

Use and Misuse of the P-Value

Emmanuel Lesaffre, Dr.Sc.

Abstract

The p-value is a widely used tool for inference in clinical studies. However, despite the numerous books and papers published on the basics of statistical inference and, thus, on the p-value, there still seems to be a need to highlight what message the p-value exactly contains (and what it does not). In this article, the basic concepts and the different misconceptions regarding the p-value will be highlighted and illustrated with a clinical trial in osteoarthritis. It will also be shown that the (95%) confidence interval is to be preferred over the p-value as a statistical inference tool.

The p-value remains the most widely utilized tool to draw inference from medical studies. However, despite its popularity, the foundational aspects of the p-value appear to be poorly understood, and misuses are abundant in the medical literature. In this paper, the basics behind the p-value are reviewed, with illustrations mainly taken from Bingham and colleagues.¹ These investigators performed two randomized, three-arm double-blinded clinical trials comparing placebo with etoricoxib 30 mg qd (ET), a selective COX-2 inhibitor, and with celecoxib 200 mg qd (CE), the standard treatment dose in the management of osteoarthritis of the knee and hip over a 12-week period. Each study consisted of two parts. In the first study, 559 patients were randomized to ET, CE, or placebo (PL). In the second study, 608 patients were randomized to one of the three medications. Patients who successfully completed part I (comparing active versus placebo treatment)

were enrolled directly into part II, an active comparator 14-week follow-up study comparing ET and CE. This paper focuses on the trial type, the classical superiority test. Tests for equivalence and non-inferiority are treated by Lesaffre in this same volume.² Three primary end points were based on the time-weighted average change from baseline over 12 weeks, using the following scales: 1. the WOMAC (Western Ontario and McMaster Universities) Pain Subscale (WOMACPA: average number of questions, 1 to 5; a large value implies severe pain); 2. the WOMAC Physical Function Subscale (WOMACPH: average of questions eight to 24; a large value implies extreme difficulty); and 3. the Patient Global Assessment of Disease Status (PGADS), measured by a visual analogue scale from 0 ("very well") to 100 mm ("very poor"). Further details can be found in the original publication.¹

The P-Value and the Statistical Test

To exemplify the concept of the p-value, take the change of WOMACPA, between ET and PL, from baseline to 12 weeks of treatment. Table 2A of Bingham and colleagues¹ reports an average value of WOMACPA of 67.4 (66.6) and 39.6 (54.2) for ET (PL) at baseline and after treatment, respectively. This translates into an average change from a baseline for ET of 27.8 and for PL of 15.4, and hence the difference in average (DA) change from baseline between ET and PL is -15.07, in favor of ET. The reported p-value is smaller than 0.001, which is explained further below.

Assume that treatments ET and PL are equally good, i.e., the true difference in averages is zero ($\Delta = 0$). If $\Delta = 0$, one can expect that DA is about zero, but also that it will fluctuate (around zero) when the study is repeated. The question is then: When is DA large enough to conclude with high certainty that ET is different in performance than PL? However, for the calculation of the p-value, we pose a different question and ask (for the current study): how rare

Emmanuel Lesaffre, Dr.Sc., is from the Biostatistical Centre, Catholic University of Leuven, Leuven, Belgium, and Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands.

Correspondence: Emmanuel Lesaffre, Dr.Sc., Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands; e.lesaffre@erasmusmc.nl.

is a result like the DA = -15.7 (or larger in absolute value) if $\Delta = 0$? The statement that PL is equally as good as the ET is called the null hypothesis and is denoted by $H_0: \Delta = 0$. A value of $p = 0.05$ means that DA belongs to the 5% extreme results that could happen under H_0 (if H_0 is true). When $p = 0.01$ ($p < 0.001$), then again the obtained DA belongs to the 5% most extreme results that could happen under H_0 (if H_0 is true) and only 1% (less than 0.1%) of the results are more extreme.

