REVIEW ARTICLE

Confidence Interval or P-Value?

Part 4 of a Series on Evaluation of Scientific Publications

Jean-Baptist du Prel, Gerhard Hommel, Bernd Röhrig, Maria Blettner

SUMMARY

Background: An understanding of p-values and confidence intervals is necessary for the evaluation of scientific articles. This article will inform the reader of the meaning and interpretation of these two statistical concepts.

Methods: The uses of these two statistical concepts and the differences between them are discussed on the basis of a selective literature search concerning the methods employed in scientific articles.

Results/Conclusions: P-values in scientific studies are used to determine whether a null hypothesis formulated before the performance of the study is to be accepted or rejected. In exploratory studies, p-values enable the recognition of any statistically noteworthy findings. Confidence intervals provide information about a range in which the true value lies with a certain degree of probability, as well as about the direction and strength of the demonstrated effect. This enables conclusions to be drawn about the statistical plausibility and clinical relevance of the study findings. It is often useful for both statistical measures to be reported in scientific articles, because they provide complementary types of information.

Dtsch Arztebl Int 2009; 106(19): 335–9
DOI: 10.3238/arztebl.2009.0335

Key words: publications, clinical research, p-value, statistics, confidence interval

Johannes Gutenberg-Universität Mainz: Zentrum für Kinder- und Jugendmedizin, Zentrum Präventive Pädiatrie: Dr. med. du Prel, MPH

Johannes Gutenberg-Universität Mainz: Institut für Medizinische Biometrie, Epidemiologie und Informatik: Prof. Dr. rer. nat. Hommel, Dr. rer. nat. Röhrig, Prof.Dr. rer. nat. Blettner

People who read scientific articles must be familiar with the interpretation of p-values and confidence intervals when assessing the statistical findings. Some will have asked themselves why a p-value is given as a measure of statistical probability in certain studies, while other studies give a confidence interval and still others give both. The authors explain the two parameters on the basis of a selective literature search and describe when p-values or confidence intervals should be given. The two statistical concepts will then be compared and evaluated.

What is a p-value?

In confirmatory (evidential) studies, null hypotheses are formulated, which are then rejected or retained with the help of statistical tests. The p-value is a probability, which is the result of such a statistical test. This probability reflects the measure of evidence against the null hypothesis. Small p-values correspond to strong evidence. If the p-value is below a predefined limit, the results are designated as "statistically significant" (1). The phrase "statistically striking results" is also used in exploratory studies.

If it is to be shown that a new drug is better than an old one, the first step is to show that the two drugs are not equivalent. Thus, the hypothesis of equality is to be rejected. The null hypothesis (H_0) to be rejected is then formulated in this case as follows: "There is no difference between the two treatments with respect to their effect." For example, there might be no difference between two antihypertensives with respect to their ability to reduce blood pressure. The alternative hypothesis (H_1) then states that there is a difference between the two treatments. This can either be formulated as a two-tailed hypothesis (any difference) or as a one-tailed hypothesis (positive or negative effect). In this case, the expression "one-tailed" means that the direction of the expected effect is laid down when the alternative hypothesis is formulated. For example, if there is clear preliminary evidence that an antihypertensive has on average a stronger hypertensive effect than the comparator drug, the alternative hypothesis can be formulated as follows: "The difference between the mean hypotensive activity of antihypertensive 1 and the mean hypotensive activity of antihypertensive 2 is positive." However, as this requires plausible assumptions about the direction of the effect, the two-tailed hypothesis is often formulated.

For example, the data from a randomized clinical study are to be used to estimate the effect strength relevant to the question to be answered. This could, for example, be the difference between the mean decrease in blood pressure with a new and with an old antihypertensive. On this basis, the null hypothesis formulated in advance is tested with the help of a significance test. The p-value gives the probability of obtaining the present test result—or an even more extreme one—if the null hypothesis is correct. A small p-value signifies that the probability is small that the difference can purely be assigned to chance. In our example, the observed difference in mean systolic pressure might not be due to a real difference in the hypotensive activity of the two antihypertensives, but might be due to chance. However, if the p-value is < 0.05, the chance that this is the case is under 5%. To permit a decision between the null hypothesis and the alternative hypothesis, significance limits are often specified in advance, at a level of significance α. The level of significance of 0.05 (or 5%) is often chosen. If the p-value is less than this limit, the result is significant and it is agreed that the null hypothesis should be rejected and the alternative hypothesis—that there is a difference—is accepted. The specification of the level of significance also fixes the probability that the null hypothesis is wrongly rejected.

