Revista Mexicana de Cardiología

Volumen 26, Número 1
Enero - Marzo 2015
pp 56 - 58

Comunicado breve

Demystification of the significance of p in statistical tests

Reinaldo Alberto Sánchez Turcios*

ABSTRACT

All statistical tests have a p value that is significant when < 0.050. This value was arbitrarily determined by RA Fisher and accepted consensually over time. Since its genesis, this value has been questioned, and nowadays it is under the careful eye of many statisticians. This issue has led to a debate among the scientific community: obtaining p significance was considered as a guarantee that the research project would be an appropriate contrast between the hypothesis and the acceptance, or rejection, of it. The purpose of this paper is to construct a discussion about p significance.

Key words: Statistic test, p value, p significance.

RESUMEN

Todas la pruebas estadísticas tienen un valor de p significativo a partir de < de 0.050, el cual fue arbitrariamente determinado por RA Fisher y aceptado por consenso a través del tiempo. Desde su génesis, este valor ha sido cuestionado y actualmente está bajo la mirada escrupulosa de muchos estadígrafos, por lo que se establece un debate en la comunidad científica donde clásicamente se consideraba obtener la significancia de p un sello de garantía, que el proyecto de investigación era capaz de aceptar o rechazar la hipótesis. El objetivo de este artículo es discutir los cuestionamientos de la significancia de p.

Palabras clave: *Pruebas estadísticas, valor de p, significancia de p.*

INTRODUCTION

It was RA Fisher, a British statistician and geneticist, who first proposed p significance formally in 1925 (which was arbitrarily determined and consensually accepted over time), it was published in Statistical Methods for Research Workers[1]. This p significance was not founded on scientific basis, but on the assumption that p was < 0.050.[2] The significance of p received strong criticism from Jerzy Neyman, a Polish Mathematician, since its inception and from the British statistician Egon Person[3] in our time. Sthepen Ziliak,[3] an economist with scientific thought, criticizes the statistic tests used in research which frequently are inadequately used. However, there are defenders of «p significance».

P significance is 5%, but it is probably the most misunderstood and misinterpreted index calculated for research in different disciplines of knowledge.[2-4] In a recent survey to medical residents published by JAMA,[4] 88% of the people surveyed could not interpret correctly p value; it is for this reason that it was considered necessary a careful look at p significance.

METHODOLOGY

In 12 papers and research reports that appeared in Science (December 14th 2014) and in 20 out of 22 papers published in the Journal of Pharmacology and Experimental Therapeutics (December 2012), p values are mentioned when describing experimental results. These values are considered valid even when several of them are at the threshold of their significance in those studies.

Nature of «p significance» in statistic tests

The significance of p is defined as the probability that Ho is true according to the research methodology whose hypothesis is proposed to demonstrate.

One of the objectives of this paper is to clarify the steps that R.A. Fisher described originally: 1. «p significance» is not definite; 2. RA Fisher tried to simplify the procedure of the test to make a judgment on the evidence, and that it would be true; 3. He stated that a set of experiments should be carried out to observe whether the results were consistent with randomness; 4. The researchers must first postulate a Ho that must be rejected, such as the existence of

* MD Endocrinology, MSc Pharmacology. Hospital General Milpa Alta. Mexico City, Mexico.

correlation, or difference between the two groups being analyzed; 5. The researchers must assume that Ho is in fact true and calculate the observed value with the data obtained, and they have to see how distant it is from a p value previously established and 6. Fisher suggested that if p had a smaller magnitude than the set value, then Ho would be false.

Once the Ho and the Ha had been specified, research should define the level of significance which is usually known as $\alpha < 0.050$. This rejects Ho, that is to say if the p value calculated with the data from the two samples is less than α value previously established, then Ho is rejected. In this step is where the problem of p interpretation resides.

DISCUSSION

The main restriction of p value is that the behavior of the results cannot be visualized simultaneously, in the long or short term (over time), of the treatments whose data will be subjected to statistic tests.

The deductive method is used, in the long term, to obtain these data. If we hypothetically carried out a repetition of the experiment, the results should be consistent with the p value obtained. When designing a new drug, many samples are analyzed for a given pathology, and the results are interpreted in a global fashion.

One of the problems is that the repetition of the drug usage in different samples is not the same; it is assumed that these data would be consistent with the p value found in preliminary studies. Other way to look at the problem is when you do a single experiment (proving Ho); this study is inductive and establishes only one result as evidence, expecting to generalize the obtained result and replicate it in other samples.[5,6] The latter proposition comes true when pretending to describe, with one short term study, the behavior in the long term of a new drug, which will replace the one in use.

