MEDwave

Topics and controversies in biostatistics
Medwave 2016 Sep;16(8):e6534 doi: 10.5867/medwave.2016.08.6534

The questioned p value: clinical, practical and statistical significance

Author: Rosa Jiménez-Paneque **[1]**

Affiliation:
[1] Subeditora y revisora estadística, Medwave

E-mail: rosa.jimenez@medave.cl

Citation: Jiménez-Paneque R. The questioned p value: clinical, practical and statistical significance. *Medwave* 2016 Sep;16(8):e6534 doi: 10.5867/medwave.2016.08.6534
Publication date: 9/9/2016
Origin: not requested
Type of review: not peer reviewed

Abstract

The use of p-value and statistical significance have been questioned since the early 80s in the last century until today. Much has been discussed about it in the field of statistics and its applications, especially in Epidemiology and Public Health. As a matter of fact, the p-value and its equivalent, statistical significance, are difficult concepts to grasp for the many health professionals some way involved in research applied to their work areas. However, its meaning should be clear in intuitive terms although it is based on theoretical concepts of the field of Statistics. This paper attempts to present the p-value as a concept that applies to everyday life and therefore intuitively simple but whose proper use cannot be separated from theoretical and methodological elements of inherent complexity. The reasons behind the criticism received by the p-value and its isolated use are intuitively explained, mainly the need to demarcate statistical significance from clinical significance and some of the recommended remedies for these problems are approached as well. It finally refers to the current trend to vindicate the p-value appealing to the convenience of its use in certain situations and the recent statement of the American Statistical Association in this regard.

Introduction

Worldwide, the value of p (p-value,) has been and continues to be a concern for researchers who write and publishers of scientific articles. Even today, after 20 years of discussion, literature addresses this issue of statistics and its applications to research and the report of results [1],[2],[3].

On the other hand, experience shows that, despite its widespread use (and misuse), many researchers do not yet know what it means. "Statisticians find it" or "statistical computer programs give it" might be common answers to the question "what is the p-value?" One cannot know everything, if it is yet very difficult to know about a particular subject imagine to further know about Statistics. Even after attending, voluntarily or mandatorily, to a course in Statistics or Biostatistics, the meaning of the p-value, if ever understood, is quickly forgotten.

What is this p-value?

The sense of this p-value is basic and intuitive. It is similar to the concept of probability that we all deal with in everyday life. For example, no one can guarantee that he or she will leave home -a typical day towards ordinary work duties- and will not have an accident even though there is no evidence that it can occur. In a situation of normal climate and normal environment in any society (which is not in war and where some natural disaster is not happening), that probability is very low and therefore, usually no one stays home for that reason.

Although the probability of an accident happening is not 0, he or she makes the decision to go out on the assumption that such probability is "very low". One could then think like this: I'll go outside but I first hypothesize "I will have an accident" (only a hypothesis), I value it and think "there are no special reasons to have an accident today (I am in good health and everything works a usually) therefore, the probability of having an accident is extremely low, so I make the decision to leave home for work and carry on”. I have rejected the hypothesis of having an accident.

For the p-value the same reasoning can be applied. This p-value identifies the likelihood that certain results occur in the case the "hypothesis one tries to reject" is fulfilled. If p is very small (say, close to zero), the hypothesis is rejected. Why do we say "hypothesis one tries to reject"? Because it is assumed that if someone is going out it is because he or she wants to go out. The same happens in research.

It is very easy to give examples of similar situations in daily life as we are constantly rejecting or not rejecting hypothesis. Even though we live in the present, we always move into the future which is unknown and we only interact with it based on probabilities that some call *subjective or intuitive* probabilities [4], because they are not strictly subject to the laws of the probability theory, but still probabilities.

Specially focusing on the evaluation of new health interventions (which provide the clearest examples for the topic of hypothesis testing); when a new treatment emerges and the moment reaches to prove its effectiveness, the researcher in general is trying to "reject the hypothesis that the new treatment is not effective" thus to introduce it in practice.

That is the simple meaning of the p-value, the probability of obtaining certain results given the hypothesis "wanted to reject" is true; if the p-value is very small, the hypothesis is rejected and the desire accomplished. This does not mean, of course, that the hypothesis is not true because, as in everyday life, soothsayers only exist in fairy tales.

How the p-value is calculated (estimated)? First the intuition and then the decision

Another issue that deserves attention and may interest the non-statistical researchers on the p-value is how it is calculated. This can be really complicated and full of mathematical calculations that are not the intention of this article. However, we can also use intuition to understand how it is obtained.

