Published in final edited form as:
Stroke. 2014 December ; 45(12): e244–e246. doi:10.1161/STROKEAHA.114.006138.

Some common misperceptions about p-values

Yuko Y. Palesch, PhD
Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Suite 303, Charleston, SC 29425

Keywords

p-value; alpha; statistical significance

A p-value<0.05 is perceived by many as the Holy Grail of clinical trials (as with most research in the natural and social sciences).Itis greatly sought after because of its (undeserved) power to persuade the clinical community to accept or not accept a new treatment into practice. Yet few, if any, of us know why 0.05 is so sacred. Literature abounds with answers to the question, “What is a p-value?” and how the value 0.05 was adopted, more or less arbitrarily or subjectively, by R. A. Fisher, in the 1920’s. He selected 0.05 partly because of the convenient fact that in a normal distribution, the five percent cutoff falls around the second standard deviation away from the mean.[1]

But little is written on how 0.05 became the standard by which manyclinical trial results have been judged. A commentary[2]ponders whether this phenomenon is similar to the results from the “monkeys in the stairs” experiment, whereby a group of monkeys were placed in a cage with a set of stairs with some fruit at the top. When a monkey went on the steps, blasts of air descended upon it as a deterrent. After a while, any monkey that attempted to get on the steps wasdissuaded by the group. Eventually, the monkeys were gradually replaced by new monkeys, but the practiceof dissuasion continued, even when the deterrent was no longer rendered. In other words, the new monkeys were unaware of the reason why they were not supposed to go up the steps, yet the practice continued.

In the following, I first review what a p-value is. Then, I address two of the many issues regarding p-values in clinical trials. The first challenges the conventional need to show p<0.05 to conclude statistical significance of a treatment effect; and the second addresses the misuse of p-values in the context of testing group differences in baseline characteristics in randomized trials. Many excellent papers and books have been published that address these topics; nevertheless, the intention of this paper is to revive and renew them (using a less statistical language) to aid the clinical investigators in planning and reporting study results.

Phone: 843-876-1917, Fax: 843-876-1923, paleschy@musc.edu.

Disclosures
The author is a DSMB member for a study of Brainsgate Ltd., and for a study by Biogen Idec Inc.

What is a p-value anyway?

We equate $p<0.05$ with statistical significance. Statistical significance isabout hypothesis testing,specifically of the null hypothesis (H_0) that means the treatment has no effect. For example, if the outcome measure is continuous, the H_0may be that the group difference in mean response (Δ) is equal to zero. Statistical significance isthe rejection of the H_0 based on the level of evidence in the study data. Note that failure to reject the H_0does not imply that $\Delta=0$ is necessarily true; just that the data from the study provide insufficient evidence to show that $\Delta\neq0$.

To declare statistical significance, we need a criterion. The alpha (also known as the Type I error probability or the significance level)is that criterion. The alpha does not change with the data. In contrast, the p-value depends on the data. Ap-value is defined as the probability of observing treatment effect(e.g., group difference in mean response)as extreme or more extreme(away from the H_0) if the $H_0$is true. Hence, the smaller the p-value, the more extreme or rare the observed data are,given the $H_0$to be true.The p-value obtained from the data is judged against the alpha. If alpha=0.05 and p=0.03, then statistical significance is achieved. If alpha=0.01, and p=0.03, statistical significance is not achieved. Intuitively, if the p-value is less than the pre-specified alpha, then the data suggest that the study result is so rare thatit does not appear to be consistent with H_0, leading to rejection of the H_0. For example, if the p-value is 0.001, it indicates that, if the null hypothesis were indeed true, then there would be only a 1 in 1,000 chance of observing data this extreme. So either very unusual data have been observed, or else the supposition regarding the veracity of the $H_0$is incorrect. Therefore, small p-values (less than alpha) lead to rejection of the H_0.

