Statistics in Medicine

Confidence intervals rather than P values: estimation rather than hypothesis testing

MARTIN J GARDNER, DOUGLAS G ALTMAN

Abstract

Overemphasis on hypothesis testing—and the use of P values* to dichotomise significant or non-significant results—has detracted from more useful approaches to interpreting study results, such as estimation and confidence intervals. In medical studies investigators are usually interested in determining the size of difference of a measured outcome between groups, rather than a simple indication of whether or not it is statistically significant. Confidence intervals present a range of values, on the basis of the sample data, in which the population value for such a difference may lie. Some methods of calculating confidence intervals for means and differences between means are given, with similar information for proportions. The paper also gives suggestions for graphical display.

Confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract.

Introduction

Over the past two or three decades the use of statistics in medical journals has increased tremendously. One unfortunate consequence has been a shift in emphasis away from the basic results towards an undue concentration on hypothesis testing. In this approach data are examined in relation to a statistical "null" hypothesis, and the practice has led to the mistaken belief that studies should aim at obtaining "statistical significance." On the contrary, the purpose of most research investigations in medicine is to determine the magnitude of some factor(s) of interest.

For example, a laboratory based study may investigate the difference in mean concentrations of a blood constituent between patients with and without a certain illness, while a clinical study may assess the difference in prognosis of patients with a particular disease treated by alternative regimens in terms of rates of cure, remission, relapse, survival, etc. The difference obtained in such a study will be only an estimate of what we really need, which is the result that would have been obtained had all the eligible subjects (the "population") been investigated rather than just a sample of them. What authors and readers should want to know is by how much the illness modified the mean blood concentrations or by how much the new treatment altered the prognosis, rather than only the level of statistical significance.

The excessive use of hypothesis testing at the expense of other ways of assessing results has reached such a degree that levels of significance are often quoted alone in the main text and abstracts of papers, with no mention of actual concentrations, proportions, etc, or their differences. The implication of hypothesis testing—that there can always be a simple "yes" or "no" answer as the fundamental result from a medical study—is clearly false and used in this way hypothesis testing is of limited value.[2]

We discuss here the rationale behind an alternative statistical approach—the use of confidence intervals; these are more informative than P values, and we recommend them for papers published in the *British Medical Journal* (and elsewhere). This should not be taken to mean that confidence intervals should appear in all papers; in some cases, such as where the data are purely descriptive, confidence intervals are inappropriate and in others techniques for obtaining them are complex or unavailable.

Presentation of study results: limitations of P values

The common simple statements "P<0·05," "P>0·05," or "P=NS" convey little information about a study's findings and rely on an arbitrary convention of using the 5% level of statistical significance to define two alternative outcomes—significant or not significant—which is not helpful and encourages lazy thinking. Furthermore, even precise P values convey nothing about the sizes of the differences between study groups. Rothman pointed this out in 1978 and advocated the use of confidence intervals,[3] and recently he and his colleagues repeated the proposal.[4]

Presenting P values alone can lead to them being given more merit than they deserve. In particular, there is a tendency to equate statistical significance with medical importance or biological relevance. But small differences of no real interest can be statistically significant with large sample sizes, whereas clinically important effects may be statistically non-significant only because the number of subjects studied was small.

Presentation of study results: confidence intervals

It is more useful to present sample statistics as estimates of results that would be obtained if the total population were studied. The lack of precision of a sample statistic—for example, the mean—which results from both the degree of variability in the factor being investigated and the limited size of the study, can be shown advantageously by a confidence interval.

A confidence interval produces a move from a single value estimate—such as the sample mean, difference between sample means, etc—to a range of values that are considered to be plausible for the population. The width of a confidence interval based on a sample statistic depends partly on its standard error, and hence on both the standard deviation and the sample size (see Appendix 1 for a brief description of the important, but often misunderstood, distinction between the standard deviation and standard error). It also

*In this paper we have preferred the notation of Mainland[1] and used P for the probability associated with the outcome of a test of the null hypothesis, and not p which is used for a proportion (see Appendix 2). Although contrary to the Vancouver convention, it is statistically more established and would also have been preferable for the statistical guidelines published in the *BMJ*.[2]

MRC Environmental Epidemiology Unit (University of Southampton), Southampton General Hospital, Southampton SO9 4XY
MARTIN J GARDNER, BSC, PHD, professor of medical statistics

Division of Medical Statistics, MRC Clinical Research Centre, Harrow, Middlesex HA1 3UJ
DOUGLAS G ALTMAN, BSC, medical statistician

Correspondence to: Professor Gardner.

depends on the degree of "confidence" that we want to associate with the resulting interval.