The calculation of the p-value is based on probability arguments and is the outcome of a statistical test. Statistical tests are designed to verify a null hypothesis, as above. In general, H_0 corresponds to a hypothesis that we wish to reject. In our example, we wish to reject that PL is as good as ET and conclude that the alternative hypothesis $H_A: \Delta \neq 0$ is true. Thus, we need to decide from which value of DA, or equivalently, from which p-value, we wish to reject H_0 . While it is customary to establish the α level as 0.05, called the significance level, as the threshold value for this decision, more specifically, the following applies:

- When $p < (\alpha =) 0.05$, the result is considered to be too extreme to believe that H_0 is true and we reject H_0 . In other words, we do *not* believe that $\Delta = 0$, and we call the result significant at 0.05, as for our example above.
- On the other hand, when $p \geq 0.05$, we conclude that the result is not that extreme and could (easily) have happened when $H_0: \Delta = 0$, implying that H_0 cannot be rejected. Thus, we conclude that the observed difference from zero is a play of chance that does not necessarily indicate $\Delta = 0$. We say that the result is not significant at 0.05. A nonsignificant result was obtained when comparing ET with CE for WOMACPH, yielding $p = 0.367$. Thus, we cannot claim that ET is different in performance than CE.

The above alternative hypothesis H_A is called two-sided, since $H_A: \Delta \neq 0$ implies either $\Delta > 0$ or $\Delta < 0$. In that case, also, the statistical test, the significance level, and the p-value are called two-sided. In a randomized controlled trial (RCT), most often the tests are two-sided, because it allows an investigator to detect whether one treatment is better or worse than the other treatment. Occasionally, the alternative hypothesis (and statistical test, significance level, p-value) is one-sided, e.g., $H_A: \Delta < 0$ or $H_A: \Delta > 0$. However, in a regulatory context, one-sided significance tests are discouraged because they do not allow for the detection of safety issues when aiming to prove a better efficacy of one treatment over the other.

There exists a large variety of statistical tests that differ with the type of response, the aim of the comparison, the experimental set-up, and the kind of question that is asked, etc. It goes beyond the objectives of this paper to provide an overview of available statistical tests. Nevertheless, some well known tests are: 1. unpaired (paired) t-tests for the comparison of the means of two independent (dependent) groups, assuming that the data have a normal distribution in each

group with equal variance; 2. the chi-square test (McNemar test) for the comparison of two proportions of two independent (dependent) groups; 3. analysis-of-variance tests for the comparison of more than two independent means; and 4. nonparametric tests, such as the Wilcoxon test and the Kruskal-Wallis test, which are not based on distributional assumptions of the data, in contrast to the t-tests, etc.

A statistical test (on $\alpha = 0.05$) is, in fact, a decision rule, which will reject H_0 in case $p < 0.05$. This rejection is incorrect in the case H_0 is, in fact, true. In such a situation, one commits an error, called a type I error. On the other hand, a type II error arises when one fails to reject H_0 ($p \geq 0.05$) when there is truly a different effect between the two treatments. The probability of committing a type I error, p (type I error), is called the false positive rate. It is exactly equal to the α significance level, and therefore is under control when a statistical test is performed (but only one; see below, multiple testing problem). The probability of committing a type II error, denoted by β , though, is not automatically controlled and to minimize this error the sample size needs to be adjusted, as we shall see below.

Misuses of the P-Value

Multiple Testing

If H_0 is true, one statistical test employing a significance level of 0.05 will reject this null hypothesis with probability exactly equal to 0.05. Thus, we run a 5% risk of claiming that H_0 is wrong when it is in fact true. Suppose 100 studies are executed, then, on average, for five studies, a wrong conclusion will be drawn under H_0 , namely that the two treatments will be claimed to have had a different effect. Also, the probability that in at least one study a wrong conclusion will be drawn will be practically equal to 1. Since false positive trial findings could lead to the approval of inefficient drugs, regulatory agencies (FDA, Food and Drug Administration, United States; EMEA, European Medicines Agency, London, United Kingdom) mandate a strict control of the overall false-positive rate (to, at most, 0.05). Thus, if multiple tests are performed, a mechanism is needed to control the probability of drawing such a false conclusion. If no such correction is performed, then one has a multiple testing problem. A multiple testing problem occurs when:

1. Two treatments are compared for several end points; the two treatments are called significant at 0.05, when at least one of the end points results in a p-value smaller than 0.05. In our osteoarthritis example, there are three primary end points (WOMACPA, WOMACPH, and PGADS), hence, there is a potential problem.
2. More than two treatments are compared. In our osteoarthritis example, there are three treatments: ET, CE, and PL.
3. Two treatments are compared in several subgroups, which happens, for example, when treatments are compared in subgroups of males, females, patients younger than 65 years, or patients older than 65 years,

etc., and it is concluded that the treatments are different when, in one or more subgroups, there is a significant result.