Main criticisms of p value

The main criticisms of p value are:

1) P values is not a usual measurement for inference.[7]
2) There are at least 12 wrong interpretations of p significance.[5,6,8,9]
3) P values do not offer exact evidence of sample differences. P values are affected by the size of the samples; if you increase the size of a sample, p value will change from significant to non-significant.[9]

Other factors that can be observed of p values misinterpretation in research papers are:[10]

a) The smaller the size of the sample, the more likely p values are wrong.
b) The smaller the sample effect size, the less likely the obtained results are truthful. (Size effect analysis).
c) If the preliminary studies are scarce and demonstrative of the studied effect association, it is more likely that the findings will be truthful.
d) The flexibility, design, definitions, favorable events, and analytic methods used in the research may lead us to false results.
e) The more controversial the study (with a larger team of researchers involved), the more likely the findings are false.

Effect size[10]

The effect size, that a researcher hypothesizes to exist between two patients subjected to two different treatments, is a value that can be referred as the heterogeneous difference in the effects of both treatments.

The effect size is the difference that the researcher proposes; whether there is a/or several effects between two treatments in a set of patients with a common pathology. Its importance in the statistical analysis is that with this value (effect size), significance and a pre-established sample size, the researcher is able to calculate the power of its statistical test, that is to say: what probability the result has to reject Ho with the available data if the statistical test has a very small power (usually less than 80%). In order to improve statistical power, the size of the sample has to be increased due to the significance level previously assigned and the effect size proposed. This latter premise is relevant since in most protocols the obtained results when rejecting Ho are not important, because the statistical analysis was carried without an acceptable power. The greater the power in a statistical design, the bigger the likelihood of Ho rejection when Ho is false.

Statistical power depends on:

1. Sample size.
2. The previously determined α significance level; it is usually from 0.050 to 0.01. It is important to

Rev Mex Cardiol 2015; 26 (1): 56-58

point out that when the effect size is very small (the researchers expect that the investigated drug will not modify the previously established treatment), the sample sizes for the fixed α significance level will be bigger.

Conversely, if one knows that the effect size is big, the sample size is relatively small: 1) < 0.20: very small, 2) 0.030-0.050: small, 3) 0.050-0.70: medium, 4) > 0.80: big. It is common to assign an effect size of 1.0 to 2.0.

Cohen,[11] the creator of this concept, makes clear that the terms: small, medium, and big have to be interpreted within the concept of the statistical analysis being concreted.

CONCLUSIONS

P significance is still valid if the originally proposed procedures by RA Fisher are executed with a strict scientific discipline when obtaining a set of p values.

Outlook. An adequate knowledge of the effect size is proposed to be able to raise an appropriate sample size with proper significance level, and statistical power of a correct analysis. It is likely that Bayesian statistics could solve the problem of the different sizes.

REFERENCES

1. Fisher RA. *Statistical methods for research workers*. 5 de Biological monographs and manuals. La Universidad de California, 1932, pp. 1-307.
2. Lew MJ. To P or not to P: on the evidential nature of P-values and their place in scientific inference stat. *ME*. 2013; 3: 27-46.
3. Nuzzo R. P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*. 2014; 506: 150-152.
4. Donna M, Windish MD, Stephen J, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA*. 2007; 298: 1010-1022.
5. Lecoutre MP, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *Int J Psychol*. 2003; 38: 37-45.
6. Hubbard R, Lindsay MR. Why p values are not a useful measure of evidence in statistical significance testing. *Theory Psychol*. 2008; 18: 69-88.
7. Lew MJ. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *Br J Pharmacol*. 2012; 166: 1559-1567.
8. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008; 45: 135-140.
9. Hurlbert SH, Lombardi CM. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neo Fisherian. *Ann Zool Fennici*. 2009; 46: 311-349.
10. Coe R. *It's the effect size, stupid. What effect size is and why it is important. School of education, University of Durham*. Paper presented at the British Educational Research Association, Annual Conference. Exter, 12-14 September, 2002.
11. Cohen J. Statistical power analysis for the behavioral sciences. New York, Academic Press. 1977, pp. 216-380.

Address for correspondence:

Reinaldo Alberto Sánchez Turcios
Tepic Núm. 113-610,
Col. Roma Sur, 06760, México, Distrito Federal.
Tel. 015552648061,
Cel. 0445543508824,
E-mail: rturcios@live.com.mx