In the Interventional Management of Stroke (IMS) III Trial that compared the efficacy of IV tPA (N=222) and IV tPA plus endovascular (N=434) treatment of acute ischemic stroke, the alpha was specified as 0.05. The unadjusted absolute group differencein the proportion of the good outcome, defined as the modified Rankin Scale score of 0-2, was 2.1% (40.8% in endovascular and 38.7% in IV tPA)[3]. Under the normal theory test for binomial proportions, this yields a p-value of0.30, meaning that if the H0were true (i.e., the treatment did not work), there would be a 30% chance of observing a difference between the treatment groups at least as large as 2.1%. Since this is not so unusual, wefail to reject H_0: $\Delta=0$ and conclude that the difference of 2.1% is not “statistically significant.”

Thinking outside the “p<0.05” box

Another interpretation of the alpha is that it is the probability of rejecting the H_0 when in fact it is true. In other words, alpha is the false positive probability.Typically, we choose alpha of 0.05, and hence, our desire to obtain $p<0.05$. There is nothing magical about 0.05. Why not consider the risk (or cost) to benefit ratio in the choice of the false positive probability the research community is willing to tolerate for a particular study? For some studies, should one consider a more conservative (like 0.01) or more liberal (like 0.10) alpha? In the case of comparative effectiveness trial, where two or more treatments, similar in cost and safety profile, that have been adopted in clinical practice are tested to identify the “best” treatment, one might be willing to risk a higher likelihood of a false positive finding

with alpha of, say 0.10. In contrast, if a new intervention to be tested is associated with high safety risks and/or that isvery expensive, one would want to be sure that the treatment is effective by minimizing the false positive probability to, say 0.01.For a certain Phase II clinical trial, where the safety and efficacy of a new treatment is still being explored, one can argue for a more liberal alpha to give the treatment a higher level of the benefit of doubt, especially when the disease or condition have only a few, if any, effective treatment options. If it should pass, it would be weeded out in a Phase III trial with a more stringent significance level. Also, if the H0is widely accepted as true (perhaps, for example, in the case of hyperbaric oxygen treatment for stroke), then one might wish to be more sure that rejecting the H0implies that the treatment is effective by using alpha of 0.01 or even lower. Of course, this means a larger study has to be conducted.

While proposing to use anything greater than an alpha of 0.05 may be challenging, especially for studies to be submitted to the US Food and Drug Administration for New Drug Application approval, scientifically sound rationaleand experienced clinical judgment should encourage one to think outside the box about the choice of the alpha.In doing so, one should ensure that scientific and ethical rationale is the driving argument for proposing a larger alpha, and not only the financial savings (as a result of smaller required sample size with a larger alpha).

P-values in the group comparison of baseline characteristics in clinical trials

Typically, primary publications of many clinical trials include in “Table 1” a long list of baseline characteristics of the study sample and their summary statistics (e.g., mean and standard deviation;median and interquartile range; or proportions). In addition, many include p-values associated with statisticaltests comparing the groups or denote with a variety of asterisks the variables where the comparison yields $p<0.05$, $p<0.01$ and $p<0.001$. Some authors assume that the journal editors require them. In the instructions to authors of prospective *New England Journal of Medicine (NEJM)* manuscripts, it states under the statistical methods:

> “For tables comparing treatment groups in a randomized trial (usually the first table in the trial report), significant differences between or among groups (i.e. $P < 0.05$) should be identified in a table footnote and the P value should be provided in the format specified in the immediately preceding paragraph. The body of the table should not include a column of P values.” (http://www.nejm.org/page/author-center/manuscript-submission; obtained on August 18, 2014)

Meanwhile, according to the current CONSORT 2010 guidelines on the publications of clinical trials:

> “Unfortunately significance tests of baseline differences are still common; they were reported in half of 50 RCTs trials published in leading general journals in 1997. Such significance tests assess the probability that observed baseline differences could have occurred by chance; however, we already know that any differences are caused by chance. Tests of baseline differences are not necessarily

> wrong, just illogical. Such hypothesis testing is superfluous and can mislead investigators and their readers. Rather, comparisons at baseline should be based on consideration of the prognostic strength of the variables measured and the size of any chance imbalances that have occurred." (http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data; obtained on August 18, 2014)

Thetwo are somewhat contradictory: one (NEJM) requiring statistical tests be performed on the baseline characteristics and the other (CONSORT) discouraging such tests.

