Nepal Journal of Epidemiology

eISSN 2091-0800

Editorial Open Access

Understanding significance and p-values

Shrikant I Bangdiwala[1]

Correspondence: Dr. Shrikant I Bangdiwala, Professor, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC, USA. Email: kant@unc.edu

Received 23 March 2016/**Revised** 24 March 2016/**Accepted** 24 March 2016

Citation: Bangdiwala SI. Understanding significance and p-values. Nepal J Epidemiol. 2016;6(1); 522-524.

On 7 March 2016, the American Statistical Association (ASA) released a statement to improve the interpretation of statistical significance and p-values and their role in scientific research. In their concluding remarks, they state "*Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning*" [1]. The ASA statement also states: "*Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value*" [1]. The p-value is thus a single index, so where does it and 'statistical significance' stand within 'good statistical practice'?

The ASA statement provided 6 guidelines for the use of p-values as part of good statistical practice:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

There has been much reaction worldwide to the ASA's statement. I would like to elaborate on point #5, on the role of p-values and statistical significance in assessing the size of an effect or the importance of a result. For this, I will rely on a historical perspective. Sir Ronald A. Fisher (1890-1962), considered as the father of modern statistical inference, introduced the idea of significance levels as a means of examining the discrepancy between the observed data and a model assumption contained in the null hypothesis [2]. Fisher (1935) stated that

"*…it is certain that the interest of statistical tests for scientific workers depends entirely from their use in rejecting hypotheses which are thereby judged to be incompatible with the observations.*" [2]

The p-value quantified the probability of getting a difference equal to or larger than the one observed, if the null hypothesis is true. Fisher viewed the p-value as an informal index of that discrepancy of the data with the assumed model (null hypothesis), and depending on its value, one could have 'weak' or 'strong' evidence against the null hypothesis. He suggested the following interpretation for the p-value:

"*If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts.*" [3]

So Fisher suggested ranges of p-values in which the evidence was weak for the null hypothesis (0.10 to 0.90) and for which the evidence was strong against the null hypothesis (below 0.02). He did not mention p-values greater 0.90 since they obviously were even weaker in supporting the null hypothesis. So what about p-values between 0.02 and 0.10? Fisher argued for continued experimentation or obtaining more information, and was himself inconsistent in claiming significance within this range.

So where does the cutpoint of significance at 0.05 come from? An interesting historical circumstance may be the explanation. In the early 20th century, when exact small-sample tests using the χ^2 , t and F statistics required tabulations for distributions other than the Gaussian (Normal) distribution, Fisher (1925) saw it convenient in his book on statistical methods for researchers to provide simple tabulations, not of the entire permutation-based distributions of the test statistics, but only of selected quantiles from those distributions[4]. He provided the quantiles at the extremes – say 10%, 5%, 2% and 1%, which were useful for researchers when testing hypotheses. The choice of simplifying the tabulations was made simply out of convenience in the era of laborious hand calculations. However, the 5% and 1% cutpoints from the tables in Fisher's classic book were assumed by many researchers to be the only choices for assessing 'significance'!

Also, when explaining the use of his χ^2 table, Fisher mentions "*We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy*." And in 1926, Fisher, when providing guidelines for agricultural experiments, states *". . . it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials. . .*" [5]. Fisher worked in agricultural experimentation, and he normally encountered the Gaussian probability distribution. He writes in the 13th edition of his 1925 classic book, page 44: "*The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not.*" Thus, despite the arbitrariness of this 'convenient' choice, the 0.05 level was seen to be the 'magical' definition of significance of a result, since 'Fisher himself said so.' Fisher (1973) in his later years was quite critical of the use of a fixed conventional level: "*... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."*[6]

The abuse of the p-value cutpoint of 0.05 has been recognized in many fields, and there is an increasing practice of not reporting p-values but effect sizes and corresponding measures of uncertainty (e.g. confidence intervals). However, the term 'significance' is still widely used to mean 'importance' and here lies my main concern. The term significance is unfortunately integral to the English language, and as Merriam-Webster online dictionary [7] defines it, it is "*the quality of being important: the quality of having notable worth or influence*." Among the expanded definition, they also define significance as: "*the quality of being statistically significant.*" Here is the major problem. Statistical significance has nothing to do with the main definition of significance – the quality of being important. It is a technical term more accurately defined to be 'when the probability that the observed data differs from the assumed model is very small [smaller than some arbitrary subjectively-selected cutpoint like 0.05].'