More generally, all probabilities are estimated from repeated observations. I shall digress to change the verb "calculate" by "estimate". "Calculate" indicates something that you can know "for sure" as to calculate the required speed that would need a vehicle to travel a given distance in a certain time. But, in practice, the probabilities are estimated (we approach them) as the probability that "it will rain today". We cannot calculate it but we can estimate it. A meteorological observatory today can have many (hundreds, perhaps thousands) observations that allow you to calculate the proportion of times it rains in "days like today" and that proportion (or percentage) is the estimate of the probability of rain today. It is surely more complicated but, with the excuse of meteorologists, can be so simplified.

So what we need for this estimate are observations. That is the basis of the need for a study to estimate the p-value which implies, in effect, collecting data. If we go to the context of the evaluation of a new health intervention, we first need observations of results of the new intervention and also observations of the effectiveness of the intervention that had been used so far. In practical and general terms, what is needed is to know whether the new intervention is more effective than the one used so far, in other words, to compare the new intervention with what existed in terms of effectiveness.

Thus, the hypothesis we want to reject is that the new intervention is as effective as the previous one. Observations about the effectiveness of the new intervention and observations on the effectiveness (result, outcome) of the intervention that existed so far ("conventional" or "control" intervention) are needed. Both "outcomes" are then compared and the difference between them is obtained. Now we proceed to estimate the probability that this difference already observed (or one larger) can occur if the new intervention was equal to the previous one and that is precisely the p-value. In other words, p is the probability that the difference observed (or one larger) occur under the hypothesis that the actual difference is zero. Once estimated that probability, if very small, I decide to reject the hypothesis of equal effectiveness, as is done with the hypothesis "I will have an accident."

Not so easy

All this has been simplified for the sake of understanding the reasoning behind the p value and its meaning. But there are several aspects that complicate this subject and are the motive for the existence of a branch of mathematics (inferential statistics) to formalize and implement all this reasoning.

- For estimates, at least theoretical models that adjust to reality and probability theory are needed, that is, probability distributions of random variables that fit each situation or reality. In fact we work with probabilistic models, and also have tools to evaluate which is the optimal model for each of the situations.
- It is also needed to decide what a "small" probability is. Researchers, along with statisticians seem to have settled for accepting a p-value (for rejection) less than 0.05 as small enough though that value is also subject of broad discussion and controversy [2],[5]. To accept it without question means we could be making the decision to reject equal efficacies when an observed (or greater) difference can occur with a probability of 0.05 in the case of equal effectiveness. That means a risk, if to accept the new intervention will be very costly, a risk of 5% may not be compatible with a good decision. Therefore, the p-value (for rejection) should adjust to the consequences of making a mistake in case of rejecting equal efficacies. For example, if the implementation of new intervention costs a million dollars, one would be prone to run a lower risk than if it only costed ten thousand dollars. So, to always take a risk of 5% is not compatible with common sense. In this respect, Ronald Fisher, the renowned statistician who introduced the p-value in the 20s of the last century, said the following [6]:

It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result.

- The way the “effectiveness differences” are estimated needs formalization and classification. Not all sorts of outcomes can be measured likewise, it depends on how we consider measuring the effect of the intervention. In principle is simple at least to understand that there is a wide range of ways to measure the effectiveness of an intervention. The outcome of one type of exercise to relieve back pain and a medicine to reduce blood cholesterol cannot be measured the same.
- We need to ensure that the differences found between interventions are due to the interventions themselves and not to the multiplicity of factors that can influence the effectiveness of an intervention. This concerns the design of research and recognition of the biases to avoid in making comparisons. For example, if we are comparing the effectiveness of a new drug to one already known with the aim of increasing survival of patients with some type of cancer, we would need patients in the group with the new drug and those in the group with the already known drug be similar regarding age. In addition, both groups of patients should be similar with respect to stage of the cancer in question and probably to a number of other features. If not, the result of the comparison might be confused by the characteristics of the groups being compared leading to a typical problem of the design of studies comparing effectiveness: confounding.
- We need observations representing all reality since we are in the field of scientific research. There is no need to know whether a new intervention is more effective than the previous one in a specific health center for a month (for instance) but in general. The practical question to pose in the best case scenario is: Shall the new intervention introduced into the arsenal of health interventions targeted at certain end, or not? Therefore, though it may seem overblown, we must accept that in general we want to infer to infinite universes in space and time.
- We need to decide how many observations of reality must be taken for rejection decision (sample size). This is an issue related to the p-value and also to the need of considering the probability of making the "opposite" error that would be not to reject the hypothesis given it is false. Bland and Altman made an illustrative discussion of this aspect and emphasize the need to not confuse the absence of rejection (p-value greater than 0.05) with the lack of difference between efficacies without considering the sample size and the probability of "not rejecting" when actually the hypothesis is not true [7].