Recall that p-values are associated with hypothesis testing. The hypotheses that are tested for these baseline characteristics evaluate whether the differences between the groups are statistically significant,butthat does not necessarily equate to clinical significance or relevanceof the difference. Note that the p-value is partially influenced by sample size. Generally, the larger the sample size, the easier it is to obtain a smaller p-value from the data for the same difference. For any study with large enough sample size, statistical significancecan beachieved; however, the observed mean group differenceis not necessarily clinically relevant. Conversely, one may note a clinically relevant difference in a baseline characteristic, but the p-value from the test may not reach statistical significance with a small sample size. Therefore, an important clinical difference in a baseline characteristic may be overlooked.

Suppose in a large (n=2,100 per group) clinical trial of acute stroke to detect a difference of 5% in good outcome between two treatment groups, the "Table 1" shows the mean baseline systolic blood pressure of 125 and 120 mmHg, each with standard deviation of 15 mmHg. The difference is 5 mmHg, and thet-testyields $p<0.01$. But one could hardly argue that this difference is clinically significant. In contrast, suppose a small study (say, n=40 in each group)to test intensive serum glucose control in acute stroke patients had enrolled subjects with history of diabetes: 20% in one group and 33% in the other. The chi-square test yields $p=0.20$ – not a statistically significant difference at the alpha of 0.05. Nevertheless, a 13% difference in the proportion of subjects with history of diabetes is likely to be a clinically important factor to consider in the analysis and interpretation of the primary outcome. In other words, p-values are meaningless at best, and potentially misleading,to ascertain whether the treatment groups arebalanced in the baseline characteristics. The same issue of seeking statistical significance without consideration for clinical relevance also applies to analyses of outcomes data. Many papers have been published in both statistical and clinical journals addressing thistopic and will not be addressed further here[4,5].

Recommendation

So what are we to do? Should we stop using p-values altogether?No, but additional information, such as the pre-specified minimum clinically important difference (MCID),the observed group differences, and their confidence intervals (CIs) will enable other investigators to better assess the level of evidence for or against the treatmenteffect since they provide a range of plausible values for the unknown true difference between the groups.

For example, in the IMS III Trial[3], the study investigators pre-specified MCID of 10%. Thereported adjusted (for baseline NIHSS score per the study analysis plan) difference was

1.5% with95% CI of (-6.1%, 9.1%).Since the 95%CI includes 0, the result is not statistically significant at alpha=0.05.In addition, if the CI had included 10%, the study result can be interpreted as “inconclusive” because 10% may be a plausible value for the true but unknown group difference; otherwise, the study could be viewed as “negative”. Therefore, such information will allow the readers to apply their knowledge, experience, and judgment on the importance and relevance of the study results beyond whether it is statistically significant or not.

Conclusion

In conclusion, R.A. Fisher did not intend for the p-value, much less $p<0.05$, to be the be-all and end-all of an experiment (or a clinical trial). He meant it as a “guide” to determine whether the study result is worthy of another look through replication. In spite of increasingly vocal criticism of oursole dependence on p-values by thebiostatisticaland even some clinical communities, it will take some time to change the culture, but the change should be embraced.

Acknowledgments

The authorwould like to thank the two anonymous reviewers for their thorough and constructive comments to clarify and improve the discussions in this paper.

Sources of Funding

This work was partially supported by NIH-U01-NS087748.

References

1. Cowles M, Davis C. On the origins of the 0.05 level of statistical significance. American Psychologist. 1982; 37:553–558.
2. Kelly M. Emily Dickinson and monkeys on the stair or what is the significance of 5% significance level. Significance. 2013; 10:21–2.
3. Broderick JP, Palesch YY, Demchuk AM, Yeatts SD, Khatri P, Hill MD, et al. Endovascular Therapy after intravenous t-PA versus tPA alone for stroke. NEJM. 2013; 368:893–903. [PubMed: 23390923]
4. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statist Med. 2002; 21:2917–2930.
5. Senn S. Seven myths of randomization in clinical trials. Statist Med. 2013; 32:1439–50.