Suppose that in a study comparing samples of 100 diabetic and 100 non-diabetic men of a certain age a difference of 6·0 mm Hg was found between their mean systolic blood pressures and that the standard error of this difference between sample means was 2·5 mm Hg—comparable to the difference between means in the Framingham study.[5] The 95% confidence interval for the population difference between means is from 1·1 to 10·9 mm Hg and is shown in fig 1 together with the original data. Details of how to calculate the confidence interval are given in Appendix 2.

Put simply, this means that there is a 95% chance that the indicated range includes the "population" difference in mean blood pressure levels—that is, the value which would be obtained by including the total populations of diabetics and non-diabetics at which the study is aimed. More exactly, in a statistical sense, the confidence interval means that if a series of identical studies were carried out repeatedly on different samples from the same populations, and a 95% confidence interval for the difference between the sample means calculated in each study, then, in the long run, 95% of these confidence intervals would include the population difference between means.

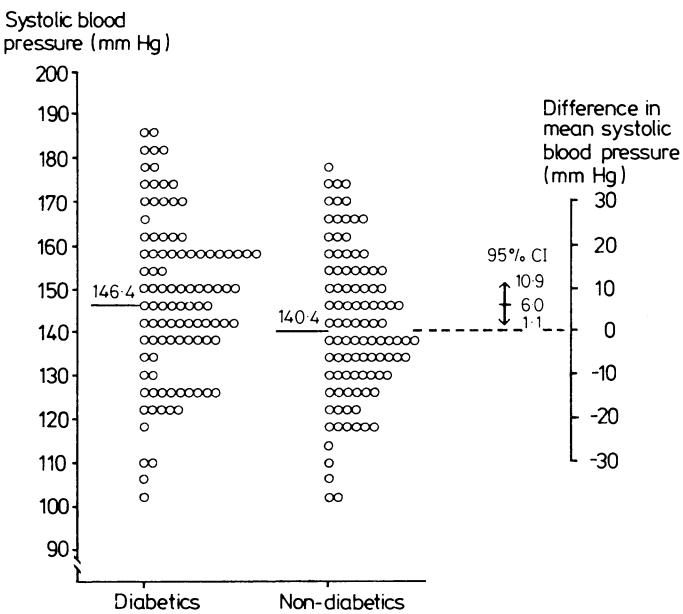

FIG 1—Systolic blood pressures in 100 diabetics and 100 non-diabetics with mean levels of 146·4 and 140·4 mm Hg respectively. The difference between the sample means of 6·0 mm Hg is shown to the right together with the 95% confidence interval from 1·1 to 10·9 mm Hg.

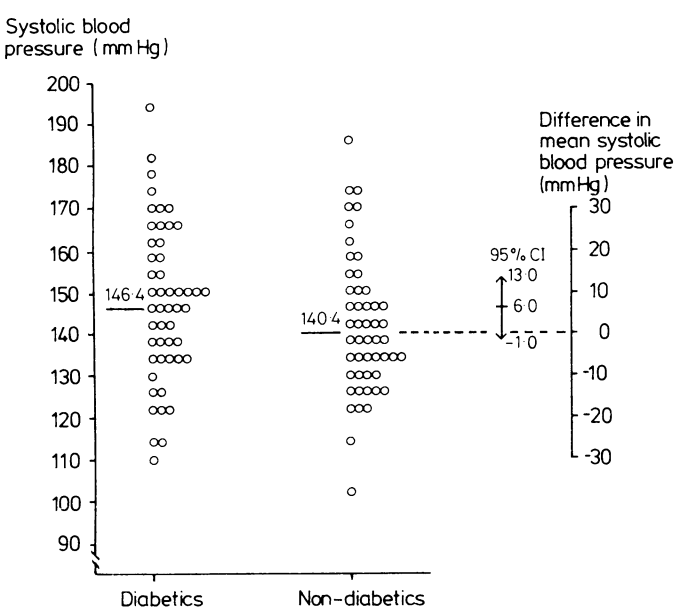

FIG 2—As fig 1 but showing results from two samples of half the size—that is, 50 subjects each. The means and standard deviations are as in fig 1, but the 95% confidence interval is wider, from −1·0 to 13·0 mm Hg, owing to the smaller sample sizes.

