Statistics on microcomputers. A non-algebraic guide to their appropriate use in biomedical research and pathology laboratory practice

2 Confidence intervals and significance tests

R A BROWN,† J SWANSON BECK*

*From the Departments of †Mathematical Sciences and *Pathology, University of Dundee, Scotland*

Populations and samples

The previous article described methods for displaying and summarising data. Usually this is not an end in itself, and research workers will wish to extend their conclusions from their samples or experiments to medical practice in general by statistical analysis. It is an essential precondition of valid statistical inference that the groups of subjects included in an investigation are selected randomly from the populations which are being compared, and its aim is to make statements about these global populations based on sample data.

One of the simplest forms of investigation is the comparison of measurements made on two well defined groups of subjects which differ in some clearly specified way—for example, medical treatment, pathological state. The subjects in each group must constitute a random sample from the corresponding population, and as far as practicable, the only major difference between the two groups should be the factor under investigation.

Before the start of an investigation it is essential to define clearly the populations to be sampled, particularly when the investigation is expected to result in recommendations of methods of treatment, diagnosis, or prognosis. Failing this, although it is still possible to carry through all stages of a statistical analysis, the final conclusions will be less valuable as they will refer to poorly defined populations. It is regrettable that many scientific reports of otherwise clearly described investigations lack unambiguous definition of criteria for group selection.

An entire population may be accessible at the time of data collection but cost precludes the measurement of every subject. In other situations, especially in disease of gradual onset, the observer can identify fully established disease, but even though he or she is aware of their presence he or she does not have the access or

aptitude to identify subjects with presymptomatic defects, such as Alzheimer's disease. In either situation the investigator will make measurements on groups of subjects who have been selected randomly from the corresponding populations. Consequently there must always be some uncertainty in any general conclusions concerning differences between the populations inferred from differences between the samples, and so the methods used to summarise the conclusions must be based on probability.

General statements often refer to *population means*; these are the values which would be obtained if every member of each population was measured and the mean was calculated for each population. The *population standard deviation* measures the spread of values over the entire population and, like the population mean, it is the value which would be obtained if it were possible to measure every individual and calculate the standard deviation of all the results. It is important to distinguish between the *population mean and standard deviation* and a *sample mean and standard deviation* to appreciate fully the discussion of fundamental concepts of statistical inference which follows later in this article. The population mean and standard deviation are inaccessible, but they can be estimated by the *sample mean and standard deviation*, calculated from a random sample of members of a population.

Statistical inference is concerned with:

- (i) making statements about population means or standard deviations, based on the limited information provided by the corresponding sample values;
- (ii) quantifying the uncertainty associated with these statements.

The normal distribution

Many standard statistical analyses of continuous data are based on the assumption that the spread of values across the population can be described by the so-called "normal" distribution. This form of distribution was

Fig 1 Normal distribution is symmetrical and unimodal: the area under any portion of the curve indicates the probability that an observation will lie within the corresponding range, for example, 68% of observations will lie within 1 SD of the mean.

originally selected as the basis of much statistical work because it is unimodal and symmetrical, with a high proportion of values relatively close to the mean and decreasing concentration of values in the “tails” of the distribution, similar to many distributions observed in practice. It is undoubtedly a convenient approximation for many sets of data, but it would be wrong to assume that this particular distribution was a biologically “natural” pattern to which all “good”

data conform—for example, the distribution of survival times of patients with cancer is not well approximated by the normal distribution.

The normal distribution curve illustrates the way in which a measurement might be expected to vary across a population (fig 1). The peak of the curve is located at the population mean, and there is a precise relation between the population mean and standard deviation and the proportion of measurements lying within any specified range. For example, the statement that 95% of values will be within 1.96 population standard deviations either side of the population mean is true for every normal distribution.

As normality is an important assumption for several statistical methods it is important to be able to assess whether a particular set of data shows evidence of serious non-normality. The common methods of statistical analysis are fairly “robust to non-normality” provided that:

- (i) the data are more or less symmetrically distributed;
- (ii) the spread of results is similar in the groups being compared;
- (iii) there are no outliers in the data.

