding is not really rare. This is possible only if the null hypothesis is false. Therefore, just as you rejected the null hypothesis that the tossed coin was unbiased (see the previous section), you reject the null hypothesis that the drug is no different from placebo. Because this (correct) reasoning is rather complicated, many prefer to explain and understand the concept in simpler but incorrect ways, as stated in the introductory paragraph to this article. Other incorrect interpretations have also been described.^[3]

INTERPRETATIONS FOR $P < 0.05$ AND $P > 0.05$

If the null hypothesis is rejected ($P < 0.05$), why cannot we conclude that just as the drug outperformed placebo in our study, the drug is truly superior to placebo in the population from which the sample was drawn? The answer is that the P value describes a probability, not a certainty. So, we can never be certain that the drug is truly superior to placebo in the population; we can merely be rather confident about it.

Next, imagine that instead of obtaining $P = 0.04$, you obtained $P = 0.14$ in the imaginary RCT described earlier. In this situation, we do not reject the null

hypothesis, based on the 5% threshold. So, can we conclude that the drug is no different from placebo? Certainly not, and we definitely cannot conclude that the drug is similar to placebo, either. After all, we did find that there was a definite difference in the response rate between drug and placebo; *it is just that this difference did not meet our arbitrary cut-off for statistical significance.* So “not significantly different” does not mean “not different from” or “similar.”

WHY IT COULD BE NECESSARY TO STOP USING A THRESHOLD FOR STATISTICAL SIGNIFICANCE

From the previous section, it is quite clear that *just as the *P* value lies along a continuum of 0 to 1, our interpretations should also lie along a continuum of differing levels of confidence (or diffidence) in the null hypothesis; we can never be certain, either way.* This means that the *P* value should be reported as an exact value and should be regarded as a continuous variable. Consequently, it should be considered fallacious to insert an arbitrary threshold to define results as significant or nonsignificant, as though significant versus nonsignificant results are in some ways categorically different the way people who are dead versus alive are categorically different. Expressed otherwise, declaring statistical significance does not improve our understanding of the data over and above what is already explained by the value of *P*.^[4] In fact, declaring significance may give us a false sense of confidence that a finding exists in the population, while rejecting significance may give us a false sense of confidence that the finding does not exist.

It follows, therefore, that it is fallacious to privilege significant results for journal publication or for media dissemination. Finally, the probability continuum is also the reason why a study which obtains a nonsignificant result *does not contradict* a study which obtains a significant result; both obtained findings that lie along a continuum, and the contradiction exists only because the findings lie on the opposite sides of an arbitrary and imaginary fence, $P < 0.05$, that we insert into this continuum. Bayesian methods are no exception to these assertions.^[5]

THE 95% CONFIDENCE INTERVAL

Imagine an RCT in which 10 of 20 patients responded to a new antidepressant drug and 11 of 22 patients responded to placebo. The response rate is exactly 50% in each group. The difference in response rates is 0%. Whatever statistical test is applied, the *P* value will be 1.00. Does this mean that we are 100% certain that

there is no difference between drug and placebo? No! What $P = 1.00$ means is that if the null hypothesis is true and if we perform the study in an identical manner a large number of times, then on 100% of occasions we will obtain a difference between groups of 0% or greater! This is actually common sense. If the drug truly has no antidepressant effect, then on some occasions the drug will outperform placebo by some margin, on other occasions placebo will outperform the drug by some margin, and perhaps on some occasions the results will be identical in the two groups; that is, on all (100%) occasions we obtain a difference between groups of 0% or greater.