So, although the reasoning behind the p-value and the interest it arouses in research can be considered relatively simple, its estimation and its implementation in practice actually contain complex and difficult to solve byways. We have noted only some of these aspects, perhaps the most notorious.

It is worth adding that the 0.05 value mentioned earlier, is the livelihood of the well-known "statistical significance". If the p-value found in a comparison of effectiveness, as that taken as example, is less than 0.05 it is said that the difference between both efficacies is significant or more precisely, statistically significant. It means that we have a high probability (95%) that the difference we have obtained (or a greater one) does not come from a reality where the effectiveness is the same (or worse); we have rejected that hypothesis. In other words, since the p-value is very small, we decided to believe that the difference found "did not happen by chance" in a context where the new intervention really is no more effective than conventional.

But the summary of issues to be addressed before decisions can be made from a p-value tells us it is neither a simple problem for statisticians nor for researchers. The good news is that many of the aforementioned questions have been addressed theoretically and its solution is substantiated and implemented through computer programs already marketed and disseminated. The design aspects to avoid bias, are also reported and discussed extensively in the literature.

The p-value problems, statistical significance and clinical significance

What is wrong then with the p-value which has been so much criticized in recent decades? [8],[9],[10]. The answer has a theoretical and philosophical base coming from the conviction that the p-value or rather, significance tests, were not devised by its founder (Ronald Fisher, 1890-1962) to be used in practice as they have been used for decades [11].

But perhaps the most simple and clear answer to this question stems from the fact that small p-values can be obtained in cases where the differences are really small, or rather, too small to be of practical importance. That is, situations are given, where a difference without any practical meaning can be "statistically significant".

Decision making in all circumstances is subject to a risk / benefit, cost / benefit or risk + cost / benefit ratio. This happens in practice and also in daily life. Continuing the paragon at the beginning of this article, the decision to leave or not is based of course on the risk (chance) of having an accident but, what is the importance of the walking out? It is not the same to go to the movies for fun than to go out for work with the risk of losing a day´s wage. We tend to accept greater risks if the benefit (or need) deriving from the decision is greater.

Something similar happens in health practice situations. It is not the same to introduce a new intervention generating an improvement of 10% compared to the one generated before, than if this introduction will generate only a 1% improvement compared to the previous one. Introducing a new intervention has a cost that is only worth paying if the potential improvement will be substantial or important within the context where it is evaluated. So the fact that the p- value, which is supposed to help making decisions, is not sensitive to the relevance or the consequences of the decisions, is a really major problem.

The fact is that, plainly speaking, the p-value depends on the size of the sample so that just increasing the sample size, statistical significance can be obtained for any difference no matter how small it may be. It is not an exceptional phenomenon, it is just natural, in practice it is not possible for two interventions, as similar their efficacies might be, to act in the same way, achieving efficiencies (or effects) exactly the same. There will always be a difference and as the sample increases, the greater the probability of finding it [12]. So the actual goal is not -nor can it be- to find the difference at all costs but to detect the difference when it may have some benefit for practice or at least for theory. Therefore, if we make the decision to adopt the new intervention just because its effectiveness is significantly higher than that of the previous one, we run the risk that the expense derived from the introduction of the new intervention, in fact is not worthwhile.

So from the 80s of last century awareness of this problem aroused and the discussion began between, on the one hand, *statistical significance* and, what many have called *clinical significance* on the other [9],[13],[14]. The latter designation, is because the subject of the comparison of efficacies is of particular importance in the evaluation of new drugs through the well-known clinical trials but could be replaced by*practical significance* since it is not only an issue of clinical trials or clinical medicine. Researchers and journal editors had to be alerted that statistical significance and the p-value were inoperative without an assessment of the practical or clinical significance of the outcomes from the comparisons.