These methods can give rise to misleading results if these requirements are not met.

A simple graphical technique known as normal plotting may be used to assess normality. When performed manually the data values in a single sample are ordered and each value is plotted against the value one would expect to observe if the data were normally

Fig 2 Survival of 60 patients with breast cancer. (a) Dot plot indicates asymmetry; (b) box and whisker plot indicates that the upper 50% of the survivals are spread over a much greater range than the lower 50%; (c) normal plot shows pronounced curvature: the data distribution is very non-normal.

Fig 3 Transcutaneous measurements of pO_2 in forearm skin of 54 normal subjects. (a) Dot plot shows no noticeable asymmetry; (b) box and whisker plot clearly indicates near symmetry of the distribution: the median is more or less centrally placed, the whiskers are of almost equal length; (c) normal plot no serious departure from normality.

distributed. If the data are approximately normally distributed then the plotted points should show a straight line trend. The method is tedious to use manually but with suitable programs a normal plot can be displayed within minutes on a microcomputer. In Statgraphics the individual observations are displayed on an appropriate scale and the best fitting straight line is included so that visual assessment of departure from linearity can be made. If the data are effectively normally distributed the plotted points will be close to the fitted line, but will show pronounced deviations when the data are seriously non-normal.

To indicate how the method works in practice, consider the data on survival times of a group of patients with breast cancer, displayed as a dot plot in fig 2a, a box and whisker plot in fig 2b, and as a normal plot in fig 2c. Mean survival at 58.5 months is small compared with the sample standard deviation (45 months), which indicates possible asymmetry. The box and whisker plot confirms this, indicating serious asymmetry, and the normal plot, which shows considerable curvature, rules out the normal distribution as an adequate approximation to these data. For comparison, fig 3 shows the corresponding displays for a set of 60 normally distributed measurements (in this case the pO_2 in normal forearm skin measured transcutaneously): the pattern of points shows no serious departure from the fitted straight line.

Estimating the population mean using a confidence interval

When a population is sampled, the sample mean is an estimate of the population mean. Other samples would produce different estimates and so it is important to be able to estimate the imprecision of an estimate from any given sample. Statisticians invented the term "confidence interval" to describe the range of plausible values for the population mean that are consistent with the observed sample mean.

Underlying the concept of the confidence interval is the realisation that the means of repeated samples will themselves have a distribution known as the *sampling distribution of the mean*. When the samples are selected from a normal population, the sampling distribution of the sample mean is normal, with the same mean as the population, but with a standard deviation which is equal to the population standard deviation divided by the square root of the sample size. The standard deviation of the sampling distribution of the mean is called the "standard error of the mean" (SEM). Because the sampling distribution is normal, 95% of all sample means will lie within 1.96 SEM of the population mean. Given the actual value of the sample mean, the following argument produces the confidence interval for the population mean: suppose the observed sample mean is, in fact, one of the 95%

MTB > describe "pO ₂ "					
pO ₂	MIN	MAX	Q1	Q3	
	67.57	100.34	76.56	88.81	
MTB > TINTERVAL for "pO ₂ "					
pO ₂	N	MEAN	STDEV	SE MEAN	95.0 PERCENT C.I.
	30	82.20	8.97	1.64	(78.85, 85.55)
MTB > TINTERVAL 99% for "pO ₂ "					
pO ₂	N	MEAN	STDEV	SE MEAN	99.0 PERCENT C.I.
	30	82.20	8.97	1.64	(77.68, 86.71)

Fig 4 Minitab display giving basic description and 95% confidence interval for the mean tcpO₂ in forearm skin of a sample of 30 normal subjects. The first command gives the basic descriptive statistics, the second calculates by default the 95% confidence interval, and the third displays the user selected 99% confidence interval.