This brings us to a question: if everything boils down to repeating the study a large number of times and getting different answers each time, can we reduce the range of uncertainty to something that could actually be helpful? Here is where 95% confidence intervals (CI) come into the picture. Means, differences between means, proportions, differences between proportions, relative risks (RRs), odds ratios, numbers needed to treat, numbers needed to harm, and other statistics that are obtained from a study are accurate only for that study. However, what we really want to know is what the values of these statistics are in the population, because we wish to generalize the results of our study to the population from which our sample was drawn. We cannot know for certain what the population values are because it is (usually) impossible to study the entire population. However, the 95% CI can help give us an idea. Whereas the 95% CI, like the *P* value, is also frequently misunderstood; here is an explanation. If we repeat a study in an identical fashion a hundred times, then 95 of the 95% CIs that we estimate in these studies would be expected to contain the population mean. So, by inference, if we examine the 95% CI that we have obtained from a single study, the probability that this particular CI contains the population mean is 95%.^[6]

In the RCT example cited earlier in this section, the response rate was 50% in each group; that is, there was no difference in the response rate between the drug and placebo. A little calculation will tell us that the RR for response is 1.00 and that the 95% CI is 0.55-1.83. That is, we are 95% confident that the population result for the response to drug versus placebo lies within the range of the drug being as much as 45% inferior to placebo to as much as 83% superior to placebo. Notice that there is no need whatsoever to bring statistical significance into the picture here. Also notice that the 95% CI provides a range of values that are possible for the population, which is far more informative than a dichotomous inference of significance versus nonsignificance.

UNCERTAINTY AND THE 95% COMPATIBILITY INTERVAL

Basing interpretations on a 0.05 or other threshold tends to provide an element of certainty to the interpretations. As already explained, this certainty is illusory because probability lies along a continuum. Furthermore, just as there are variations within a data set, there will be variations across replicatory studies, even across hypothetical replications. We can never be certain about which data set and which set of conclusions provide the best fit to the population. So, taking the discussion to its logical end, Amrhein *et al.*^[5] and Wasserstein *et al.*^[4] suggested that instead of drawing dichotomous conclusions that imply certainty, scientists should embrace uncertainty.

In this context, as one possible solution, Amrhein *et al.*^[5] offered the suggestion of reconceptualizing 95% CI as compatibility intervals. That is, all values within the 95% CI are compatible with the data recorded in the study; the point estimate (e.g., a mean or a RR), *regardless of “statistical significance,”* is the most compatible, and other values in the CI are progressively less compatible (but nevertheless still compatible) the greater their distance from the point estimate. Explained somewhat simplistically, this means that (provided the study was well-designed, well-conducted, and well-analyzed) the point estimate obtained in the study has the best chance of being the population value, and that all the other values in the 95% CI also have a chance of being the population value, with progressively decreasing likelihood the greater the distance from the point estimate.

Explained with the help of an example, consider the RCT in which we found that the RR for a response to the study drug (vs. placebo) was 1.00 (95% CI, 0.55-1.83). We should not interpret this finding as nonsignificant; rather, we should consider that the most likely interpretation is that the drug is no better or worse than placebo, and that lower efficacy (to the most extreme and least likely value of 45% worse) and higher efficacy (to the most extreme and least likely value of 83% better) possibilities are also compatible with the data recorded in the study. The reader is once again reminded that statistical significance does not enter the picture anywhere.

If the 95% CI for an RR is 0.95–2.20, the traditional interpretation would have been “not significant,” but a better interpretation would be that the results are mostly compatible with an increase in risk. Similarly, if the 95% CI for an RR is 0.65–1.05, the traditional interpretation would again have been “not significant,” but the better interpretation is that the results are mostly compatible with a decrease in risk. In this regard,

Amrhein *et al.*^[5] remind readers that even a 95% CI describes probabilities; it does not exclude the possibility that the population value lies outside the compatibility range. It must also be remembered that the 95% CI is an estimate; it is not a definitive statement of where the population parameter probably lies.