The sample size affects the size of the standard error and this in turn affects the width of the confidence interval. This is shown in fig 2, which shows the 95% confidence interval from samples with the same means and standard deviations as before but only half as large—that is, 50 diabetics and 50 non-diabetics. Reducing the sample size leads to less precision and an increase in the width of the confidence interval, in this case by some 40%.

The investigator can select the degree of confidence associated with a confidence interval, though 95% is the most common choice—just as a 5% level of statistical significance is widely used. If greater or less confidence is required different intervals can be constructed: 99%, 95%, and 90% confidence intervals for the data in fig 1 are shown in fig 3. As would be expected, greater confidence that the population difference is within a confidence interval is obtained with wider intervals. In practice, intervals other than 99%, 95% or 90% are rarely quoted.

Some methods of calculating confidence intervals for means, proportions, and their differences are given in Appendix 2. Confidence intervals can also be calculated for other statistics, such as regression slopes and relative risks.[6] When the observed data cannot be regarded as having come from a Normal distribution the situation is not always straightforward (see Appendix 2).

Confidence intervals convey only the effects of sampling variation on the precision of the estimated statistics and cannot control for non-sampling errors such as biases in design, conduct, or analysis.

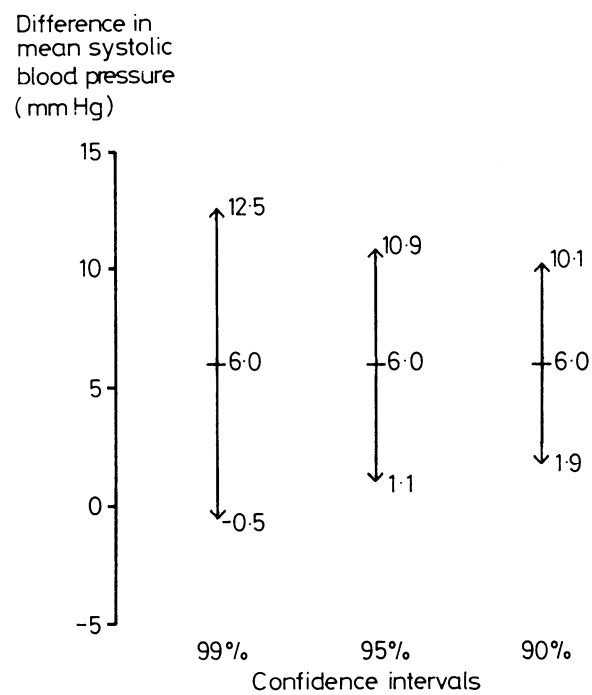

FIG 3—Confidence intervals associated with differing degrees of "confidence" using the same data as in fig 1.

Confidence intervals and statistical significance

There is a close link between the use of a confidence interval and a two sided hypothesis test. If the confidence interval is calculated then the result of the hypothesis test can be inferred at an associated level of statistical significance. The right hand scale in fig 1 includes the point that represents a zero difference in mean blood pressure between diabetics and non-diabetics. This zero difference between means corresponds to the value examined under the "null hypothesis" and, as fig 1 shows, it is outside the 95% confidence interval. This indicates that a statistically significant difference between the sample means at the 5% level would result from applying the appropriate unpaired *t* test. Fig 3, however, shows that the P value is greater than 1% because zero is inside the 99% confidence interval, so $0{\cdot}01<P<0{\cdot}05$. By contrast, had zero been within the 95% confidence interval this would have indicated a non-significant result at the 5% level. Such an example is shown in fig 2 for the smaller samples.

The 95% confidence interval covers a wide range of possible population mean differences, even though the sample difference between means is **different from zero at the 5% level of statistical significance. In particular,**

the 95% confidence interval shows that the study result is compatible with a small difference of around 1 mm Hg as well as with a difference as great as 10 mm Hg in mean blood pressures. Nevertheless, the difference between population means is much more likely to be near to the middle of the confidence interval than towards the extremes. Although the confidence interval is wide, the best estimate of the population difference is 6·0 mm Hg, the difference between the sample means.