that fall within 1.96 SEM of the population mean, then the population mean cannot be more than 1.96 SEM away from the sample mean; consequently the 95% confidence interval for the population mean is the range (sample mean—1.96 SEM to sample mean + 1.96 SEM). We cannot be absolutely certain, however, that the sample mean is one of the 95% of sample means that are within 1.96 SEM of the population mean; there is a 5% chance that it is one of those that are not. So the statement "the population mean is within 1.96 SEM of the sample mean" will be correct on 95 occasions out of 100, which is why the interval is called a 95% confidence interval. There is a minor technical difficulty because the SEM requires the population standard deviation value, which will be unknown in any practical situation; the sample standard deviation, however, can be used instead when calculating the SEM, with a correction to take into account the uncertainty arising from estimation of the population standard deviation by the sample standard deviation. This changes the number 1.96 to a somewhat larger value, depending on the sample size; it is 2.23 for samples of size 10, 2.09 for a sample of size 20, and 2.00 for samples of size 60 (values for all sample sizes may be obtained from a table of percentage points of the *t*-distribution¹).

Of course, there is no particular reason why the 95% probability level should be especially singled out; a higher value such as 99% might be preferable as then there is only one chance in 100 that the confidence interval fails to bracket the value of the population mean. By convention, the commonly used confidence intervals are 90%, 95%, and 99%. The higher the value selected for the probability that a given confidence interval will include the population mean, the wider that interval will be. The point is illustrated as follows: 30 normal subjects were selected at random to estimate mean po₂ in normal forearm skin, and the sample mean was 82.20 with a sample standard

deviation of 8.97. Output from a Minitab analysis is shown in fig 4. The 95% confidence interval for the population mean tcpO₂ stretches from 78.85 to 85.55; the corresponding 99% confidence interval is wider, stretching from 77.68 up to 86.71: the wider range of probable values for the population mean is the price which must be paid for being more confident that this range includes the population mean value.

Comparative investigations

Many medical research investigations entail comparisons between subjects selected from various well defined subpopulations, with the aim of detecting medically or scientifically significant differences between the overall distribution of certain variables in the different subpopulations. The simplest comparative investigation entails two subpopulations, which may be, for example, a population of normal controls and a population of patients with a particular disease. Such an investigation would be carried out by randomly selecting roughly the same number of subjects from each subpopulation and making the appropriate measurement on each subject. We will refer to this as the *two sample comparison*. A second, similar, type of investigation is known as the *paired comparison* investigation, in which the same measurement is made on each of several subjects on two occasions—for example, before and after treatment—and so each subject acts as his or her own control for the purpose of assessing changes brought about by treatment. One must be careful to distinguish between the two types of comparative investigation as the appropriate statistical analyses are not interchangeable.

The analysis of the two sample comparison assumes that:

- (i) the measurements in each subpopulation are normally distributed;

Fig 5 Haemoglobin concentration in men and women (80 subjects in each group).

Fig 6 White cell counts in men and women (80 subjects in each group).

(ii) the difference between the subpopulations manifests itself only as a difference between the corresponding population means;
(iii) the standard deviation is the same in both subpopulations.

In the next article we shall describe alternative analyses for use in situations where these assumptions do not hold.

A confidence interval for the difference between the population means provides the answer to the question of whether the data show that the population means are different and indicates how large the difference is likely to be.

The procedure is very similar to that used to determine a confidence interval in the single sample case. As an example, the data presented in fig 5 were obtained in an investigation that compared measurements of haemoglobin concentration in the blood of 80 healthy men and a similar group of 80 healthy women. The sample mean and standard deviation were 154.8 g/l and 24.9 g/l for the men and the mean was 140.2 g/l and the standard deviation 28.1 g/l for the women. The difference between the sample means was 14.6 g/l and the standard error of this difference was 4.2 g/l. The 95% confidence interval for the difference between the population means stretched from 6.3 g/l to 22.9 g/l. This is interpreted as indicating that one can be 95% sure that average haemoglobin concentration in men can be expected to be between 6.3 g/l and 22.9 g/l higher than in women. When there is only a small difference between population means—for example, in leucocyte counts, a similar investigation might produce the data illustrated in fig 6, with sample means of $6.97 \times 10^9/l$ and $7.04 \times 10^9/l$ and standard deviations of $1.05 \times 10^9/l$ and $0.95 \times 10^9/l$. A 95% confidence interval for the difference between the population mean for men and the population mean for women stretches from $-0.38 \times 10^9/l$ up to $0.24 \times 10^9/l$. In this case the supposition that there is no difference between the population means cannot be

Fig 7 Effect of sample size on the width of a confidence interval. The bars represent confidence intervals for the difference in mean haemoglobin concentration using the means and standard deviations from the data of fig 5, but assuming sample sizes ranging from 10 up to 500 in each group.

ruled out as zero falls inside the confidence interval.

