ELSEVIER

Netherlands Journal of Medicine 49 (1996) 177–182

The Netherlands
JOURNAL OF
MEDICINE

Seminar

Confidence intervals and *P*-values [1]

Th. Stijnen *

Department of Epidemiology and Biostatistics and Consultation Centre for Clinical Research, Erasmus University and University Hospital, PO Box 1738, 3000 DR Rotterdam, Netherlands

Received 22 May 1996; accepted 24 May 1996

Keywords: Confidence interval; *P*-value; Standard error

1. Introduction

Over the past 20 or 30 years the use of statistics in medical research has grown enormously. Important concepts in medical statistics are *P*-value and confidence interval. The use of hypothesis testing and *P*-values still predominates in the medical literature where confidence intervals are often more informative. In the last decade editorial boards of several medical journals have promoted the use of confidence intervals instead of *P*-values [1–6]. In this Seminar a short introduction to the concepts *P*-value and confidence interval is given, and the advantage of the latter is discussed. Since the standard error plays an important role in this context, this concept is introduced first.

2. Standard error

The concept of standard error concerns the precision of an estimate of an unknown population parameter. The smaller the standard error, the more precise the estimate. An implicit assumption in almost all statistical analyses is that the data may be considered as a random sample from a certain population. Based on the data in the sample, an estimate is computed of an unknown characteristic of interest of the population (often a population mean or a percentage, in general a population parameter). For example, suppose that in a random sample of size $n = 200$ from the population of Dutch women above 65 years the body weight (x) is measured. The unknown mean in the population is called μ. An estimate for it is of course the sample mean, $\tilde{x}$. The standard error of $\tilde{x}$ is a measure for the precision of $\tilde{x}$ as an estimate of μ. Since μ is a mean, this standard error is called the 'standard error of the mean' (SEM). The SEM is computed as $s/\sqrt{n}$, where s is the sample standard deviation. The variability of a variable in the population, in our example body weight, is usually described by the standard deviation σ. An estimate for this is the sample standard deviation s. Also the precision of s as an estimate for σ could be characterized by a standard error, although this is rarely done in the medical literature.

Note that the meaning of the word *parameter* here is different from the terminology often used in clinical practice. Physicians often use the term 'parameter' where a variable is meant. In the statisti-

* Tel.: (+31-10) 4087390; fax: (+31-10) 4365933.

[1] Seminars in Clinical Epidemiology and Decision Analysis. Series Editor: D.E. Grobbee, MD, PhD, Department of Epidemiology, Erasmus university, and Consultation Centre for Clinical Research, Erasmus University and University Hospital, Rotterdam, Netherlands.

0300-2977/96/$15.00 Copyright © 1996 Elsevier Science B.V. All rights reserved.
PII S0300-2977(96)00041-1

cal terminology a parameter cannot vary between persons, but is a fixed, usually unknown, number that refers to a certain characteristic of a population.

The standard error of an estimate of a percentage is determined as follows. As an example we consider the situation where a sample of size $n = 200$ is available from the population of Dutch women above the age of 65, and one is interested in the (unknown) prevalence (π) of osteoporosis. Suppose 20 women in the sample are observed to have osteoporosis. Then, of course, the best estimate for π is the sample prevalence $p = 20/200 = 10\%$. The corresponding standard error is given by $SE(p) = \sqrt{(p(100 - p)/n)} = \sqrt{(10 \times 90/200)} = 2.1\%$.

Often the question of what to report is raised in publications—the standard error of the mean or the standard deviation? The answer is straightforward, if one realises the different aims of the two concepts: the standard deviation describes variability in the population, while the standard error characterizes how precise the estimate of the mean is. However, it is not surprising that the two concepts are often interchanged, because the standard error can be computed from the standard deviation and the other way around (via $SEM = s/\sqrt{n}$). Moreover, the standard error is in fact also a standard deviation, not of the variable itself, but of the distribution of the sample mean of that variable. Note that the sample mean has a distribution: if sampling is repeated many times, then, due to sampling variability, the sample mean will vary from sample to sample. So there exists an (imaginary) 'population' of parameter estimates. The precision of the sample mean as an estimate for the population mean is determined by the extent of variability in this 'population': the smaller this variability, the better the precision of the estimate. This variability in turn can be characterized by its standard deviation. In fact, the standard error is just defined as the estimate of the standard deviation of the distribution of the sample mean. Therefore sometimes the terminology 'standard deviation of $\bar{x}$' is used instead of standard error of $\bar{x}$.