4. Two treatments are compared at several time points, e.g., when, at regular time intervals, it is verified whether there is a different effect between two treatments.
5. Several statistical tests are employed to evaluate the difference between two treatments, and one infers a significant difference if one of the tests yields a significant outcome, etc.

Possible solutions for the problem of multiple testing are the following: A first and evident solution is to choose only one primary end point such that the risk is automatically 0.05. A next solution could be, if multiple end points need to be considered, then to avoid multiple testing, one could construct a combined end point based on clinical or statistical reasoning. For instance, in a trial to treat acute myocardial infarction (MI) patients, three binary end points might be of interest: cardiac mortality (Y/N), re-infarction (Y/N), and rehospitalization (Y/N). Then one could construct a combined binary end point that is equal to "1" if at least one of the three events has happened and "0" if none happened. A statistical approach to combining the three end points would consist of calculating a weighted average of the three binary endpoints, with the weights determined using a statistical procedure. A third way to deal with multiple testing is to correct for it. Suppose that $\alpha = 0.05$ for each test separately. If we claim that the two treatments are statistically significantly different when minimally even one of the three p-values is smaller than 0.05, then the risk of wrongly claiming that the two treatments have a different efficacy is approximately equal to the sum of the individual risks and, thus, approximately equal to 0.15 (in fact, smaller than 0.15). In order to control the overall error rate (also called the familywise error rate), one option is to reduce the significance level for each test to $0.05/3$, thereby reducing the familywise error rate to maximally 0.05. This is called the Bonferroni adjustment. When this adjustment is applied to a general number of tests, say k , each individual p-value should be smaller than α/k in order to claim an overall statistical significant difference at 0.05 between the two treatments. In Bingham and coworkers,¹ the important results remain significant after a Bonferroni adjustment.

More accurate approaches than the Bonferroni adjustment are those of Simes, Holm, Hochberg, and Hommel, of which some belong to so-called closed testing procedures (see Moyé³). In the particular case of repeatedly testing in time, i.e., when performing interim analyses, dedicated procedures have been worked out to control the overall type I error. The two most popular procedures are those of O'Brien-Fleming and Pocock (see Proschan and associates⁴).

Interpretation of a Nonsignificant Result and Sample Size Calculation

In case of a nonsignificant result ($p \geq 0.05$), the null hypothesis is not rejected. In that case, the study is called negative.

Despite such a nonsignificant result, some investigators claim a trend in the data toward a significant result, thereby wrongly giving the impression that with a somewhat bigger sample size the study would certainly have been positive. Further, too often one interprets a nonsignificant outcome as a proof that there is no difference between the two treatments. There is the following saying: absence of evidence is not evidence of absence. In other words, it is not that if we are unable to show a difference that there is, indeed, no difference. In fact, one can never show the equal effect of two treatments with statistical tools. For this reason, we avoided stating above that "H₀ is accepted" in case of a nonsignificant result, rather we stated that we could "not reject H₀."

Negative studies often occur because the sample size is too small, and thereby the probability of making a type II error is too large. Thus, one needs to ensure at the planning of the study to have a small β , or equivalently, a large $1-\beta$, called the power of the study. For a given sample size, the power depends on: 1. the primary endpoint; 2. the statistical test; 3. significance level; and, most importantly, 4. the assumed clinically relevant difference Δ_S ("S" stands for superiority, notation consistent with the report by Lesaffre²) between the two treatments. Power also depends on other factors, such as the control rate, when comparing two proportions or the pooled standard deviation, in the case of comparing two means. In practice, one determines a sample size that results in a power of at least 0.80. This implies that with an 0.80 probability, one will be able to reject the null hypothesis for the given Δ_S , assuming that the other values were chosen correctly. The practical calculation of a sample size is highly technical and is best done using a software package.

In Bingham and colleagues,¹ the following statement is given with respect to the sample size: "With 200 patients each in the etoricoxib and celecoxib groups and 100 patients in the placebo group, each study provided an overall power of at least 87% to satisfy the primary hypothesis of non-inferiority between actives and of actives demonstrating superiority over placebo." Their description of the sample size falls a bit short, however, in that they did not report enough details for repeating the calculations, e.g., for the comparison of the two drugs (ET and CE) versus placebo; Δ_S was not specified.