The confidence interval is recommended as superior to any other method of statistical analysis in this situation because it gives a clear indication of the size of the difference between the population means, together with an equally clear indication of the imprecision in our knowledge of the actual difference. Thus there can be no doubt in the reader's mind as to whether a reported difference is medically important as the probable size of the difference is quoted. Nor can there be much doubt about whether the sample sizes were large enough to detect worthwhile differences in an investigation which reports a "negative" finding—that is, a confidence interval which includes zero. For example, if a difference of 5 units were medically important, but the confidence interval reported in a particular investigation stretched from -20 up to $+40$, its width indicates that variability is so great that a much larger investigation would be

required to reduce the width of the interval to a level where one could be optimistic about detecting a difference of 5 units.

As an indication of how sample size affects the width of a confidence interval, fig 7 shows the end points of 95% confidence intervals for investigations with the same means and standard deviations as in fig 5, and calculated from samples varying in size from 10 up to 500. The gain, in terms of decreased width, from using larger sample sizes is quite pronounced up to 100 per group, but declines sharply thereafter.

The analysis of results arising from a paired comparison investigation can be summarised briefly as no new ideas are involved. Recall that two measurements are made on each subject to estimate the difference between the population means before and after treatment. The appropriate method of analysis is to subtract the first measurement from the second for each subject and calculate a confidence interval for the mean of these differences. If the confidence interval does not straddle zero this will indicate that treatment does affect the variable of interest.

Confidence intervals compared with significance tests

There is a strong association between the confidence interval approach and an alternative commonly used in medical literature, based on the idea of a "significance test", in which the difference between the sample means is divided by the standard error. The result is conventionally denoted by the symbol "*t*" and the method is known as the "*two sample t-test*" (in the paired comparison, it is the mean of the differences divided by its standard error which is referred to as "*t*" and the corresponding test is the "*paired t-test*").

The significance testing approach is based on the following argument: suppose the population means are equal, then we would expect the sample means to be roughly equal, apart from the inevitable difference resulting from sampling, and so their difference will be "close" to zero. The standard error of the difference is the appropriate scale factor for judging whether the observed difference is small enough for it to be regarded as evidence that the population means are equal. Thus "*t*" is the appropriate statistic for making the assessment. If "*t*" is large then this is taken as evidence against the supposition that the population means are equal. A "large" value is one which has a low probability of occurring when the population means are equal. This probability is called the "p-value" in the medical literature and the "significance level" by statisticians.

The logic of the *t*-test is as follows. If the p value associated with the result is small, say 0.01, then there are two possibilities:

(i) the population means *are equal* but the investiga-

tion has produced data which have only one chance in 100 of occurring or;

(ii) the population means are not equal.

The conventional approach has been to regard an investigation as having produced evidence in favour of a genuine difference whenever the p value is less than 0.05, and then the result is reported as being "significant".

The shortcomings of the significance testing approach are well known to statisticians and have been pointed out on several occasions.^{2,3} Firstly, the result of an investigation is being forced into a framework where all that matters is whether a result is or is not significant. This is surely not wise, as what really matters is whether an investigation has been able to detect a medically important difference between the population means and how large that difference is likely to be. Secondly, the significance test approach pretends that all situations can easily be neatly categorised as ones where the population means are or are not exactly equal. In reality there is likely to be some difference between the population means and so a very large investigation will stand a good chance of detecting the difference, no matter how small and medically unimportant it may be. The method can easily lead to the attitude that one should interpret, for example, a "significant ($p < 0.01$)" result as meaning that a difference which has a practical importance has been detected.

