**References:** Are all major references included?
- Are all references cited completely and in the desired format of the journal?
- Are references chosen directly related to the study?

The peer review process is generally similar for all journals. Once an author submits a manuscript, it is initially reviewed by an editor of the journal to determine its suitability according to the guidelines set by the editorial policy. The manuscript could be rejected without additional review if the content does not fall within the scope of the journal, if it does not follow editorial policy and procedural guidelines, or it has already been accepted in another journal (in press). If the manuscript is not rejected when first received, it is then sent out for review to a minimum of two additional reviewers in the journal's list of reviewers who are considered experts in the content of the paper. This process is usually a closed review adopted by most journals and can be a single-blinded review where the reviewers' identities are withheld from the authors but the reviewers are aware who wrote the paper they are evaluating, or a double-blinded review where the identity of the authors is also concealed during the review process.[5] When the chosen reviewers have accepted their assignment, they are given a time period to review the paper, usually with the help of a checklist similar to the sample given above. The reviewers return their recommendations and report to the editor who assesses them collectively and then makes a decision whether to reject the manuscript outright, to withhold judgment pending major or minor revisions, to accept it pending satisfactorily completed revisions, or to accept it as written (which is rare).[1] For a manuscript requiring revisions, the authors have to submit the revised manuscript incorporating the recommendations of the reviewers. Once the manuscript has been revised satisfactorily, it is accepted and prepared for publication that may take several months.

The review process generally does not change the basic nature of the submitted manuscript; rather, it assists the authors in improving the presentation of their work. This can only happen when knowledgeable reviewers take time to participate in the peer review process and evaluate submissions with care and sensitivity.[1]

*For in-depth discussion of the peer review process, please refer to reference 1.*

REFERENCES
1. Voight ML, Hoogenboom BJ. Publishing your work in a journal: understanding the peer review process. *Int J Sports Phys Ther* 2012;7:453-460.
2. Burnham JC. The evolution of editorial peer review. *JAMA* 1990;263:1323-1329.
3. International Committee of Medical Journal Editors. *Recommendation for the conduct, reporting, editing, and publication of scholarly work in medical journals*. www.icmje.org (updated December 2015).
4. Gannon F. The essential role of peer review. *EMBO Reports* 2001;2:743.
5. Ware M. Peer review: Recent experience and future directions. *New Rev Information Networking* 2011;16:23-53.

# WHEN TO USE P VALUES & CONFIDENCE INTERVALS FOR REPORTING INTERGROUP COMPARISONS

Patricia M. Khu, MD, MS
Presented at APAME 2015 Manila

Reporting research results usually requires the investigator to subject the collected data to a statistical procedure determining the degree to which the data are consistent with the specific hypothesis under investigation. This is the test of significance for the p value.

There are six features common to significance tests.[1] First, there is a hypothesis about the population; that there is no difference between the two groups to be compared or the null hypothesis ($H_o$). Second, the sample taken from the population is random. Third, there is a set of comparable events (2 x 2 tables). Fourth, the probability distribution of the test statistic is based on the assumption that the null hypothesis ($H_o$) is true and the sampling uncertainty is random. Fifth, there is a ranking of all possible outcomes in a set of comparable events according to their consistency with the null hypothesis. Lastly, the probability that sample uncertainty, called chance, would produce outcome no more consistent with $H_o$ than the outcome observed is calculated. This probability is called the significance level of the data with respect to $H_o$.

The resulting p value obtained is the likelihood

that the result observed is due to random occurrence if $H_o$ is true. It usually does not take on an exact value; rather, it is more correctly denoted as a probability greater than or less than a given value. The significance level is the value of p at which we are willing to reject $H_o$ even if it is correct. The most commonly accepted level of significance is $\alpha = 0.05$ or 5%; this significance limit is usually specified in advance. This means that the probability of observing the results obtained even if there were truly no treatment effect (if $H_o$ was true) is less than 5%. In other words, it is quite possible that we would be wrong in rejecting the null hypothesis but this would happen only 5 times out of 100 (or 1 out of 20) over repeated studies using different samples of the same size.[2]

Smaller p values correspond to stronger evidence that the results are significant and the probability is small that the difference is due to chance. There is also less likelihood of committing a Type I ($\alpha$) error that occurs when we conclude from the significant findings that there is an effect (reject $H_o$) when in fact there is no true effect.[3] It is said that approximately 1 in 20 significant findings will be spurious or arising from chance.[3]