This example therefore shows the lack of precision of the observed sample difference between means as an estimate of the population value, and this is clear in each of the three confidence intervals shown in fig 3. It also shows the weakness of considering statistical significance in isolation from the numerical estimates.

The confidence interval thus provides a range of possibilities for the population value, rather than an arbitrary dichotomy based solely on statistical significance. It conveys more useful information at the expense of precision of the P value. However, the actual P value is helpful in addition to the confidence interval, and preferably both should be presented. If one has to be excluded, however, it should be the P value.

Suggested mode of presentation

In content, our only proposed change is that confidence intervals should be reported instead of standard errors. This will encourage a move away from the current emphasis on statistical significance. For the major finding(s) of a study we recommend that full statistical information should be given, including sample estimates, confidence intervals, test statistics, and P values—assuming that basic details, such as sample sizes and standard deviations, have been reported earlier in the paper. The major findings would include at least those related to the original hypothesis(es) of the study and those reported in the abstract.

For the above example the textual presentation of the results might read:

> The difference between the sample mean systolic blood pressures in diabetics and non-diabetics was 6·0 mm Hg, with a 95% confidence interval from 1·1 to 10·9 mm Hg; the t test statistic was 2·4, with 198 degrees of freedom and an associated P value of P=0·02.

In short:

> Mean 6·0 mm Hg, 95% CI 1·1 to 10·9; t=2·4, df=198, P=0·02.

The exact P value from the t distribution is 0·01732, but one or two significant figures are enough[2]; this value is seen to be within the range 0·01 to 0·05 determined earlier from the confidence intervals. Often a range for P will need to be given because only limited figures are available in published tables—for example, 0·3<P<0·4.

The two extremes of a confidence interval are sometimes presented as confidence limits. However, the word "limits" suggests that there is no going beyond and may be misunderstood because, of course, the population value will not always lie within the confidence interval. Moreover, there is a danger that one or other of the "limits" will be quoted in isolation from the rest of the results, with misleading consequences. For example, concentrating only on the upper figure and ignoring the rest of the confidence interval would misrepresent the finding by exaggerating the study difference. Conversely, quoting only the lower limit would incorrectly underestimate the difference. The confidence interval is thus preferable because it focuses on the range of values.

The same notation can be used for presenting confidence intervals in tables. Thus, a column headed "95% confidence interval" or "95% CI" would have rows of intervals: "1·1 to 10·9", "48 to 85", etc. Confidence intervals can also be incorporated into figures, where they are preferable to the widely used standard error, which is often shown solely in one direction from the sample estimate. If individual data values can be shown as well, which is usually possible for small samples, this is even more informative. Thus in fig 1, despite the considerable overlap of the two sets of sample data, the shift in means is shown by the 95% confidence interval excluding zero. For paired samples the individual differences can be plotted advantageously in a diagram.

The example given here of the difference between two means is common. Although there is some intrinsic interest in the mean values themselves, inferences from a study will be concerned mainly with their difference. Giving confidence intervals for each mean separately is therefore unhelpful, because these do not usually indicate the precision of the difference or its statistical significance.[7,8] Thus, the major contrasts of a study should be shown directly, rather than only vaguely in terms of the separate means (or proportions).

For a paper with only a limited number of statistical comparisons related to the initial hypotheses confidence intervals are recommended throughout. Where multiple comparisons are concerned, however, the usual problems of interpretation arise, since some confidence intervals will exclude the "null" value—for example, zero difference—through sampling variation alone. This mirrors the situation of calculating a multiplicity of P values, where not all statistically significant differences are likely to represent real effects.[9] Judgment needs to be exercised over the number of statistical comparisons made, with confidence intervals and P values calculated, to avoid misleading both authors and readers.[2]

Conclusion

We have argued that the excessive use of hypothesis testing at the expense of more informative approaches to data interpretation is an unsatisfactory way of assessing and presenting statistical findings from medical studies. We prefer the use of confidence intervals, which present the results directly on the scale of data measurement. We have also suggested a notation for confidence intervals which is intended to force clarity of meaning.

Confidence intervals, which also have a link to the outcome of hypothesis tests, should become the standard method for presenting the statistical results of major findings.