NO TO *P* AND NO TO A THRESHOLD FOR STATISTICAL SIGNIFICANCE

P values and the concept of statistical significance have been questioned for long.^[7] In 2016, the American Statistical Association (ASA) released a statement on statistical significance and *P* values.^[8] The statement asserted that *P* values were never intended to substitute for scientific reasoning. The statement highlighted six points: (1) *P* values can provide an indication of how compatible or incompatible the data are with a specified statistical model. (2) Taken alone, the *P* value is not a good test of a hypothesis or a good evaluation of a model. (3) *P* values do not estimate the probability that a hypothesis is true or the probability that chance is responsible for the findings. (4) *P* values, including those that meet arbitrary criteria for statistical significance, do not indicate an effect size or the importance of a result. (5) scientific conclusions and decision-making should not be based only on whether or not the *P* value falls below an arbitrary threshold; and (6) drawing proper inferences requires complete reporting and transparency. The ASA added that other statistical estimates, such as CIs, need to be included; and that Bayesian approaches need to be used, and false discovery rates need to be considered. Some of these points have already been explained; the rest are out of the scope of this article, and the reader is referred to the original statement.

Doing away with *P* and a threshold for statistical significance will, however, be hard. This is because estimating *P* and declaring statistical significance (or its absence) has become the cornerstone of empirical research, and if changes are to be made herein, textbooks, the education system, scientists, funding organizations, and scientific journals will all need to make a sea change. This could take years or decades if indeed it ever happens. The motivation to effect the change will be small, because *P* values are easy to calculate and use, alternatives are not easy to either understand or use, and, besides, there is no consensus on what the alternatives must be.^[4]

IN FAVOR OF RETAINING DICHOTOMOUS DISTINCTIONS

There is a small but definite role for the retention of the $P < 0.05$ threshold for statistical significance.

Dichotomous interpretations of research findings need to be made when action is called for, such as whether or not to approve a drug for marketing.^[9] Preset rules are required in such situations; uncertainty, recommended by Armhem *et al.*,^[5] cannot be embraced because, then, no decision would be possible. In such circumstances, study findings will need to meet or exceed expectations, and so a threshold for statistical significance needs to be retained. However, to protect the integrity of science and reduce false-positive findings, there may be a case to set the bar higher, such as at $P < 0.005$.^[10] In fact, in genetics research, reduction in the false-positive risk is achieved by setting the bar very high, such as at $P < 0.00000001$ or lower. If a threshold for significance were to be completely discarded, as many now demand, then there is a risk that study results will be interpreted in ways that suit the user's interest; that is, bias will receive a free pass.^[11] Setting a threshold for P is also necessary for sample size estimation and power calculations.

There are other circumstances, too, when a threshold for P may be required. An example is for industry quality control, or for risk tolerance. Consider a man who uses a parachute; he would like to be far more than 95% certain that the parachute will open.^[1] Thresholds will also be required as a filter when choosing variables for further investigation, as in brain imaging or genome analyses.^[4]

RECOMMENDATIONS

The P value should be interpreted as a continuous variable and not in a dichotomous way. So, we should not conclude that just because the P value is < 0.05 or some other predetermined threshold, the study hypothesis is true. Likewise, we should not say that just because $P > 0.05$ or some other predetermined threshold, the study hypothesis is false. These are, in any case, wrong interpretations of what the P value means.

Whereas a threshold for statistical significance could be useful to base decisions upon, its limitations should be recognized. It may be wise to set a threshold that is lower than 0.05 and to examine the false-positive rate associated with the study findings. It is also important to examine whether what has been accepted as statistically significant is clinically significant.

Examining a single estimate and the associated P value is insufficient. It is necessary to assess as much as possible about the estimate. Besides absolute values, 95% CIs should be examined as compatibility intervals, and the precision of this interval should be considered. Measures of effect size, such as standardized mean

deviation, RR, and numbers needed to treat, and the confidence (compatibility) intervals associated with these measures of effect size should also be reported.

All findings should be interpreted in the context of the study design, including the nature of the sample, the sample size, the reliability and validity of the instruments used, and the rigor with which the study was conducted.