The remedy

Finding a remedy (alternative) for these contradictions detected forty years after using the p-value in the various hypothesis testing needed in research work and decision making, is still the topic of several articles and methodological discussions. Several alternative solutions to this problem have appeared in the world of research and statistics and reflected in scientific literature.

A radical approach is to change the whole system of statistical hypothesis testing to evaluate the effectiveness of new technologies or to test hypotheses in any field of scientific knowledge. This intends to be a new paradigm that promotes replacing frequentist techniques (as are often called these methods of testing hypotheses we have mentioned) by *Bayesian* techniques [15]. Bayesian methods to test hypotheses, are not based on the p-value obtained from a study in which all factors that may affect compliance with the hypothesis, except the one studied, are controlled. The encouraged idea is that the accumulated experience on a hypothesis can and should contribute to its verification. With previous experience (plus probability theory, of course) "a priori probability" is adopted and, with the data provided by a new study the "posterior probability" that would lead to decision making is attained. This is roughly the idea underlying Bayesian techniques that constitute today a broad array struggling to prevail as the main method of statistical inference but, apparently, has not been as successful as it might have been expected. Presciently in 1987, in an article in JAMA, Browner and Newman delved into the similarity between hypothesis testing in research and evaluation of diagnostic tests in the clinic scenario where *a priori* and *a posteriori* probabilities become particularly relevant [16].

Another remedy, perhaps the most widely accepted today, promotes the use of the so-called confidence intervals. These intervals are supposedly able to quantify our confidence in the results of a study and, above all, allow us to approach the magnitude of the *effect size* [9]. The term "confidence" deserves an explanation, we can return to our going-out example at the beginning of this paper. As we have left hundreds of times under certain conditions (climatic, personal, social and other) and have not had any accidents, we are confident that this time (since the above conditions are present) we will not have it either. Very often confidence and probability are confused with each other but it is clear that we do not know the probability of having an accident, we could estimate it but cannot know it. Actually we are confident that we will not have an accident because the probability of having it (given the conditions) is very low. However in statistics the term "confidence" replaces “probability” once the sample has been obtained and we have results, basically because the probability is a term linked to the unknown. Once you have calculated the so-called confidence interval, we cannot speak of chance because the fact already happened.

More specifically in the example of comparing efficiencies between new and an existing intervention, the idea is to obtain confidence intervals for the difference between the two interventions. It means to find a method with a high probability of success (i.e., a method that provides a high probability that the difference actually is between the two boundaries of the interval). Once the interval is obtained we speak of confidence, that is, instead of saying "this is the interval obtained with a method which provides high probability that the difference really is in this range," one says "this is the confidence interval ". It should be added that for a confidence interval to have utility, it cannot have a very large breadth. What could be the gain of knowing, with high confidence that the difference in the effectiveness of two interventions lies between 1 and 50% improvement? That can surely is known without any study. To obtain narrow confidence intervals adequate sample sizes are needed. Therefore, to obtain confidence intervals does not eliminate the need to search for the required sample size, it only gives us more information than just the p-value.

Nevertheless, substituting p-values by confidence intervals is not actually to change "evil for good" but to add information to the result. Confidence intervals are obtained by means of same statistical techniques as the p-value to the point that in fact both are equivalent [17]. Therefore it is misleading to think of confidence intervals as something different from the p-value, it only constitutes extra information. Ronald Fisher never proposed that this value (p) was taken as a rigid rule to make a decision but as a valuable aid to implement statistical inference [18],[19].

Other alternatives have been proposed in the literature but, in general, are more difficult to understand and assimilate. For example, the use of likelihood ratios to compare the degree of adjustment of two models to reality [5],[20]. In the general effectiveness comparison example (new intervention to conventional), we would have a model supposing the new intervention to be more effective than the conventional, another model states both efficiencies are equal. Data are obtained (the corresponding study is conducted) and the model that best fits the data is accepted. It seems attractive, but has not succeeded in replacing hypothesis testing although it is a basic method for assessing different regression models and is used extensively in other areas of knowledge.

Shall we resume the use of p-values?

As a final thought, and as an example that the famous expression of Socrates "I know one thing: that I know nothing" is and probably will be forever valid, I must point out that currently several voices in the field of science, intend to restore the use the p-value. The most direct appear to have been those of Joseph Fleiss in 1986 [21] and Clarice R. Weinberg in 2001 [22]. The latter in her commentary published in the journal Epidemiology, responds to the tendency to qualify the use of p-value as "politically incorrect" that flourished in the late 90s of the last century. Both are abundant in examples of many situations among the applications of statistical data analysis techniques, where it may be convenient to use p-values. Mention may be made to those situations involving joint evaluation of multiple hypotheses or questioning the relevance of certain model.

