Clin Orthop Relat Res (2010) 468:2297–2298
DOI 10.1007/s11999-010-1402-9

IN BRIEF

In Brief

The P Value

What Is It and What Does It Tell You?

Frederick Dorey PhD

Published online: 25 May 2010

Background

In medical papers today there usually are several statements based on the result of hypothesis tests presented, along with the associated p values. For example, a recent article by van Raaij et al. [1] compared the use of laterally wedged insoles with valgus braces for reduction of pain or improving function in selected patients with osteoarthritis. One of the statements made in that randomized study was that "At 6 months, 71% of patients in the insole group complied with the treatment, which was greater ($p = 0.015$) than 45% for the brace group" [1].

Question

How does this hypothesis test address the issue of compliance between these two approaches, what information is supplied by the associated p value, and how should it be interpreted?

Discussion

The primary purpose of an hypothesis test is to decide if the results of a study, based on a small sample, provide enough evidence against the null hypothesis (denoted by H_0), so that it is reasonable to believe that in a larger target population, H_0 is false, thus accepting the associated alternative hypothesis (denoted by H_1) as being true. The null hypothesis for this situation states that there is no meaningful clinical difference between the two treatment approaches in terms of the percent compliance in the target population [1]; formally stated, the expected difference between the percent compliance in the two samples should be zero. The alternative hypothesis is that there is a meaningful difference in percent compliance between the two treatments in the target population. van Raaij et al. reported a large difference of 26% between the two treatments [1]. The hypothesis test is designed to help determine if a 26% difference is so large and the resulting p value of 0.015 so small that we should reject H_0.

First and foremost, a p value is simply a probability. However, it is a conditional probability, in that its calculation is based on an assumption (condition) that H_0 is true. This is the most critical concept to keep in mind as it means that one cannot infer from the p value whether H_0 is true or false. More specifically, after we assume H_0 is true, the p value only gives us the probability that, simply owing to the chance selection of patients from the larger (target) population, the clinical experiment resulted in a difference in the samples, as large or larger, than the actual 26% observed [1]. If a resulting small p value suggests that chance was not responsible for the observed difference of 26% and the randomization of patients, as in this case [1], makes the presence of bias unlikely, then the most likely conclusion is that in the target population the treatments must produce different compliance results.

Thus a p value is simply a measure of the strength of evidence against H_0. A study with a $p = 0.531$ has much less evidence against H_0 than a study with a $p = 0.058$. However, a study with a $p = 0.058$ provides similar evidence as a study with a $p = 0.049$ and a study with a $p = 0.049$ also has much less evidence than one with a

F. Dorey (✉)
Department of Pediatrics, Children's Hospital Los Angeles,
4650 Sunset Boulevard, Mailstop 54, Los Angeles,
CA 90027, USA
e-mail: fdorey@chla.usc.edu

$p = 0.015$. Although a very small p value does provide strong evidence that H_0 is not true, a very large p value, even as large as 0.88, does not provide real evidence that H_0 is true. For example, the alternative hypothesis might in fact still be true but owing to a small sample size, the study did not have enough power to detect that H_0 was likely to be false. This notion, referred to as the power of the test, will be discussed later.

Authors sometimes take a formal approach in evaluating the results of an hypothesis test. An artificial cut point is chosen, called the significance level, and the result is called statistically significant if the p value is less than the significance level leading to the rejection of the null hypothesis. Although 5% usually is taken as the significance level, there is no real scientific reason for choosing that versus any other small value. Always rejecting H_0 when p is less than 5% results in an incorrect rejection of the null hypothesis 5% of the time. However, as there is no real practical difference between a p value of 0.06 and 0.045 from a probability point of view, it is difficult to understand why this rigorous interpretation has become the standard today. In the study by van Raaij et al. [1], the result is statistically significant at the 5% level as $p = 0.015$. However, if a similar difference of 26% had been found in a study with only 24 patients with insoles and 22 patients with braces, the associated p value (chi square test) would have been 0.081, a result that would be called not statistically significant. That would not have meant that there was no difference between the two treatments, but only that, with the given small sample size there is not enough evidence to reject H_0.

Myths and Misconceptions

There are several misconceptions associated with the interpretation of a p value. One of the most common ones is that the p value gives the probability that H_0 is true. As mentioned earlier, as the p value is calculated based on an assumption that H_0 is true it cannot provide information regarding whether H_0 is in fact true. This argument also shows that first, p cannot be the probability that the alternative hypothesis is true. Second, the p value is very dependent on the sample size. Third, it is not true that the p value is the probability that any observed difference is simply attributable to the chance selection of subjects from the target population. The p value is calculated based on an assumption that chance is the only reason for observing any difference. Thus it cannot provide evidence for the truth of that statement. The concept of a p value is not simple and any statements associated with it must be considered cautiously. A wealth of information and references concerning these and other misinterpretations of p values can be found on the WEB. Finally, it is important to reemphasize that if the result of an hypothesis test is that a difference was not statistically significant, it does not mean that there is no difference between the treatment groups in the target population.

Conclusion

The only question that the p value addresses is, does the experiment provide enough evidence to reasonably reject H_0. The actual p value always should be indicated when presenting the results of a clinical study, as the p value as a probability, provides a continuous measure of the evidence against H_0. In the study by van Raaij et al. [1], randomization of the patients, the observed difference of 26% between the treatments, and the very small p value of 0.015 suggest that rejection of the null hypothesis is reasonable. Finally, the question of just how much difference might exist between the treatments in the target population is not directly addressed by the p value. Although 26% is a reasonable estimate of that difference, a confidence interval is more appropriate to address that question.

Reference

1. van Raaij TM, Reijman M, Brouwer RW, Bierma-Zeinstra SM, Verhaar JA. Medial knee osteoarthritis treated by insoles or braces: a randomized trial. *Clin Orthop Relat Res*. 2010 Feb 23. (Epub ahead of print)