P-values alone do not permit any direct statement about the direction or size of a difference or of a relative risk between different groups (1). However, this would be particularly useful when the results are not significant (2). For this purpose, confidence limits contain more information. Aside from p-values, at least a measure of the effect strength must be reported—for example, the difference between the mean decreases in blood pressure in the two treatment groups (3). In the final analysis, the definition of a significance limit is arbitrary and p-values can be given even without a significance limit being selected. The smaller the p-value, the less plausible is the null hypothesis that there is no difference between the treatment groups.

Confidence limits—from the dichotomous test decision to the effect range estimate

The confidence interval is a range of values calculated by statistical methods which includes the desired true parameter (for example, the arithmetic mean, the difference between two means, the odds ratio etc.) with a probability defined in advance (coverage probability, confidence probability, or confidence level). The confidence level of 95% is usually selected. This means that the confidence interval covers the true value in 95 of 100 studies performed (4, 5). The advantage of confidence limits in comparison with p-values is that they reflect the results at the level of data measurement (6). For instance, the lower and upper limits of the mean systolic blood pressure difference between the two treatment groups are given in mm Hg in our example.

The size of the confidence interval depends on the sample size and the standard deviation of the study groups (5). If the sample size is large, this leads to "more confidence" and a narrower confidence interval. If the confidence interval is wide, this may mean that the sample is small. If the dispersion is high, the conclusion is less certain and the confidence interval becomes wider. Finally, the size of the confidence interval is influenced by the selected level of confidence. A 99% confidence interval is wider than a 95% confidence interval. In general, with a higher probability to cover the true value the confidence interval becomes wider.

In contrast to p-values, confidence intervals indicate the direction of the effect studied. Conclusions about statistical significance are possible with the help of the confidence interval. If the confidence interval does not include the value of zero effect, it can be assumed that there is a statistically significant result. In the example of the difference of the mean systolic blood pressure between the two treatment groups, the question is whether the value 0 mm Hg is within the 95% confidence interval (= not significant) or outside it (= significant). The situation is equivalent with the relative risk; if the confidence interval contains the relative risk of 1.00, the result is not significant. It would then have to be examined whether the confidence interval for the relative risk is completely under 1.00 (= protective effect) or completely above it (= increase in risk).

Figure 1 shows the difference for the example of the mean systolic blood pressure difference between two groups. The confidence interval for the mean blood pressure difference is narrow with small variation within the sample (= low dispersion) *(figure 1b)*, low confidence level *(figure 1d)* and large sample size *(figure 1f)*. In this example, there is no significant difference between the mean systolic blood pressures in the groups if the dispersion is high *(figure 1c)*, the confidence level is high *(figure 1e)* or the sample size is small *(figure 1g)*, as the value zero is then contained in the confidence interval.

Although point estimates, such as the arithmetic mean, the difference between two means or the odds ratio, provide the best approximation to the true value, they do not provide any information about how exact they are. This is achieved by confidence intervals. It is of course impossible to make any precise statement about the size of the difference between the estimated parameters for the sample and the true value for the population, as the true value is unknown. However, one would like to have some confidence that the point estimate is in the vicinity of the true value (7). Confidence intervals can be used to describe the probability that the true value is within a given range.

If a confidence interval is given, several conclusions can be made. Firstly, values below the lower limit or above the upper limit are not excluded, but are improbable. With the confidence limit of 95%, each of these probabilities is only 2.5%. Values within the confidence limits, but near to the limits, are mostly less probable than values near the point estimate, which in our example with the two antihypertensives is the difference in the mean values of the reduction in blood pressure in the two treatment groups in mm Hg. Whatever the size of the confidence interval, the point estimate based on the

sample is the best approximation to the true value for the population. Values in the vicinity of the point estimate are mostly plausible values. This is particularly the case if it can be assumed that the values are normally distributed.

A frequent procedure is to check whether confidence intervals include a certain limit or not and, if they do not, to regard the findings as being significant. It is however a better approach to exploit the additional information in confidence intervals. Particularly with so-called close results, the possibility should be considered that the result might have been significant with a larger sample.