If the p value is less than the pre-defined limit, the result is designated as "statistically significant" and $H_o$ is rejected and the alternative hypothesis ($H_1$) is accepted. P value alone, however, does not give any direct statement about the direction or the size of the difference or relative risk between different groups.[4] A directional test is a one-tailed test that looks for a treatment difference in one direction only, such as the study drug is more effective than the control and not vice versa. A non-directional test is a two-tailed test that looks for treatment difference in either direction, either the study drug is more effective than the control or the controlled treatment is better than the study drug. Some statisticians believe that p value is more useful when the results are not significant.[4]

*What does "not significant" really mean?* When the test statistic is bigger than 95% of the values that would occur if the treatment has no effect, the null hypothesis is rejected and we conclude that treatment has an effect and is statistically significant. When the test statistic is not big enough to reject the null hypothesis, we conclude that the test failed to demonstrate an effect and report as not statistically significant. It has been observed that many researches discussed the results not statistically significant as if there is no treatment effect when in fact the test just failed to show an effect.

The other error, Type II ($\beta$), is much more common and occurs when we conclude from non-significant findings that there is no effect (reject $H_1$) when in fact there is a real effect. This can happen when the sample size is too small to detect a significant difference when one exists. Power calculations are designed to minimize this error. The ability to detect a treatment effect with a given level of confidence depends on 3 parameters; namely, 1) the size of the treatment effect; 2) the variability within the population; 3) the size of the samples used in the study.[4] Bigger samples are better able to detect an effect compared to smaller samples. Thus, studies on therapies involving few subjects (small sample) that failed to show an effect may lack the statistical power to detect the effect. Conversely, in large databases with numerous variables, very large sample sizes will tend to pick up statistically significant differences in variables, even if the difference is minute.[4] Hence, it is prudent to always consider what is being compared, the cost of treatment, the potential side effects, and the overall benefit to the population under study.

*Statistical versus clinical significance*. Statistical significance may not always translate into clinical significance. Doing significance testing simply asks whether the data collected in a study are compatible with the notion of no difference between the two groups compared. When we reject equivalence, this does not mean that we accept that there is an important difference. A large study may identify as statistically significant a small difference. As clinicians, we also have to consider the clinical relevance. In assessing the importance of significant results, the size of the effect (not just the size of the significance) also matters.[4]

Using confidence interval (CI) in reporting research results gives not just the size and direction of the effect, but also the level of confidence that the point estimate or true parameter is within the confidence limits. The point estimate provides the best approximation to the true value, but does not provide any information on how exact it is. This is provided by the confidence interval that described the probability that the true value is within a given range. The 95% CI is usually selected; meaning that the interval covers the true value in 95 out of 100 studies performed.

The size of the confidence interval depends on the sample size and the standard deviation of the study groups. A larger sample size will have a narrower CI and the conclusion is more certain. A small sample size will have a wider CI, higher dispersion, and the conclusion is less certain. Values within the CI but near the confidence limits are less probable than values near the point estimate. Values below the lower limit or above the upper limit are not excluded but are improbable and with 95% CI, each probability is only 2.5%.[4]

The size of the confidence interval is also influenced by the selected level of confidence; 99% CI is wider than 95% CI, indicating that the wider the interval the higher the probability of including the true value.

Conclusion about statistical significance is also possible in confidence interval; if the zero value is not within the interval, it is said to be statistically significant.

In summary, confidence interval provides information on the statistical significance, the direction and strength of the effect, allowing decision on clinical relevance of the results. P value, on the other hand, allows quick decision whether the value is statistically significant or not, but can be misleading, leading to decisions solely based on statistics.[4]

*For in-depth discussion of p values and confidence intervals, please refer to the references below where most of the information in this article were obtained.*

REFERENCES

1. Glantz ST. *Primer of Biostatistics,* 7th ed. Singapore: McGraw Hill; 2012.
2. Dawson B & Trapp RG. *Basic & Clinical Biostatistics,* 4th ed. New York: McGraw Hill; 2004.
3. Browner WS. *Publishing & Presenting Clinical Research,* 3rd ed. Phila, PA: Lippincott Williams & Wilkins; 2012.
4. Dawson GF. *Easy Interpretation of Biostatistics: The Vital Link to Applying Evidence in Medical Decisions.* Phil, PA: Saunders, Elsevier; 2008.