BENTHAM OPEN

CrossMark

The Open Cardiovascular Medicine Journal

Content list available at: www.benthamopen.com/TOCMJ/

DOI: 10.2174/1874192401610010196

RESEARCH ARTICLE

Insights in Hypothesis Testing and Making Decisions in Biomedical Research

Varin Sacha[1] and Demosthenes B. Panagiotakos[2,*]

[1] *Collège de Villamont, Lausanne, Switzerland*
[2] *School of Health Science and Education, Harokopio University, Athens, Greece*

Received: December 2, 2015 Revised: February 22, 2016 Accepted: March 5, 2016

Abstract: It is a fact that *p* values are commonly used for inference in biomedical and other social fields of research. Unfortunately, the role of *p* value is very often misused and misinterpreted; that is why it has been recommended the use of resampling methods, like the bootstrap method, to calculate the confidence interval, which provides more robust results for inference than does *p* value. In this review a discussion is made about the use of *p* values through hypothesis testing and its alternatives using resampling methods to develop confidence intervals of the tested statistic or effect measure.

Keywords: Biostatistics, Confidence intervals, Evidence based medicine, *p* value, Resampling.

BRIEF HISTORY OF HYPOTHESIS TESTING

At first it has to be clarified that a "significance test" is different to a "hypothesis test". Many textbooks, especially in social and biomedical sciences, mix these two approaches to a logically flawed mishmash, which is referred as "null-hypothesis significance test". However, null-hypothesis significance test is a combination of ideas developed in the 1920s and 1930s, primarily by Ronald Fisher (1925) and Jerzy Neyman & Egon Pearson (1933) [1]. These two testing approaches are not philosophically compatible even if they are technically related. Fisher developed tests of significance as an inferential tool. The main reason was to walk away from the subjectivism inherent to Bayesian inference (*i.e.*, namely in the form of giving equal prior probabilities to hypotheses) and substitute a more objective approach. However, Fisher's tests also depend on two other important elements: research methodology (Fisher pioneered experimental control, random allocation to groups, *etc.*) and small samples. Neyman & Pearson liked Fisher's approach, although lacked a strong mathematical foundation. As their theory progressed, the approach stopped being an improvement on Fisher's approach and became a different approach. The main differences between Fisher and Neyman & Pearson approaches are both philosophical and technical. Philosophically, Neyman and Pearson's approach assumes a known hypotheses, and it is based on repeated sampling from the same population, focuses on decision making, and aims to control decision errors in the long run. Thus, it can be considered as less inferential and more deductive. Technically, Neyman and Pearson's approach uses Fisher's tests of significance, but also incorporates other elements, like effect sizes, Type II errors, and the power of the statistical test. Neyman and Pearson also incorporated other methodological improvements, such as random sampling [2 - 15].

Significance test and hypothesis test are based on the assumption of a (statistical) null hypothesis, *i.e.*, a statement that there is no relationship, *e.g.*, no difference between treatment effects on an outcome. This is a mere technical requirement giving a statistical context that is required to apply probabilistic calculations. In reference to the approach suggested by Fisher, a "significance test" considers only the null hypothesis and gives a *p* value which is a continuous empirical measure of the "significance of the results" (given the considered null hypothesis). This measure has no

* Address correspondence to this author at the 46 Paleon Polemiston St. Glyfada, Attica, 166 74, Greece; Tel.: +30210-9603116; Fax: +30210-9600719; E-mail: d.b.panagiotakos@usa.net

particular meaning and it is not calibrated to some kind of relevance. It is just a value between 0 and 1, referring on how likely is to observe "more extreme results" given the null hypothesis. According to the approach suggested by Neyman & Pearson, a "hypothesis test" is actually a test about an alternative hypothesis, which refers to a "minimally relevant effect" (and not about "some non-zero effect" as the null hypothesis). These tests are designed to control error-rates and allow a balance on the expected cost/benefit ratios that are associated with the actions taken based on the test results. To perform such tests, it must be specified a minimally relevant effect and also acceptable error rates. After the experiment or the study is conducted, the decision is actually about rejecting (or not) a hypothesis. So either the "null hypothesis" is not rejected, which means that the assumed effect was not relevant, or the alternative hypothesis is accepted, which means that the effect was relevant. Note that there is no point where the "truthfulness" of an effect is discussed. This does not matter in statistical hypothesis testing. The only thing that matters is what actions are taken based on an effect that is considered relevant [2 - 15].