It is explained above that the standard error can be interpreted as the standard deviation of the sample mean. A more useful and more concrete interpretation is directly related to the concept of the confidence interval, which is introduced in the next section.

3. Confidence intervals

The aim of a confidence interval is to give a range of possible values for the unknown population parameter of interest, say μ, which are reasonably consistent with the data observed in the sample. Values outside this interval are implausible values for μ. Values inside the interval are more likely, where values in the centre of the interval are more likely than peripheral values. In the sequel the concept of the confidence interval is made more precise and it will be seen that the standard error plays a major role. As an example the situation is considered where one is interested in the estimation of a population mean, based on a sample from the population.

It was explained above that the standard error is defined as the estimate of the standard deviation of the distribution of the sample mean. A well-known theorem from probability theory, the Central Limit Theorem, says that the distribution of a sample mean, irrespective of the distribution of the variable in the population, can be well approximated by what is called a 'normal' or 'Gaussian' distribution. The larger the sample size, the better the resemblance to a normal distribution. A normal distribution is characterized by two parameters—its mean μ and its standard deviation σ (see Fig. 1 for an illustration)—and has the property that the probability of a value between $\mu - k\sigma$ and $\mu + k\sigma$ only depends on k, not on μ or σ. For example, the probability of a value

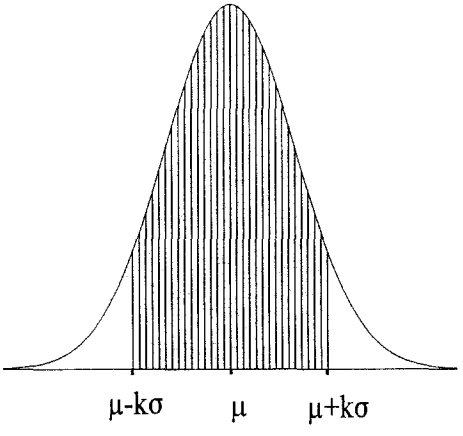

Fig. 1. The normal or Gaussian distribution with mean μ and standard deviation σ. The probability of a value between $\mu \pm k\sigma$ only depends on k (see Table 1).

Table 1
The normal distribution: probability of a value between $\mu \pm k\sigma$ for some selected values of k

k	Probability between $\mu \pm k\sigma$
0.5	0.38
1	0.68
1.5	0.87
1.64	0.90
1.96	0.95
2.33	0.98
2.58	0.99
3.29	0.999

A more extensive table can be found in Refs. [8–10].

between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 95%. In Table 1 this probability is given for a number of possible choices of k. So the distribution of $\tilde{x}$ is approximately normal with mean μ and standard deviation equal to the SEM. The probability that $\tilde{x}$ takes a value between $\mu \pm 1.96 \times$ SEM, therefore, is about 95%. Note that $\tilde{x}$ between $\mu \pm 1.96 \times$ SEM implies that μ lies between $\tilde{x} \pm 1.96 \times$ SEM. Thus the probability is approximately 95% that $\tilde{x} - 1.96 \times \text{SEM} < \mu < +1.96 \times \text{SEM}$. This interval is called an (approximate) '95% confidence interval' for μ. Roughly speaking, this means that there is an about 95% chance that the unknown parameter value μ lies in the interval. Strictly speaking, this holds before the experiment is done. Once the experiment has been done, μ is or is not in the interval, and we can no longer talk about a probability. This is why one talks about 'confidence'. More exactly formulated, a 95% confidence interval means that if the sampling were repeated many times, in the long run 95% of the confidence intervals would contain the unknown value of μ.