Important international journals of medical science, such as the Lancet and the British Medical Journal, as well as the International Committee of Medical Journal Editors (ICMJE), recommend the use of confidence intervals (6). In particular, confidence intervals are of great help in interpreting the results of randomized clinical studies and meta-analyses. Thus the use of confidence intervals is expressly demanded in international agreements and in the CONSORT statement (8) for reporting randomized clinical studies and in the QUORUM statement (9) for reporting systematic reviews.

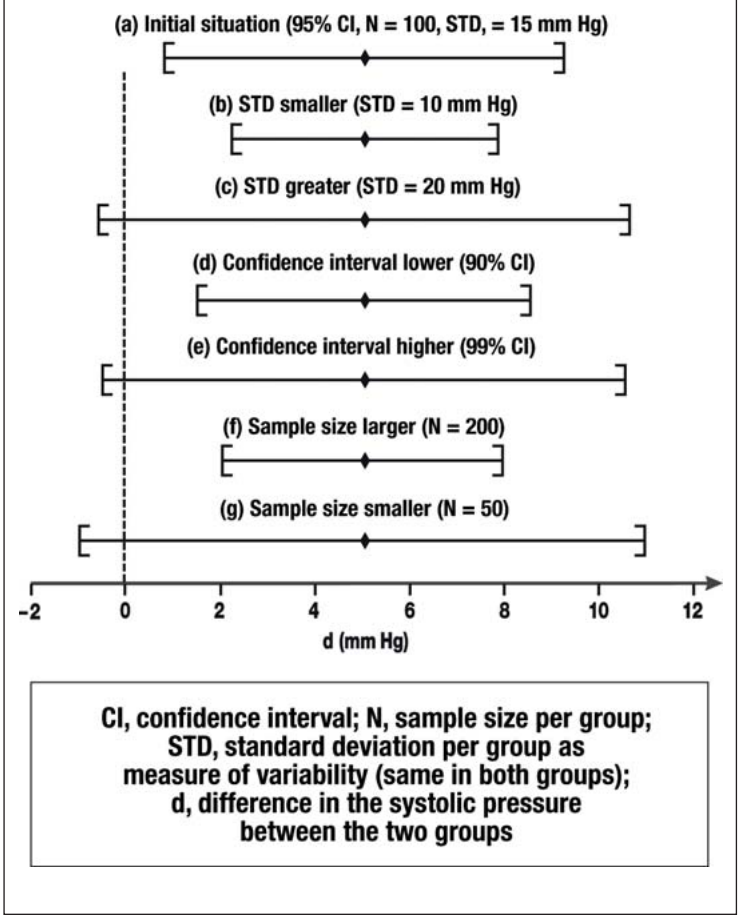
Statistical significance versus clinical relevance

A clear distinction must be made between statistical significance and clinical relevance (or clinical significance). Aside from the effect strength, p-values incorporate the case numbers and the variability of the sample data. Even if the limit for statistical significance is laid down in advance, the reader must still judge the clinical relevance of statistically significant differences for himself. The same numerical value for the difference may be "statistically significant" if a large sample is taken and "not significant" if the sample is smaller. On the other hand, results of high clinical relevance are not automatically unimportant if there is no statistical significance. The cause may be that the sample is too small or that the dispersion in the samples is too great—for example, if the patient group is highly heterogenous. For this reason, a decision for significance or lack of significance on the basis of the p-value alone may be simplistic.

This can be illustrated using the example of systolic blood pressure. *Figure 2* specifies a relevance limit r. A systolic blood pressure difference of at least 4 mm Hg between the two groups is then defined as clinically relevant. If the blood pressure difference is neither statistically significant nor clinically relevant *(figure 2a)* or statistically significant and clinically relevant *(figure 2b)*, interpretation is easy. However, statistically significant differences in blood pressure may lie under the limit for clinical relevance and are then of no clinical importance *(figure 2c)*. On the other hand, there may be real and clinically important differences in systolic blood pressure between the treatment groups, even though statistical significance has not been achieved *(figure 2d)*.