We acknowledge the collaboration of the editorial staff of the *British Medical Journal* in the development of this paper and its proposals. We also thank the people who kindly read and constructively criticised the manuscript during its development and Miss Brigid Grimes for her careful typing.

Appendix 1: Standard deviation and standard error

When numerical findings are reported, regardless of whether or not their statistical significance is quoted, they are often presented with additional statistical information. The distinction between two widely quoted statistics —the standard deviation and the standard error—is, however, often misunderstood.[10-14]

The standard deviation is a measure of the variability between individuals in the level of the factor being investigated, such as blood alcohol concentrations in a sample of car drivers, and is thus a descriptive index. By contrast, the standard error is a measure of the uncertainty in a sample statistic. For example, the standard error of the mean indicates the uncertainty of the mean blood alcohol concentration among the sample of drivers as an estimate of the mean value among the population of all car drivers. The standard deviation is relevant when variability between individuals is of interest; the standard error is relevant to summary statistics such as means, proportions, differences, regression slopes, etc.[2]

The standard error of the sample statistic, which depends on both the standard deviation and the sample size, is a recognition that a sample is most unlikely to determine the population value exactly. In fact, if a further sample is taken in identical circumstances almost certainly it will produce a different estimate of the same population value. The sample statistic is therefore imprecise, and the standard error is a measure of this imprecision. By itself the standard error has limited meaning, but it can be used to produce a confidence interval, which does have a useful interpretation.

Appendix 2: Methods of calculating confidence intervals

Formulas for calculating confidence intervals (CIs) are given for means, proportions, and their differences. There is a common underlying principle of subtracting and adding to the sample statistic a multiple of its standard error (SE). This extends to other statistics, such as regression coefficients, but is not universal.

CONFIDENCE INTERVALS FOR MEANS AND THEIR DIFFERENCES

Confidence intervals for means are constructed using the t distribution if the data have an approximately Normal distribution. For differences between two means the data should also have similar standard deviations (SDs) in each study group. This is implicit in the example given in the text and in the worked example below.

Single sample

The confidence interval for a population mean is derived using the mean ($\bar{x}$) and its standard error from a sample of size n. For this case the $SE=SD/\sqrt{n}$. Thus, the confidence interval is given by:

$$\bar{x}-(t_{1-\alpha/2}\times SE) \quad \text{to} \quad \bar{x}+(t_{1-\alpha/2}\times SE),$$

where $t_{1-\alpha/2}$ is the appropriate value from the t distribution with n−1 degrees

of freedom associated with a "confidence" of $100(1-\alpha)\%$. For a 95% CI α is 0·05, for a 99% CI α is 0·01, and so on. Values of t can be found from tables in statistical textbooks or *Documenta Geigy*.[15] For a 95% CI the value of t will be close to 2·0 for samples of 20 upwards but noticeably greater than 2·0 for smaller samples.

Two samples

Unpaired case—The confidence interval for the difference between two population means is derived in a similar way. Suppose $\bar{x}_1$ and $\bar{x}_2$ are the two sample means, s_1 and s_2 the corresponding standard deviations, and n_1 and n_2 the sample sizes. Firstly, we need a "pooled" estimate of the standard deviation, which is given by:

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}.$$

From this the standard error of the difference between the two sample means is:

$$SE_{diff} = s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

The confidence interval is then:

$$\bar{x}_1-\bar{x}_2-(t_{1-\alpha/2}\times SE_{diff}) \quad \text{to} \quad \bar{x}_1-\bar{x}_2+(t_{1-\alpha/2}\times SE_{diff}),$$

where $t_{1-\alpha/2}$ is taken from the t distribution with n_1+n_2-2 degrees of freedom.

If the standard deviations differ considerably then a common pooled estimate is not appropriate unless a suitable transformation of scale can be found. Otherwise obtaining a confidence interval is more complex.[6]

Paired case—This includes studies of repeated measurements—for example, at different times or in different circumstances on the same subjects—and matched case-control comparisons. For such data the same formulas as for the single sample case are used to calculate the confidence interval, where $\bar{x}$ and SD are now the mean and standard deviation of the individual within subject or patient-control differences.