Major Problems Using the *p* Values as Result of a Hypothesis Test

Many investigators, in various research fields refer to Neyman & Pearson hypothesis tests and their associated *p* values. Indeed, the *p* value is a widely used tool for inference in studies. However, despite the numerous books, papers and other scientific literature published on this topic, there still seems to be serious misuses and misinterpretations of the *p* value. According to Daniel Goodman, "a *p* value is the right answer to the wrong question" [1]. A summary is given by Joseph Lawrence that presented at least four different major problems associated with the use of the *p* values [16]:

1. "*P* values are often misinterpreted as the probability of the null hypothesis, given the data, when in fact they are calculated assuming the null hypothesis to be true."
2. "Researchers often use *p* values to “dichotomize” results into “important” or “unimportant” depending on whether *p* is less or greater than a significance level, *e.g.*, 5%, respectively. However, there is not much difference between *p*-values of 0.049 and 0.051, so that the cut off of 0.05 is considered arbitrary."
3. "*P* values concentrate attention away from the magnitude of the actual effect sizes. For example, one could have a *p* value that is very small, but is associated with a clinically unimportant difference. This is especially prone to occur in cases where the sample size is large. Conversely, results of potentially great clinical interest are not necessarily ruled out if $p > 0.05$, especially in studies with small sample sizes. Therefore, one should not confuse statistical significance with practical or clinical importance."
4. "The null hypothesis is almost never exactly true. In fact it is hard to believed that the null hypothesis, H_o: $\mu = \mu$, is correct! Since the null hypothesis is almost surely false to begin with, it makes little sense to test it. Instead, it should rational to start with the question “by how much are the two treatments different?"

There are so many major problems related to *p* values that most statisticians now recommend against their use, in favour of, for example, confidence intervals. In a previous publication entitled “The value of *p*-value in biomedical research” alternatives for evaluating the observed evidence were briefly discussed [17]. Here, a thorough review on hypothesis testing is presented.

Hypothesis Testing *Versus* Confidence Intervals

Researchers from many fields are very familiar with calculating and interpreting the outcome of empirical research based solely on the *p* value [18]. The commonly suggested alternative to the use of the hypothesis tests is the use of confidence intervals [19 - 26]. As it has been suggested by Wood (2014), “*the idea of confidence intervals is to use the data to derive an interval within a specified level of confidence that the population parameter will lie with confidence*" [19]. Two-sided hypothesis tests are dual to two-sided confidence intervals. A parameter value is in the $(1-\alpha)x100\%$ confidence interval if-and-only-if the hypothesis test whose assumed value under the null hypothesis is that parameter value accepts the null at level α. The principle is called the duality of hypothesis testing and confidence interval [20]. Thus, there is a one-to-one relationship between one-sided tests and one-sided confidence intervals. In addition, there is an exact relationship only if the standard error used in both the confidence intervals and the statistical tests, is identical.

However, many statisticians nowadays avoid using any hypothesis tests, since their interpretations may vary and the derived *p* values cannot, generally, be interpreted in meaningful ways. Moreover, it is adopted that by calculating the confidence interval, researchers may have “insights” to the nature of their data and the evaluated associations, whereas *p* values tell absolutely nothing. Criticism against hypothesis testing, dating for most of them more than 50 years ago, suggests that "they (hypotheses tests) are not a contribution to science" (Savage, 1957 in Gerrodette, 2011, p. 404) or "a serious impediment to the interpretation of data" (Skipper & *et al.*, 1967, in Gerrodette, 2011, p. 404), or "worse than irrelevant" (Nelder, 1985 in Gerrodette, 2011, p. 404) or "completely devoid of practical utility" (Finney, 1989, in Gerrodette, 2011, p. 404) [1].