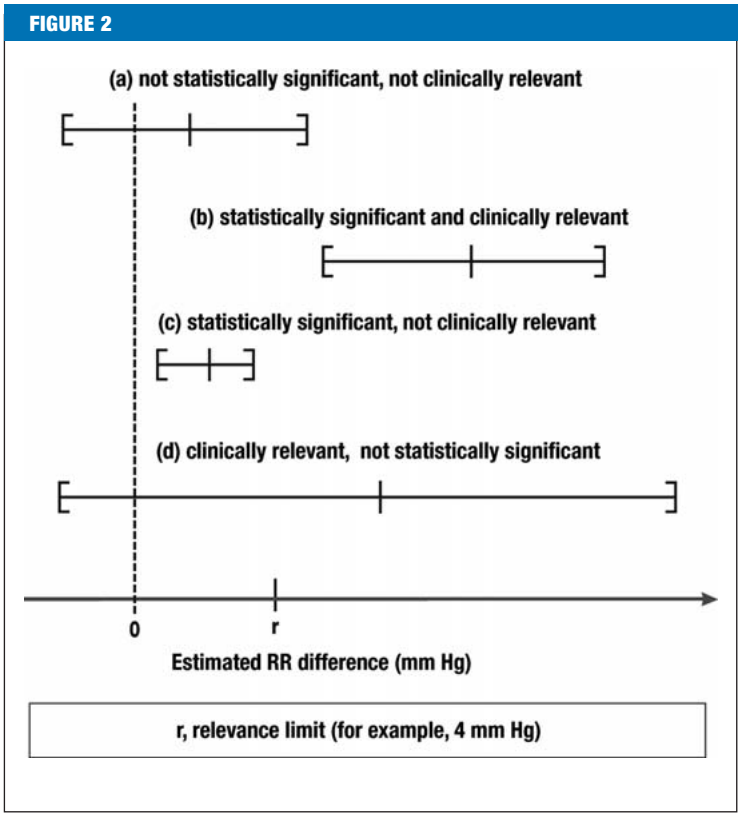
Unfortunately, statistical significance is often thought to be equivalent to clinical relevance. Many research workers, readers, and journals ignore findings which are potentially clinically useful only because they are not statistically significant (4). At this point, we can criticize the practice of some scientific journals of preferably publishing significant results. A study has shown that this is mainly the case in high–impact factor journals (10). This can distort the facts ("publication bias"). Moreover, it can often be seen that a non-significant difference is interpreted as meaning that there is no difference (for example, between two treatment groups). A p-value of >0.05 only signifies that the evidence is not adequate to reject the null hypothesis—for example, that there is no difference between two alternative treatments. This does not imply that the two treatments are equivalent. The quantitative compilation of comparable studies in the form of systematic reviews or meta-analyses can then help to identify differences which had not been recognized because the number of cases in individual studies had been too low. A special article in this series is devoted to this subject.

FIGURE 1



Using the example of the difference in the mean systolic blood pressure between two groups, it is examined how the size of the confidence interval (a) can be modified by changes in dispersion (b, c), confidence interval (d, e), and sample size (f, g). The difference between the mean systolic blood pressure in group 1 (150 mm Hg) and in group 2 (145 mm Hg) was 5 mmHg. Example modified from (6)

FIGURE 2



Statistical significance and clinical relevance

P-values versus confidence intervals—What are the differences?

The essential differences between p-values and confidence intervals are as follows:

- The advantage of confidence intervals in comparison to giving p-values after hypothesis testing is that the result is given directly at the level of data measurement. Confidence intervals provide information about statistical significance, as well as the direction and strength of the effect (11). This also allows a decision about the clinical relevance of the results. If the error probability is given in advance, the size of the confidence interval depends on the data variability and the case number in the sample examined (12).
- P-values are clearer than confidence intervals. It can be judged whether a value is greater or less than a previously specified limit. This allows a rapid decision as to whether a value is statistically significant or not. However, this type of "diagnosis on sight" can be misleading, as it can lead to clinical decisions solely based on statistics.
- Hypothesis testing using a p-value is a binary (yes-or-no) decision. The reduction of statistical inference (inductive inference from a single sample to the total population) to this level may be simplistic. The simple distinction between "significant" and "non-significant" in isolation is not very reliable. For example, there is little difference between the evidence for p-values of 0.04 and of 0.06. Nevertheless, binary decisions based on these minor differences lead to converse decisions (1, 13). For this reason, p-values must always be given completely (suggestion: always to three decimal places) (14).
- When a point estimate is used (for example, difference in means, relative risk), an attempt is made to draw conclusions about the situation in the target population on the basis of only a single value for the sample. Even though this figure is the best possible approximation to the true value, it is not very probable that the values are exactly the same. In contrast, confidence intervals provide a range of possible plausible values for the target population, as well as the probability with which this range covers the real value.
- In contrast to confidence intervals, p-values give the difference from a previously specified statistical level α (15). This facilitates the evaluation of a "close" result.
- Statistical significance must be distinguished from medical relevance or biological importance. If the sample size is large enough, even very small differences may be statistically significant (16, 17). On the other hand, even large differences may lead to non-significant results if the sample is too small (12). However, the investigator should be more interested in the size of the difference in therapeutic effect between two treatment groups in clinical studies, as this is what is important for successful treatment, rather than whether the result is statistically significant or not (18).