Worked example: two unpaired samples

Blood pressure levels were measured in 100 diabetic and 100 non-diabetic men aged 40-49 years. Mean systolic blood pressures were 146·4 mm Hg (SD 18·5) among the diabetics and 140·4 mm Hg (SD 16·8) among the non-diabetics, giving a difference between sample means of 6·0 mm Hg.

Using the formulas given above the pooled estimate of the standard deviation is:

$$s = \sqrt{\frac{(99\times 18{\cdot}5^2)+(99\times 16{\cdot}8^2)}{198}} = 17{\cdot}7 \text{ mm Hg},$$

and the standard error of the difference between the sample means is:

$$SE_{diff} = 17{\cdot}7\sqrt{\frac{1}{100}+\frac{1}{100}} = 2{\cdot}50 \text{ mm Hg}.$$

To calculate the 95% CI the appropriate value of $t_{0\cdot975}$ with 198 degrees of freedom is 1·97. Thus the 95% CI is given by:

$$6{\cdot}0-(1{\cdot}97\times 2{\cdot}50) \quad \text{to} \quad 6{\cdot}0+(1{\cdot}97\times 2{\cdot}50)$$

that is, from 1·1 to 10·9 mm Hg, as shown in fig 1.

Suppose now that the samples had been of only 50 men each but that the means and standard deviations had been the same. Then the pooled standard deviation would remain 17·7 mm Hg, but the standard error of the difference between the sample means would become:

$$SE_{diff} = 17{\cdot}7\sqrt{\frac{1}{50}+\frac{1}{50}} = 3{\cdot}54 \text{ mm Hg}.$$

The appropriate value of $t_{0\cdot975}$ on 98 degrees of freedom is 1·98, and the 95% CI is calculated as:

$$6{\cdot}0-(1{\cdot}98\times 3{\cdot}54) \quad \text{to} \quad 6{\cdot}0+(1{\cdot}98\times 3{\cdot}54)$$

that is, from −1·0 to 13·0 mm Hg, as shown in fig 2.

For the original samples of 100 each the appropriate values of $t_{0\cdot995}$ and $t_{0\cdot95}$ with 198 degrees of freedom to calculate the 99% and 90% CIs are 2·60 and 1·65, respectively. Thus the 99% CI is calculated as:

$$6{\cdot}0-(2{\cdot}60\times 2{\cdot}50) \quad \text{to} \quad 6{\cdot}0+(2{\cdot}60\times 2{\cdot}50)$$

that is, from −0·5 to 12·5 mm Hg (fig 3), and the 90% CI is given by:

$$6{\cdot}0-(1{\cdot}65\times 2{\cdot}50) \quad \text{to} \quad 6{\cdot}0+(1{\cdot}65\times 2{\cdot}50)$$

that is, from 1·9 to 10·1 mm Hg (fig 3).

Sample sizes and confidence intervals

In general increasing the sample size will reduce the width of the confidence interval. If we assume the same means and standard deviations as in the example fig 4 shows the resulting 99%, 95%, and 90% confidence intervals for the difference in mean blood pressures for sample sizes of up to 500 in each group. The benefit, in terms of narrowing the confidence interval, of a further increase in the number of subjects falls sharply with increasing sample size.