Example 1. Suppose a sample of size $n = 50$ is drawn from the population of Dutch women above 65 years and body weight is measured. The mean of body weight in the sample is $\tilde{x} = 67.3$ kg and the standard deviation is $s = 5.2$ kg. The standard error then is $\text{SEM} = s/\sqrt{n} = 5.2/\sqrt{50} = 0.73$. The approximate 95% confidence interval for the unknown mean body weight in the population is: $\tilde{x} \pm 1.96 \times \text{SEM} = 67.3 \pm 1.96 \times 0.73 = 67.3 \pm 1.4 = 65.9$ kg to 68.7 kg. The conclusion is that based on this study one can be 95% 'confident' that the unknown μ lies between these bounds.

Example 2.

Suppose one is interested in the prevalence π of osteoporosis in the population of Dutch women above 65 years. In a sample of size $n = 200$ women 24 cases of osteoporosis are observed. Then the estimated prevalence is $p = 24/200 = 12\%$. The corresponding standard error is $\text{SE}(p) = \sqrt{(12 \times (100 - 12)/200)} = 2.3\%$, and the approximate 95% confidence interval for π is: $p \pm 1.96 \times \text{SE}(p) = 12 \pm 1.96 \times 2.3 = 12 \pm 4.5 = 7.5\%$ to 16.5%.

In most cases the level of confidence is chosen to be 95%. There is however no special reason for this choice. Another level of confidence may be obtained by choosing another value for k (Table 1). For instance, a 99% confidence interval is obtained by taking $k = 2.58$, and a 90% confidence interval by $k = 1.64$.The interval $\tilde{x} \pm \text{SEM}$ gives a 68% confidence interval. The larger the chosen level of confidence, the wider the confidence interval will be. The width of a confidence interval depends on the level of confidence and the magnitude of the standard error. The width of the confidence interval depends on the sample size n through the standard error of the mean $s/\sqrt{n}$. If n increases, the width decreases proportionally to $1/\sqrt{n}$: i.e., to halve the length of the interval, the sample size must be increased by a factor of 4.

So far, *approximate* confidence intervals have been presented. On the assumption that the distribution of x in the population is Gaussian, it is possible to construct *exact* confidence intervals for μ. The difference from the approximate one is that Student's t-distribution is used to determine k instead of the normal distribution. It is beyond the scope of this Seminar to go into the details of this. For sample sizes above $n = 30$, the difference between exact and approximate is negligible. Also for a percentage exact confidence intervals can be computed. However, the approximate approach described above is usually sufficiently accurate.

4. *P*-values

In the medical literature very often, appropriately or inappropriately, P-values are reported. A P-value always corresponds to a hypothesis (the 'null hy-

pothesis') concerning the unknown value of a population parameter. Simply said, a P-value can be interpreted as a measure of evidence in the data *against* the null hypothesis. The smaller the P-value, the less plausible the null hypothesis is given the observed data. Therefore, if a P-value is mentioned, the reader first of all should ask what the corresponding null hypothesis is, otherwise one cannot interpret the P-value. Sometimes the P-value is even irrelevant because the reader is not interested in the corresponding hypothesis.

In the sequel the construction, definition and interpretation of the P-value are discussed. This will be done with the following example, based on Example 2. Suppose it is known that the prevalence of osteoporosis in the corresponding Belgian population is 6%. In the study of the example a prevalence of 12% was found with standard error 2.3%. On the basis of the data observed in the study, is it possible to claim that the true prevalence in the Netherlands is higher than in Belgium? In statistical terminology the null hypothesis (H_0) becomes $\pi = 6\%$ and one wishes to quantify the evidence in the data against this null hypothesis. In order to do this, a 'test statistic' is used. In general, this test statistic has the following form:

$$Z = \frac{\text{Parameter estimate} - \text{Parameter value if } H_0 \text{ true}}{\text{Standard error}}$$