Conclusion

Taken in isolation, p-values provide a measure of the statistical plausibility of a result. With a defined level of significance, p-values allow a decision about the rejection or maintenance of a previously formulated null hypothesis in confirmatory studies. Only very restricted statements about effect strength are possible on the basis of p-values. Confidence intervals provide an adequately plausible range for the true value related to the measurement of the point estimate. Statements are possible on the direction of the effects, as well as its strength and the presence of a statistically significant result. In conclusion, it should be clearly stated that p-values and confidence intervals are not contradictory statistical concepts. If the size of the sample and the dispersion or a point estimate are known, confidence intervals can be calculated from p-values, and conversely. The two statistical concepts are complementary.

Conflict of interest statement
The authors declare that there is no conflict of interest as defined by the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 23 July 2008, revised version accepted on 21 August 2008.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Bland M, Peacock J: Interpreting statistics with confidence. The Obstetrician and Gynaecologist 2002; 4: 176–80.
2. Houle TT: Importance of effect sizes for the accumulation of knowledge. Anesthesiology 2007; 106: 415–7.
3. Faller, H: Signifikanz, Effektstärke und Konfidenzintervall. Rehabilitation 2004; 43: 174–8.
4. Greenfield ML, Kuhn JE, Wojtys EM: A statistics primer. Confidence intervals. AmJ Sports Med 1998; 26: 145-9. No abstract available. Erratum in: Am J Sports Med 1999; 27: 544.
5. Bender R, Lange St: Was ist ein Konfidenzintervall? Dtsch Med Wschr 2001; 126: 41.
6. Altman DG: Confidence intervals in practice. In: Altman DG, Machin D, Bryant TN, Gardner MJ. BMJ Books 2002; 6–9.
7. Weiss C: Intervallschätzungen. Die Bedeutung eines Konfidenzintervalls: In: Weiß C: Basiswissen Medizinische Statistik. Springer Verlag 1999; 191–2.
8. Moher D, Schulz KF, Altman DG für die CONSORT Gruppe: Das COSORT Statement: Überarbeitete Empfehlungen zur Qualitätsverbesserung von Reports randomisierter Studien im Parallel-Design. Dtsch Med Wschr 2004; 129: 16-20.
9. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF: Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999; 354: 1896–900.
10. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR: Publication bias in clinical research. Lancet 1991; 337: 867–72.
11. Shakespeare TP, Gebski VJ, Veness MJ, Simes J: Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and riskbenefit contours. Lancet 2001; 357: 1349–53. Review.
12. Gardner MJ, Altman DG: Confidence intervals rather than P-values: estimation rather than hypothesis testing. Br Med J 1986; 292: 746–50.
13. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S: Basic statistics for clinicians: 1. hypothesis testing. CMAJ 1995; 152: 27–32. Review.
14. ICH 9: Statisticlal Principles for Clinical Trials. London UK: International Conference on Harmonization 1998; Adopted by CPMP July 1998 (CPMP/ICH/363/96)
15. Feinstein AR: P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998; 51: 355–60.
16. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S: Basic statistics for clinicians: 2. interpreting study results: confidence intervals. CMAJ 1995; 152: 169–73.
17. Sim J, Reid N: Statistical inference by confidence intervals: issues of interpretation and utilization. Phys Ther 1999; 79: 186–95.
18. Gardner MJ, Altman DG. Confidence intervals rather than P values. In: Altman DG, Machin D, Bryant TN, Gardner MJ: Statistics with confidence. Confidence intervals and statistical guidelines. Second Edition. BMJ Books 2002; 15–27.

Corresponding author
Dr. med. Jean-Baptist du Prel, MPH
Zentrum für Kinder- und Jugendmedizin
Zentrum Präventive Pädiatrie Mainz
Langenbeckstr. 1
55101 Mainz, Germany
duprel@zpp.klinik.uni-mainz.de