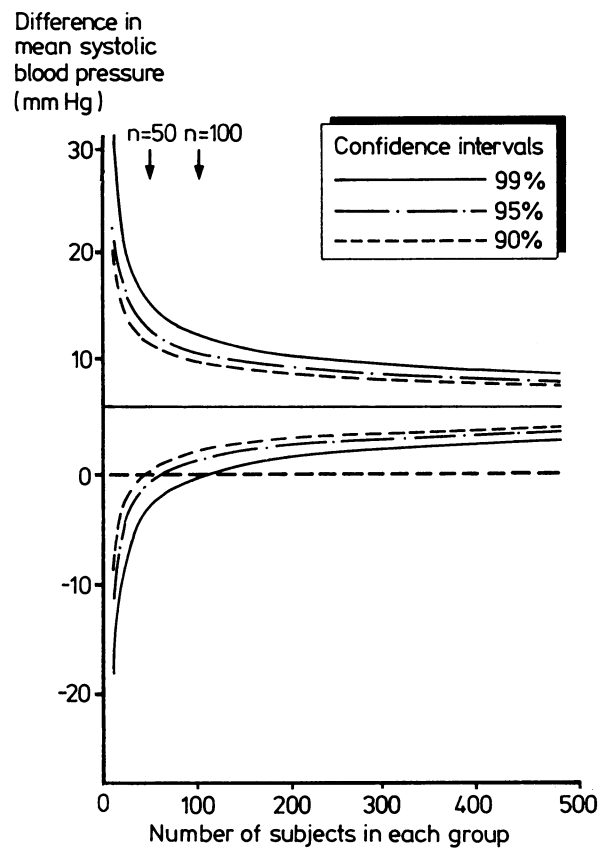

FIG 4—Confidence intervals resulting from the same means and standard deviations as in fig 1 and given in the worked example, but showing the effect on the confidence interval of sample sizes of up to 500 subjects in each group. The two horizontal lines show: --- zero difference between means, —— study difference between means of 6·0 mm Hg. The arrows indicate the confidence intervals shown in figs 1-3 for sample sizes of 100 and 50 in each group.

Non-Normal data

The sample data may have to be transformed on to a different scale to achieve approximate Normality. The most common reason is because the distribution of the observations is skewed, with a long "tail" of high values. The logarithmic transformation is the most frequently used.

For a single sample a mean and confidence interval can be constructed from the transformed data and then transformed back to the original scale of measurement. This is preferable to presenting the results in units of, say, log mm Hg. With highly skewed or otherwise awkward data the median may be preferable to the mean as a measure of central tendency and used with non-parametric methods of analysis. Confidence intervals can be calculated for the median.[15]

For the case of two samples only the logarithmic transformation is suitable. For paired or unpaired samples the confidence interval for the difference in the means of the transformed data has to be transformed back. For the log transformation the anti-log of the difference in sample means on the transformed scale is an estimate of the ratio of the two population (geometric) means, and the anti-logged confidence interval for the difference gives a confidence interval for this ratio. Other transformations do not lead to sensible confidence intervals when transformed back.

CONFIDENCE INTERVALS FOR PROPORTIONS AND THEIR DIFFERENCES

Confidence intervals for proportions, or differences between two proportions, can be constructed similarly. The formulas given below should not be used for small samples—for example, fewer than 50 in each group and proportions outside the range 0·1 to 0·9. A continuity correction can be incorporated,[16] as is sometimes done for the χ^2 test of the difference between proportions in a 2×2 table.

Single sample

If p is the observed proportion of subjects with some feature in a sample of size n then the standard error of p is $SE=\sqrt{p(1-p)/n}$. The $100(1-\alpha)\%$ confidence interval for p is given by:

$$p-(N_{1\ \alpha/2}\times SE) \quad \text{to} \quad p+(N_{1\ \alpha/2}\times SE),$$

where $N_{1\ \alpha/2}$ is the appropriate value from the standard Normal distribution for the $100(1-\alpha/2)$ percentile found in widely available tables. Thus, for a 95% CI $N_{1\ \alpha/2}=1{\cdot}96$; this value does not depend on the sample size, as it does for means.

Two samples

Unpaired case—The confidence interval for the difference between two population proportions is constructed round p_1-p_2, the difference between the observed proportions in the two samples. The standard error of p_1-p_2 in this case is:

$$SE_{diff}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}.$$

The confidence interval is then given by:

$$p_1-p_2-(N_{1\ \alpha/2}\times SE_{diff}) \quad \text{to} \quad p_1-p_2+(N_{1\ \alpha/2}\times SE_{diff}),$$

where $N_{1\ \alpha/2}$ is found as for the single sample case.

Paired case—Suppose that a sample of n subjects has twice been examined for the presence or absence of a particular feature. The data can be tabulated thus:

Feature at time 1	Feature at time 2	Number of subjects
Present	Present	a
Present	Absent	b
Absent	Present	c
Absent	Absent	d
Total		n

Then the proportions of subjects with the feature on the two occasions are $p_1=(a+b)/n$ and $p_2=(a+c)/n$, and the difference between them is $p_1-p_2=(b-c)/n$. The standard error of this difference is:

$$SE_{diff}=\frac{1}{n}\sqrt{b+c-\frac{(b-c)^2}{n}}.$$

The confidence interval for p_1-p_2 is then given as:

$$p_1-p_2-(N_{1\ \alpha/2}\times SE_{diff}) \quad \text{to} \quad p_1-p_2+(N_{1\ \alpha/2}\times SE_{diff}),$$

where $N_{1-\alpha/2}$ is found as for the single sample case.