In this case this is: $Z = (12 - 6)/2.3 = 2.61$. It is clear that if H_0 is not true, then Z will tend to have a large outcome (in absolute value), while if H_0 is true the test statistic Z will tend to have a small outcome (in absolute value). So a large observed value of Z indicates that the null hypothesis might not be true, while a small outcome of Z supports the null hypothesis. How extreme is the observed outcome $Z = 2.61$? In order to judge that, we consider the distribution of Z if the null hypothesis is true. From probability theory it is known that this distribution is well approximated by a standard normal distribution: i.e., the Gaussian distribution with $\mu = 0$ and $\sigma = 1$. In Fig. 2 the graph of this distribution is given. As a measure of how extreme the observed value 2.61 of Z is under H_0, the shaded area in the figure is taken. This is the (two-sided [2]) P-value. So the P-value is the probability on the null hypothesis that the test statistic takes a value at least as extreme as the value actually observed. From Table 2 it can be seen that the P-value corresponding to $Z = 2.61$ is smaller than 0.01. This means that something has happened that was very unlikely under the null hypothesis, something that had a probability of less than 1%. This is why one argues that the null hypothesis is not true if the P-value is small enough. As a threshold value usually $P = 0.05$ is chosen. As soon as the P-value is smaller than this threshold, called the 'significance level', one claims that H_0 is

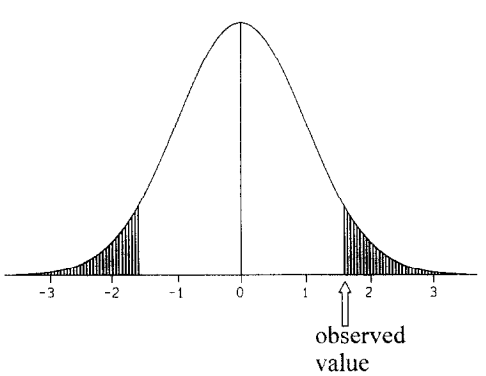

Fig. 2. The approximate distribution of the test statistic Z if the null hypothesis is true: the standard normal distribution. The shaded area is the P-value.

Table 2
Some selected possible outcomes of the test statistic Z and the corresponding P-value

Z	P-value
0.00	1
0.67	0.50
1.28	0.20
1.64	0.10
1.96	0.05
2.33	0.02
2.58	0.01
3.29	0.001

See Refs. [8–10] for a more extensive table.

[2] The P-value is called two-sided because large negative as well as large positive values are considered to be evidence that the null hypothesis is not true. If only large negative or only large positive values indicate that the null hypothesis is not true, the one-sided P-value might be taken. The use of one-sided P-values is very rare in medical articles. See Ref. [7] for a recent discussion of this topic.

not true. Of course, this might be incorrect, but the chance that this happens while H_0 in fact is true, is small (i.e., at most 5%). If one considers the risk of rejecting H_0 as too high, one can of course choose a smaller significance level (e.g., 0.01). If the P-value is smaller than the chosen significance level, one says that the test result is 'statistically significant'. In our example it might be claimed that the true prevalence of osteoporosis in the Dutch population is higher than 6%, and one says that the observed prevalence, $p = 12\%$, is statistically significantly higher than 6%.

In general, there is a close relationship between hypothesis testing and a confidence interval: the 95% confidence interval exists of all values of π_0 for which the P-value of the test for H_0: $\pi = \pi_0$ is larger than 5%. In other words, the parameter values outside the confidence interval are statistically significant, while the parameter values inside the interval are non-significant.

Example 3. In order to investigate the difference in efficacy between two treatments A and B, 100 patients were treated in a double-blind randomized clinical trial with treatment A and 100 patients with treatment B. The data are summarized in the table below.