On the other hand, an investigation which results in a "not significant ($p > 0.05$)" difference risks being misinterpreted as providing conclusive evidence that a truly zero difference between the population means has been established. Nothing could be further from the truth. The correct interpretation is that any underlying difference is too small to have been detected by this particular investigation (the investigation may in fact have been too small to detect really worthwhile differences). One cannot use statistical analysis to prove conclusively that there is absolutely no difference between a pair of population means.

By comparison, the confidence interval approach is superior. As it gives a range of plausible values for the difference between the population means, the knowledgeable subject specialist is able to judge whether the difference being reported is of practical importance. In fact, the confidence interval and the significance test are closely related. If the 95% confidence interval for the mean difference does not bracket zero, then the *t*-test will produce a result which is significant at the 5% level and perhaps at an even lower level. Conversely, if the *t*-test results in $p > 0.05$ then the 95% confidence interval will include zero, but the interval gives a more useful interpretation of the "not significant" result: the true difference may not be zero but could be any value within the interval.

Two-sample analysis results				
Sample statistics:	Number of Obs.	HbM	HbF	Pooled
	Average	80	80	160
	Variance	154.817	140.238	147.528
	Std. Deviation	622.298	789.773	706.035
	Median	24.9459	28.1029	26.5713
Conf. interval for diff. in means:		95 Percent		
(Equal Vars.)	Sample 1-Sample 2	6.28021	22.8798	158 D.F.
(Unequal Vars.)	Sample 1-Sample 2	6.27931	22.8807	155.8 D.F.
Conf. interval for ratio of variances:		95 Percent		
	Sample 1 Sample 2	0.50532	1.22864	
Hypothesis test for HO: Diff = 0		Computed t statistic = 3.4703		
vs Alt: NE		Sig. Level = 6.7024E-4		
at Alpha = 0.05		so reject HO		

Fig 8 Screen display produced by Statgraphics in response to a request for a two sample *t*-test analysis of the haemoglobin data of fig 5.

As both the *t*-test and the confidence interval approach are available in all good statistical packages it is advisable to combine the two. To illustrate this we compared the data of fig 5 using the "two sample" analysis from Statgraphics. The screen output is shown in fig 8; the major descriptive statistics for the two samples are shown in the top of the field. Note that both the standard deviations and the variances, which are merely the squares of the standard deviations, are displayed. Then the 95% confidence interval for the difference between the population means is given (the user can choose an alternative level such as 99%): the display next shows the results of two forms of the interval, the first being the one we have already described, which assumes equal population standard deviations; the second is an alternative that is appropriate when this assumption is clearly invalid. The user may choose whichever is more appropriate, depending on the variances shown in the upper part of the panel. Thereafter the user has an opportunity for calculating a confidence interval for the ratio of the population variances at a probability level of his or her choice; this is a further guide to which of the alternative intervals to use. If the confidence interval for the ratio of variances brackets the value one then the equal variances confidence interval is appropriate. The bottom section of the panel is written in computer statistics jargon, but we can interpret the computer output into plain English as follows: "the data provide evidence of a difference between mean haemoglobin concentrations in men and women. The difference

between the sample means for two groups of 80 subjects was 14.6 g/l, this difference is significant at $p < 0.001$, and a 95% confidence interval for the difference between healthy men and women stretches from 6.3 g/l to 22.9 g/l".

This final display can be "dumped" to the printer for a permanent record at a single key stroke. With other simple manipulations dot plots or box and whisker plots can be displayed and recorded permanently with a graph plotter. A bench-based laboratory scientist can thus obtain reliable comparisons between groups of experimental observations and can produce various forms of display to clarify the interpretation quickly on a microcomputer. Used properly, this facility will extend the range and improve the efficiency of many types of laboratory work for a very small financial outlay.

References

- 1 Diem K, Lentner C, eds. *Documenta Geigy, Scientific Tables*. Basle: Geigy, 1970.
- 2 Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746-50.
- 3 Rothman K. A show of confidence. *N Engl J Med* 1978;299:1362-3.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, PO Box 120, Dundee DD1 9SY, Scotland.