In epidemiology, it is common to use the term 'clinical significance' to imply importance, but often this is mistaken with 'statistical significance.' Statistical significance is not equivalent to scientific or clinical importance, relevance, meaningfulness, or other such synonyms. As stated by Bangdiwala (2009)[8], "*meaningful relationships can or cannot be statistically significant, just like non-meaningful relationships can or cannot be statistically significant.*" Smaller p-values do not necessarily imply the presence of larger effects, or effects that are more important, and larger p-values do not imply a small or unimportant effect. As illustrated in Table 1, a small effect can have a small p-value if the sample size is large or the variability is low, and a large effect can have a large p-value if the sample size is small or the variance is too large. With small sample sizes, meaningless effects can be statistically significant if the variability is low (Case #1), but are easily not significant with a slight increase in imprecision (Case #2). Large effects with high precision can have significance in very small sample sizes (Case #3), but large effects in small sample sizes can be non-significant with higher variability (Case #4). With large sample sizes, unimportant effects are significant despite slightly higher variability (Case #5), and require a lot of imprecision to get a non-significant result if the effect observed is large (Case #6).

Concluding recommendations

Since the p-value is a single index, following the ASA's statement, we strongly support that it cannot and should not be considered as the sole basis for scientific reasoning. Given the misuses and misconceptions concerning p-values, the recommendation is to present the estimate of the effect, provide a measure of uncertainty of the estimation (e.g. confidence interval), and interpret the results in terms of scientific importance.

Secondly, whether a p-value exceeds or not an arbitrary threshold (such as 0.05) cannot and should not be considered as defining the importance of the result. The technical statistical term 'significance' has been hijacked by the scientific and research community, and it is time it is rescued by us the statisticians.

The word ‘significance’ should only be used when referring to probability statements after a formal statistical test, i.e. reserved for use only in its statistical context. Other words in the English language can be used when wanting to highlight the importance of a result: important, meaningful, big, great, large, fantastic, crucial, influential, relevant, vital, awesome, and so forth.

Table 1: Hypothetical example data to illustrate the relationship of sample size, observed effects, and variability of the measurements, with p-values and statistical significance when testing the null hypothesis of no difference in means against a two-sided alternative of a difference, where a ‘meaningful’ or important difference is arbitrarily set at δ=5

Case	Sample size per group (n_1, n_2)	Observed difference δ in sample means (effect)	Standard deviation of the measurement per group (σ_1, σ_2)	Standardized effect (t statistic)	p-value
#1 – small sample size, meaningless effect, high precision	10, 10	1	1, 1	2.24	0.0382
#2 – small sample size, meaningless effect, lower precision	10, 10	1	2, 2	1.12	0.2783
#3 – very small sample size, meaningful effect, high precision	3, 3	6	1, 1	7.35	0.0018
#4 – small sample size, meaningful effect, low precision	10, 10	6	7, 7	1.92	0.0713
#5 – large sample size, meaningless effect, lower precision	100, 100	1	3, 3	2.36	0.0194
#6 – large sample size, meaningful effect, very low precision	100, 100	6	25, 25	1.70	0.0913

Conflict of interest:

The author declares that he has no financial or non-financial potential conflicts of interest.

Author’s affiliation:

[1] Professor, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC, USA.

Reference:

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose, The American Statistician. 2016; DOI:10.1080/00031305.2016.1154108 http://dx.doi.org/10.1080/00031305.2016.1154108

2. Fisher RA. Statistical tests. Nature. 1935;147:474. http://dx.doi.org/10.1038/136474b0

3. Lehmann EL. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association. 1993;88(424):1242-1249. http://dx.doi.org/10.1080/01621459.1993.10476404

4. Fisher RA. Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh. 1925. PMid:17246289

5. Fisher RA. The Arrangement of Field Experiments, Journal of the Ministry of Agriculture of Great Britain. 1926;33:503-513.

6. Fisher RA. Statistical Methods and Scientific Inference, 3rd ed., Collins Macmillan, London. 1973.

7. Merriam-Webster online dictionary http://www.merriam-webster.com/dictionary/significance [accessed 25 March 2016].

8. Bangdiwala SI. Significance or importance: what is the question? International Journal of Injury Control and Safety Promotion. 2009;16(2):113–114. http://dx.doi.org/10.1080/17457300902909296 PMid:19941207