	Treatment		
Outcome	A	B	Total
Cured	30	40	65
Not cured	70	60	135
Total	100	100	200

The question is whether on the basis of these data it is justified to claim that treatment B is more effective than A. The statistical null hypothesis then is H_0: $\pi_B - \pi_A = 0$, where π_A and π_B are the cure probabilities corresponding to treatment A and B, respectively. The estimated cure probability is 30% for A and 40% for B. The corresponding standard errors are respectively 4.6% and 4.9%. The estimate of the difference between the cure probabilities is $40\% - 30\% = 10\%$. To determine the standard error of this estimate, the statistical rule is used that says that the standard error of a difference is equal to the square root of the sum of the squared standard errors. So in this case the standard error is $\sqrt{(4.6^2 + 4.9^2)} = 6.7\%$ and the test statistic becomes: $Z = 10/6.7 = 1.49$. From Table 2 it follows that the P-value is somewhere between 10% and 20%. (Roughly this means that the probability of an observed difference between the cure probabilities larger than or equal to 10% occurs with a probability larger than 10%, assuming that the two cure probabilities in fact are equal). Therefore the two cure probabilities are not significantly different, and, on the basis of these data, it is not justified to claim that B has a higher cure rate than A.

5. Discussion

Although the P-value has its merit as a measure of evidence against a certain null hypothesis, it has its limitations. This is certainly the case if it is given in simple statements as '$P < 0.05$', '$P > 0.05$' or '$P = \text{NS}$', as is frequently done. This reduces the result of a study to a simple 'yes' or 'no' answer, which is clearly an oversimplification and not very informative. For example, the evidence against the null hypothesis is much stronger if $P = 0.01$ than if $P = 0.05$, although both results are significant at level 0.05. Therefore, it is always advised to report the exact P-values. The main limitation of the P-value is that it does not explicitly refer to the magnitude of the effect. Consider as an example a clinical trial comparing an experimental treatment with placebo treatment. If the estimated treatment effect is statistically significant, then the effect is not necessarily also clinically relevant. For example, the estimated treatment effect might be small but is nevertheless statistically significant, due to a large sample size or a small variability in outcome measure. On the other hand, a statistically non-significant treatment effect does not necessarily imply that the treatment effect is zero. The true treatment effect might be of clinically relevant size, but statistical significance is not reached due to a too small number of patients or a high variability in the outcome variable. This is illustrated by the example in the previous section. The cure rate under treatment B is 10% higher than for treatment A, but the difference is not statistically significant. However, a true difference in cure rate of 10% would probably have been very

clinically relevant, and it is obvious that even larger differences cannot be excluded.

The confidence interval does not have these limitations. By giving a range of values, on the basis of the study data, in which the true value for the treatment effect may lie, it is more informative and easier to interpret. This is again illustrated by the example in Section 4. The 95% confidence interval for the true difference in cure rates between treatment A and B is $10 \pm 1.96 \times 6.7 = 10 \pm 13.1\% = (-3.1\%, 23.1\%)$. This confidence interval gives an adequate description of the uncertainty about the true difference. A higher cure rate of 20% or more in favour of treatment B is not excluded by the interval. So it is immediately clear from the interval that the result of the trial is inconclusive.

In this Seminar it was only possible to touch upon the most important biostatistical concepts used in medical articles. For a more detailed discussion and a broader introduction to biostatistical methods in clinical research, the reader is referred to Refs. [8–10].

References

[1] Langman MJS. Towards estimation and confidence intervals. Br Med J 1986;292:716.
[2] Gardner MJ, Altman DG. Confidence intervals rather than *P*-values: estimation rather than hypothesis testing. Br Med J 1986;292:746–750.
[3] Anonymous. Report with confidence [Editorial]. Lancet 1987;i:488.
[4] Bulpitt CJ. Confidence intervals. Lancet 1987;i:494–497.
[5] Rothman KJ, Yankauer A. Confidence intervals vs. significance tests: quantitative interpretation (Editors note). Am J Public Health 1986;76:587–588.
[6] International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Br Med J 1988;296:401–405.
[7] Walker MW. Low power and striking results – a surprise but not a paradox. N Engl J Med 1995;332:1091–1092.
[8] Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
[9] Campbell MJ, Machin D. Medical statistics: a commonsense approach. Chichester: John Wiley and Sons, 1990.
[10] van Houwelingen JC, Stijnen Th, van Strik R. Inleiding in de medische statistiek. Utrecht: Wetenschappelijke uitgeverij Bunge, 1993.