Nevertheless, and despite all the criticism, the hypothesis tests and their associated *p* values are still widely prevalent. According to Lesaffre (2008) [21], it is important to note that a 95% confidence interval bears more information than a *p* value, since the confidence interval has a much easier interpretation and allows better comparability of results across different trials. Moreover, in meta-analyses, the confidence interval is the preferred tool for making statistical inference. According to Wood (2104) [19], a (1-α)x100% confidence interval provides directly the strength of the effect, as well as the uncertainty due to sampling error, in an obvious way by providing the width of the interval. The information displayed is not trivial or obvious like the NHST conclusions may be, and misinterpretations seem far less likely than for NHSTs. Thus, the use of the confidence intervals has the potential to avoid many of the widely acknowledged problems of NHSTs and *p* values [19]. Moreover, several high-impact journals, especially in health sciences and other fields, as well as Societies (*e.g.*, American Psychological Association's (APA) Task Force on Statistical Inference (TFSI)) have strongly discouraged the use of *p* values to prefer point and interval estimates of the effect size (*i.e.*, odds ratios, relative risks, *etc*), instead of *p* values, as an expression of uncertainty resulting from limited sample size and also encouraging the use of Bayesian methodology [21 - 22]. It is not surprising to note that, a century following its introduction many researchers still poorly understand the exact meaning of *p* value, resulting in many miss-interpretations [17].

Advantages of The Confidence Interval *Versus p* Value

It is now common belief that researchers should be interested in defining the size of the effect of a measured outcome, rather than a simple indication of whether it is or not statistically significant [23]. On the basis of the sample data, confidence intervals present a range of alternative values in which the unknown population value for such an effect is likely to lie. Indeed, confidence intervals give different information and have different interpretation than *p* values, since they specify a range of alternative values for the actual effect size (since they present the results directly on the scale of the measurement), while *p* values don't. Moreover, confidence intervals make the extent of uncertainty salient, which a *p* value cannot do. Since the mid 1980's, Gardner & Altman suggested that "*a confidence interval produces a move from a single value estimate - such as the sample mean, difference between sample means, etc – to a range of values that are considered to be plausible for the population*" [24].

Resampling Techniques

It is known from basic statistics that many statistical criteria (*e.g.*, t-test) are asymptotically normally distributed, but the normal distribution may not be always a good approximation to their actual sampling distribution in the empirical samples derived from experiments, clinical trials or observational surveys. Indeed, the validity of the traditional statistical inference is mostly based on a theorem known as the Central Limit Theorem, which stipulates that, under fairly general conditions, the sampling distribution of the test statistic can be approximated by a normal distribution or under more limited assumptions by the t- or chi-square distributions. Based on these assumptions confidence intervals and *p* values are then calculated; however, with a considerable level of doubts and concerns.

The point of resampling method is to not rely on the Gaussian assumptions. Resampling is a methodology suggested in early 1940s in order to estimate the precision of statistics, like means, medians, proportions, odds ratios, relative risks, *etc.*, by using *k*-subsets of size *m* (< n) of the originally collected data (*i.e.*, jackknife method) or drawing a random set of data with replacement from the original set (*i.e.*, bootstrap method). Indeed, when the Gaussian assumptions are not true, the validity of the classical inferential statistics tends to be undermined. It is in these situations that the resampling methods really come to the rescue. The main idea of resampling is to obtain an empirical distribution of the test statistics based on what it is observed and use it to approximate the true, but unknown, distribution of the test statistic. An important advantage of this approach is that it could be applied for many statistics (*e.g.*, means, median, *etc.*) and effect size measures (*e.g.*, correlation coefficients, odds ratios, relative risks, *etc.*) with the use of computer software. Specifically, there are different types of resampling methods, *i.e.*, bootstrap, jackknife, cross-validation (also called rotation estimation and permutation test, or randomization exact test). In classical parametric test the observed statistics are compared to the theoretical sampling distributions, while in resampling methods we start from theoretical distributions, which makes them innovative approaches [25]. Among all resampling methods, bootstrap is certainly the most frequently used procedure [26]. So, the resampling methods can be a substantial improvement over the traditional inference, since a confidence interval for the true value of unknown statistic or effect size measure has a much more concrete interpretation than has the p value from a statistical test, although there is still no guarantee.

However, at this point it should be mentioned that it is often the sampling distribution of various effect sizes to be highly skewed, thus, the traditional confidence intervals will not work well, since they will always be skewed, too. Symmetrical confidence intervals are appropriate for a few things such as means and linear regression coefficients, but they are inappropriate for many other measures [27]. So, it is better not to assume a symmetric confidence interval for a measure of association, and to start from the assumption that they are not normally distributed. The empirical distribution derived for example from the bootstrap method does not assume that the distribution is symmetrical.

CONCLUSION

In conclusion, it could be recommend for inferencial purposes, to present the results from studies using confidence interval of the statistics and effect size measures of interest, rather than hypothesis test and its associated p value. Moreover, depending on the statistics of interest, bootstrap techniques or another resampling methods are also recommended, because these techniques are independent of the shape of the underlying distribution and can easily performed using software.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared None.