FURTHER READING

Readers who are enthusiastic may refer to a special supplement of the *American Statistician*, published in 2019, titled "Statistical Inference in the 21st Century: A World Beyond $P < 0.05$." This issue contains 43 articles on the subject, some of which are technical but many of which are understandable to the average medical scientist. Whereas the concepts of P and statistical significance are not altogether rejected, and whereas there is no consensus on what the best alternative is, many proposals have been made. These include transforming P values into S-values, deriving second-generation P values, using an analysis of credibility, combining P values with a computed false-positive risk, combining sufficiently small P values with sufficiently large effect sizes, the use of a confidence index, the use of statistical decision theory, and, as already discussed, the use of compatibility intervals.

The articles in this special issue are arranged in five sections: Getting to a post " $P < 0.05$ " era; interpreting and using P ; supplementing or replacing P ; adopting more holistic approaches; and reforming institutions: changing publication policies and statistical education. The editorial in the special issue^[4] presents a useful summary of each article, provided by the authors of the articles.

Last but not least, readers are also strongly encouraged to consult the article by Goodman^[3] which lists 12 *misperceptions* about the P value. These are as follows: if the P value is 0.05, the null hypothesis has a 5% chance of being true; a nonsignificant P value means that (for example) there is no difference between groups; a statistically significant finding (P is below a predetermined threshold) is clinically important; studies that yield P values on opposite sides of 0.05 describe conflicting results; analyses that yield the same P value provide identical evidence against the null hypothesis; a P value of 0.05 means that the observed data would be obtained only 5% of the time if the null hypothesis were true; a P value of 0.05 and a P value less than or equal to 0.05 have the same meaning; P values are better written as inequalities, such as $P < 0.01$ when $P = 0.009$; a

P value of 0.05 means that if the null hypothesis is rejected, then there is only a 5% probability of a Type I error; when the threshold for statistical significance is set at 0.05, then the probability of a Type I error is 5%; a one-tail *P* value should be used when the researcher is uninterested in a result in one direction, or when a value in that direction is not possible; and scientific conclusions and treatment policies should be based on statistical significance.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCES

1. Gauvreau K, Pagano M. Why 5%? Nutrition 1994;10:93-4.
2. Kyriacou DN. The enduring evolution of the P-value. JAMA 2016;315:1113-5.
3. Goodman S. A dirty dozen: Twelve P-value misconceptions. Semin Hematol 2008;45:135-40.
4. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “ $p < 0.05$.” Am Stat 2019;73(Suppl. 1):1-19.
5. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305-7.
6. Andrade C. A primer on confidence intervals in psychopharmacology. J Clin Psychiatry 2015;76:e228-31.
7. Nuzzo R. Scientific method: Statistical errors. Nature 2014;506:150-2.
8. Wasserstein RL, Lazar NA. The ASA's statement on *P* values: Context, process, and purpose. Am Stat 2016;70:129-33.
9. Ioannidis JPA. The importance of predefined rules and prespecified statistical analyses: Do not abandon significance. JAMA 2019; Apr 4. doi: 10.1001/jama.2019.4582. [Epub ahead of print].
10. Ioannidis JPA. The proposal to lower P-value thresholds to 0.05. JAMA 2018;319:1429-30.
11. Ioannidis JPA. Retiring statistical significance would give bias a free pass. Nature 2019;567:461.

“Quick Response Code” link for full text articles

The journal issue has a unique new feature for reaching to the journal's website without typing a single letter. Each article on its first page has a “Quick Response Code”. Using any mobile or other hand-held device with camera and GPRS/other internet source, one can reach to the full text of that particular article on the journal's website. Start a QR-code reading software (see list of free applications from <http://tinyurl.com/yzlh2tc>) and point the camera to the QR-code printed in the journal. It will automatically take you to the HTML full text of that article. One can also use a desktop or laptop with web camera for similar functionality. See <http://tinyurl.com/2bw7fn3> or <http://tinyurl.com/3ysr3me> for the free applications.