Worked example: two unpaired samples

Response to treatment was assessed among 160 patients randomised to either treatment A or treatment B with the following results:

Response	Treatment A	Treatment B
Improvement	61	45
No improvement	19	35
Total	80	80

The proportions whose condition improved were $p_A=0{\cdot}76$ and $p_B=0{\cdot}56$ (61/80 and 45/80) for treatments A and B respectively, which indicates a preferential improvement proportion of 0·20 for treatment A. In terms of percentages 76% of patients on treatment A improved compared with 56% on treatment B, suggesting that an extra 20% of patients would improve if given A rather than B.

The standard error of the difference $p_A-p_B=0{\cdot}20$ from the formula for the unpaired case is:

$$\sqrt{\frac{0{\cdot}76\times0{\cdot}24}{80}+\frac{0{\cdot}56\times0{\cdot}44}{80}}=0{\cdot}073.$$

The 95% CI is then given by:

$$0{\cdot}20-(1{\cdot}96\times0{\cdot}073) \quad \text{to} \quad 0{\cdot}20+(1{\cdot}96\times0{\cdot}073)$$

that is, from 0·06 to 0·34. Thus, although the best estimate of the difference in the percentage of patients improving is 20%, the 95% CI ranges from 6% to 34%, showing the imprecision due to the limited sample size.

The usual χ^2 test for these data gives a numerical value of: $\chi^2=7{\cdot}16$, df=1, P=0·007, for which the level of statistical significance is consistent with the 99% CI (using $N_{0{\cdot}995}=2{\cdot}58$) of 0·01 to 0·39.

Technical note

Although for quantitative data and means there is a direct correspondence between the confidence interval approach and a *t* test of the null hypothesis at the associated level of statistical significance, this is not exactly so for qualitative data and proportions. The reason is related to the use of different estimates of the standard error for the usual tests of the null hypothesis from those given here for constructing confidence intervals. The lack of direct correspondence is small and should not result in changes of interpretation. In addition, more accurate confidence intervals can sometimes be obtained by using estimates of the standard error of the sample statistic at the confidence limits themselves—such as derived by Cornfield for relative risks.[17]

References

1 Mainland D. Statistical ritual in clinical journals: is there a cure?—I. *Br Med J* 1984;**288**:841-3.
2 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;**286**:1489-93.
3 Rothman K. A show of confidence. *N Engl J Med* 1978;**299**:1362-3.
4 Poole C, Lanes S, Rothman KJ. Analysing data from ordered categories. *N Engl J Med* 1984;**311**:1382.
5 Kannel WB, McGee DL. Diabetes and cardiovascular risk factors: the Framingham study. *Circulation* 1979;**59**:8-13.
6 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971.
7 Browne RH. On visual assessment of the significance of a mean difference. *Biometrics* 1979;**35**:657-65.
8 Altman DG. Statistics and ethics in medical research: VI—presentation of results. *Br Med J* 1980;**281**:1542-4.
9 Jones DR, Rushton L. Simultaneous inference in epidemiological studies. *Int J Epidemiol* 1982;**11**:276-82.
10 Gardner MJ. Understanding and presenting variation. *Lancet* 1975;i:230-1.
11 Feinstein AR. Clinical biostatistics XXXVII: demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clin Pharmacol Ther* 1976;**20**:617-31.
12 Bunce H, Hokanson JA, Weiss GB. Avoiding ambiguity when reporting variability in biomedical data. *Am J Med* 1980;**69**:8-9.
13 Altman DG. Statistics in medical journals. *Statistics in Medicine* 1982;**1**:59-71.
14 Brown GW. Standard deviation, standard error: which "standard" should we use? *Am J Dis Child* 1982;**136**:937-41.
15 Diem K, Lentner C, eds. *Documenta Geigy. Scientific tables*. 7th ed. Basle: Geigy, 1970.
16 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. Chichester: Wiley, 1981:29-30.
17 Breslow NE, Day NE. *Statistical methods in cancer research: volume 1—the analysis of case-control studies*. Lyon: International Agency for Research on Cancer, 1980:133-4.

(Accepted 8 January 1986)