REFERENCES

[1] Gerrodette T. Inference without significance: measuring support for hypotheses rather than rejecting them. Mar Ecol (Berl) 2011; 32: 404-18. [http://dx.doi.org/10.1111/j.1439-0485.2011.00466.x]

[2] Fisher RA. Inverse probability and the use of likelihood. Proc Camb Philos Soc 1932; 28: 257-61. [http://dx.doi.org/10.1017/S0305004100010094]

[3] Fisher RA. Statistical Methods for Research Workers. 12th ed. Edinburgh: Oliver and Boyd 1954.

[4] Fisher RA. Statistical methods and scientific induction. J R Stat Soc Series B Stat Methodol 1955; 17: 69-78.

[5] Fisher RA. The Design of Experiments. 7th ed. Edinburgh: Oliver and Boyd 1960.

[6] Fisher RA. Statistical Methods and Scientific Inference. 3rd ed. London: Collins 1973.

[7] Neyman J. Basic ideas and some recent results of the theory of testing statistical hypotheses. JR Stat Soc 1942; 105: 292-327. [http://dx.doi.org/10.2307/2980436]

[8] Neyman J. First Course in Probability and Statistics. New York: Henry Holt 1953.

[9] Neyman J. The problem of inductive inference. Commun Pure Appl Math 1955; III: 13-46. [http://dx.doi.org/10.1002/cpa.3160080103]

[10] Neyman J. Note on an article by Sir Ronald Fisher. J R Stat Soc Ser B Stat Methodol 1956; 18: 288-94.

[11] Neyman J, Fisher RA. R. A. Fisher (1890--1962): an appreciation. Science 1967; 156(3781): 1456-60. [http://dx.doi.org/10.1126/science.156.3781.1456] [PMID: 17741062]

[12] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika 1928; 20A: 175-240.

[13] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond 1933; A 231: 289-337.

[14] Schmid bauer H, Rösch A. Bus 701: Advanced Statistics. Istanbul: Istanbul Bilgi Üniversitesi 2010; 1-30. Available from: www.hs-stat.com/courses/BUS701 /slides/bus701_ch15_v2009-01-20.pdf

[15] Hald A. A History of parametric statistical inference from Bernoulli to Fisher, 1713 to 1935. Denmark: Department of applied mathematics and statistics, University of Copenhagen 2004; pp. 1-199.

[16] Lawrence J. Review: Frequentist inferences for means and proportions. Available at: http://www.medicine.mcgill.ca /epidemiology/joseph/courses/EPIB-621/MeanProp.pdf

[17] Panagiotakos DB. Value of p-value in biomedical research. Open Cardiovasc Med J 2008; 2: 97-9.

[http://dx.doi.org/10.2174/1874192400802010097] [PMID: 19430522]

[18] Lakens D, Evers ER. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci 2014; 9(3): 278-92.
[http://dx.doi.org/10.1177/1745691614528520] [PMID: 26173264]

[19] Wood M. *P* values, confidence intervals, or confidence levels for hypotheses. 2014. Available at SSRN from : http://dx.doi.org/10.2139/ssrn.2393927

[20] Casella G, Berger RL. Statistical Inference. 2nd ed. Boston, USA: Cengage Learning 2002.

[21] Lesaffre E. Use and misuse of the *p*-value. Bull NYU Hosp Jt Dis 2008; 66(2): 146-9.
[PMID: 18537787]

[22] Cumming G, Finch S. Inference by eye, confidence intervals and how to read pictures of data. Am Psychol 2005; 60(2): 170-80.

[23] Altman DG, Machin D, Bryant TN, Gardner MJ, Eds. Statistics with Confidence. 2nd ed. London: BMJ Books 2000.

[24] Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986; 292(6522): 746-50.
[http://dx.doi.org/10.1136/bmj.292.6522.746] [PMID: 3082422]

[25] Yu CH. Resampling methods: concepts, applications, and justification. Pract Assess Res Eval 2003; 8(19): 1-23.

[26] Tu W. Resampling Methods. In: El-Shaarawi A-H, Piegorsch W, Eds. Encyclopedia of environmetrics. Chichester: John Wiley & Sons Ltd 2013.

[27] Newcombe RG. Confidence Intervals for Proportions and Related Measures of Effect Size. Boca Raton, Floride: CRC Press 2012.
[http://dx.doi.org/10.1201/b12670]