The prevailing idea is that the p-value must be taken for what it really is, an aid to decision-making that cannot be abstracted from the context where decisions take place and much less taken as a proxy for all practical/scientific reasoning surrounding any appraisal of knowledge and its applications. The very recent statement of the American Statistical Association (ASA) is eloquent and clear in this regard [23].

Conclusion

The p-value together with statistical significance and practical significance of results found in a study have been the subject of discussion in the scientific research scenario for decades. The balance undoubtedly leans today towards confirming that, in most cases, there should not be interest in a mere statistical significance, and that any significance test (if performed) must be accompanied by a demonstration of its potential practical significance. However it is a comprehensive and current issue because there is still no general consensus on what might be the better, less risky and faster, way to infer from the sample, which represents a study, to the population where questions are posed.

Notes

Acknowledgement
To Vivienne C. Bachelet, for her critical review of the manuscript and contributions to the clarity of the text.

From the editor
The author originally submitted this article in Spanish and subsequently translated it into English. The*Journal* has not copyedited the English version.

Conflicts of interest
The author declares there is no conflict of interest related to the topic being addressed.

Funding
The author declares she has not received any funding for the preparation of this article.

References

1. Bertolaccini L, Viti A, Terzi A. Are the fallacies of the P value finally ended? J Thorac Dis. 2016 Jun;8(6):1067-8. | CrossRef | PubMed |
2. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods Nature Research. 2015 Feb 26;12(3):179-85. | Link |
3. Huber W. A clash of cultures in discussions of the P value. Nat Methods. Nature Research. 2016 Jul 28;13(8):607-607. | Link |
4. Woolfson M. M. 1.1. Probability and Everyday Speech. Everyday probability and statistics health, elections, gambling and war. 1st ed. London: Imperial College Press; 2008: 5-6.
5. Dixon P. The p-value fallacy and how to avoid it. Can J Exp Psychol. 2003 Sep;57(3):189–202.| PubMed |
6. Fisher RA. The Design of Experiments. 2nd ed. Edinburgh: Oliver and Boyd; 1937.
7. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995 Aug 19;311(7003):485. | PubMed |
8. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008 Jul;45(3):135-40. | CrossRef | PubMed |
9. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986 Mar 15;292(6522):746-50. | Link |
10. Goodman SN, Hopkins J. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999;130:995-1004. | Link |
11. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51(4):355-60. | PubMed |
12. Silva Ayçaguer LC. [Confidence intervals and p values]. Medwave. 2014;14(1):e5894. |CrossRef | PubMed |
13. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. J Am Stat Assoc. 1987 Mar;82(397):112-22. | Link |
14. Blackwelder WC. "Proving the null hypothesis" in clinical trials. Control Clin Trials. 1982;3(4):345-53. | PubMed |
15. Silva LC. 6.5.3. El advenimiento de la era bayesiana. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997:156–7.
16. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA. 1987;257(18):2459-63. | PubMed |
17. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ. 2011;343:d2304. | PubMed |

18. Buyse M, Hurvitz SA, Andre F, Jiang Z, Burris HA, Toi M, et al. Statistical controversies in clinical research: statistical significance-too much of a good thing ….Ann Oncol. 2016 May;27(5):760-2. | CrossRef | PubMed |
19. van Helden J. Confidence intervals are no salvation from the alleged fickleness of the P value. Nat Methods. Nature Research; 2016 Jul 28;13(8):605-6. | Link |
20. Glover S, Dixon PLikelihood ratios: a simple and flexible statistic for empirical psychologists. Psychon Bull Rev. 2004 Oct;11(5):791-806. | PubMed |
21. Fleiss JL. Significance tests have a role in epidemiologic research: reactions to A. M. Walker. Am J Public Health. 1986 May;76(5):559-60. | PubMed |
22. Weinberg CR. It's time to rehabilitate the P-value. Epidemiology. 2001 May;12(3):288-90. |PubMed |
23. Wasserstein RL, Lazar NA. The ASA's Statement on p - Values: Context, Process, and Purpose. Am Stat. Taylor & Francis; 2016 Apr 2;70(2):129-33. | Link |

Author address:
[1] Villaseca 21 Of. 702
Ñuñoa
Santiago
Chile