Statistically Significant versus Clinically Relevant

Table 1 in Bingham and coworkers¹ shows the demographics of the two studies at baseline and evaluates the balance between the three arms by showing the descriptive statistics in each arm. No statistical test was applied to check for imbalance, and this is the correct approach. Indeed, at baseline none of the patients had received treatment. Since at that time the patients were only randomized to treatments without having received treatment, one can be sure that the null hypothesis of no treatment effect applies. In that case, one can predict how many statistical tests will be significant

at 0.05, i.e., with 100 independent tests, on average, five. Clearly, significant differences are to be expected, and, therefore, significance tests to verify balance at baseline are useless.

A statistically significant result is not necessarily clinically relevant. Let us take the following fictive example. Two drugs (A and B) are compared to treat acute MI patients, and the response is a 1-year mortality. A study has been set-up with twice 400 patients (i.e., 400 patients in two treatment groups). The results were 2% mortality for A and 10% for B, yielding a highly significant result with a chi-square test ($p < 0.01$). However, suppose that, in the same settings, twice 100,000 patients had been treated and the results were: A: 0.002% and B: 0.0010%. Again, a highly significant p-value would have been obtained with the chi-square test, but the net clinical benefit of A versus B is now 100 times lower than in the previous study. In fact, one can show that for each (small) $\Delta (\neq 0)$, there is a sample size such that H_0 is rejected with high probability.

The (95%) Confidence Interval

While the p-value is still the most popular tool with which to critically interpret the results of an RCT, its difficulties in use and interpretation have been gradually recognized in the international literature. Indeed, high quality journals, such as the New England Journal of Medicine and The Lancet, nowadays prefer the reporting of the 95% confidence interval (CI).

In our example, the observed difference in averages (DA = -15.07) is an estimate of Δ , the true difference in averages. But, clearly, DA is only an estimate, and we are still uncertain about the true difference. The 95% confidence interval (CI) is an expression of that uncertainty. More specifically, the 95% CI is the interval that contains, with 0.95 probability, the true difference (Δ). Thus, when this interval is narrow, we have a good idea of Δ . For WOMACPA, Bingham and associates¹ reported the 95% CI for the following differences: 1. CE – ET: [-7.02, 0.77]; 2. ET – PL: [-19.72, -10.41]; and 3. CE – PL: [-16.57, -7.32]. The interpretation of the 95% CI is much easier than that of the p-value. For example, the third CI tells us that CE treatment induces a truly higher decrease on the average WOMACPA than placebo, and that Δ lies with a 0.95 chance between -16.57 and -7.32.

Further, there is a relationship between the 95% CI and

the p-value. Namely, when the 95% CI of Δ does not contain 0, then the corresponding p-value is less than 0.05 and vice-versa. Thus, we can conclude from the above 95% CIs that the treatments CE and ET are not statistically significantly different (95% CI of Δ contains 0), while ET is statistically significantly different (and better) than the placebo (95% CI of Δ does not contain 0).

It is important to realize that a 95% CI bears more information than a p-value. Indeed, both methods allow the evaluation of the clinical relevance of the obtained difference in means and its statistical significance. However, the CI has a much easier interpretation and allows better comparability of results across different clinical trials. The 95% CI is also the preferred tool of statistical inference in meta-analyses.

Some Final Remarks

It is surprising to see that, almost a century following its introduction, the p-value is still poorly understood by many clinicians. Thus, it must be recognized that the p-value is a highly abstract concept not well understood. The CI, on the other hand, is easier to understand and gives, as well, a much better insight into the observed clinical results. Therefore, it is reassuring to see that an increasing number of journals require reporting of CIs.

Disclosure Statement

The author has no financial or proprietary interest in the subject matter or materials discussed, including, but not limited to, employment, consultancies, stock ownership, honoraria, and paid expert testimony.

References

1. Bingham CO III, Sebba AI, Rubin BR, et al. Efficacy and safety of etoricoxib 30 mg and celecoxib 200 mg in the treatment of osteoarthritis in two identically designed, randomized, placebo-controlled, non-inferiority studies. *Rheumatology*. 2007;46:496-507.
2. Lesaffre E. Superiority, equivalence and non-inferiority trials. *Bull NYU Hosp Jt Dis*. 2008;66(2):150-4.
3. Moyé LA. *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*, New York: Springer-Verlag, 2003.
4. Proschan MA, Lan KKG, Wittes JT. *Statistical Monitoring of Clinical Trials: A Unified Approach*, New York: Springer-Verlag, 2006.