J. Paediatr. Child Health (2001) **37**, 72–77

Statistics for Clinicians

4: Basic concepts of statistical reasoning: Hypothesis tests and the *t*-test

JB CARLIN[1,4] and LW DOYLE[2–4]

[1]*Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Parkville,* [2]*Division of Newborn Services, Royal Women's Hospital, Melbourne, and the Departments of* [3]*Obstetrics and Gynaecology, and* [4]*Paediatrics, University of Melbourne, Parkville, Victoria, Australia*

In the previous article in this series,[1] we introduced the concept of *sampling variability,* and showed how it underlies the calculation and interpretation of *standard errors* and *confidence intervals (CIs).* An example that was used to illustrate these ideas was the comparison of verbal IQ at age 5 years between a group of children who were born with extremely low birthweight (< 1000 g) and another group of children of birthweight 1000–1500 g (see Fig. 3 of the previous article). Although there was a mean difference of 5.5 IQ points (94.7 *vs* 100.2, more than one-third of a standard deviation) between the two groups, 95% CIs for the two mean values overlapped quite substantially. A natural question is whether the observed difference reflects a real underlying difference or not. The overlapping CIs perhaps suggest that the answer is 'no', but the reader should be warned that examining overlap of CIs is not equivalent to formal hypothesis testing of differences between groups, which is the traditional approach to this sort of question. (We shall return later to the relationship between hypothesis testing and confidence intervals.)

This article explains the basic logic of hypothesis tests and in particular discusses the *t*-test, one of the more common statistical tests, encountered in about one-third of original articles in the *Journal of Paediatrics and Child Health*.[2]

HYPOTHESIS TESTS AS SIGNAL DETECTION

Those who like to listen to their stereo will be familiar with the concept of a signal-to-noise ratio – the stereo sounds clearer the *louder the signal* and the *lower the noise* in the generated signal. The basic structure of the statistical test is similar to a signal-to-noise ratio. The 'signal' corresponds to a measure of the difference between the groups. The 'noise' is essentially the *standard error* (of the difference), which quantifies the sampling variability and thereby the statistical *uncertainty* of the comparison measure. The standard error of a difference measure is directly related to the inherent variation within the groups, and inversely to the square root of the sample size (n). The formula for many test statistics can then be expressed informally as:

$$\text{Test statistic} \quad \alpha \quad \frac{\text{Signal}}{(\text{Variation within Groups})/\sqrt{n}}$$

(where the symbol 'α' means 'proportional to').

The larger the value of this signal-to-noise ratio, the more likely it is that the observed difference reflects a true underlying difference. Large values correspond to what is termed 'statistical significance' (more on this below). It is easy to see that statistical significance can arise with any combination of a large difference between the groups, or with less inherent variation within the groups, or with a larger sample size. For the *t*-test, the signal is the size of the difference between the two group means, and the 'noise' denominator is the standard error of the difference.

To illustrate some of these points graphically, the simulated data in Fig. 1 (panels A–D) compare two groups on a continuous scale of measurement (remember, it is always useful to plot your data when comparing groups). The data within the panels vary in (i) the size of the difference between the two groups, and (ii) the variation within the groups. In panel A, it is hard to tell if the difference between the two groups is statistically significant, because there is a small average difference between them and large variation within each group, with the result that the 95% CI for each group overlap substantially. However, as either the difference between the groups increases (panel B), or the variation within the groups diminishes (panel C), or both (panel D), the overlap between the 95% CIs reduces (panels B,C) or disappears (panel D) and the two groups look more clearly different. We will use these data later to calculate *t*-tests for the respective panels to confirm our visual impression of statistical significance or non-significance.

HYPOTHESIS TESTS AND *P* VALUES: BASIC PRINCIPLES

In statistical terms, we make sense of the question 'Is this difference real?' by using the concept of sampling variability and comparing the size of difference observed with the range of values that *might have been expected if there were no true difference.* What do we mean by 'no true difference'? This refers to the underlying *population parameters*, as distinct from our *sample statistics*, which are based on the observed data. As always, statistical inference is meaningful only when

Correspondence: Associate Professor LW Doyle, Department of Obstetrics and Gynaecology, The University of Melbourne, Parkville 3010, Australia. Fax: (03) 9347 1761; email: lwd@unimelb.edu.au

Reprints not available.

Accepted for publication 14 November 2000.

we keep a clear distinction between the underlying (unobservable) true reality (for example, the mean difference in 5-year verbal IQ between the populations of *all* children in the two birthweight groups to whom we might wish to generalize our findings) and the 'noisy' results that we find in our current study. So, loosely speaking, hypothesis tests create a probabilistic index (the *P* value) that provides a measure of the 'suspicion' with which we may view the conservative hypothesis that there is no true difference.

More specifically, the method of hypothesis testing starts with an assumption that the groups that are being compared come from the same population, or at least from populations with the same mean. This is called the 'null' hypothesis; sometimes written as 'H_0'. For comparison, there is the alternative hypothesis, H_1 say, that the two groups come from populations with different means. Referring back to the signal-to-noise ratio analogy, H_0 can be identified as the case where the data contain only noise and no signal, and H_1 is the case where there is an important signal amongst the noise. The statistician then calculates a *test statistic*, which gives a measure of the observed difference in the sample data, in a standardized form whose sampling distribution may be calculated under the null hypothesis. In particular, using a computer package or a set of statistical tables, one may obtain the *P value*, defined as:

$$P = \text{Probability}\left(\begin{array}{l}\text{we would observe a test statistic at}\\ \text{least as extreme as the value found in}\\ \text{our data, assuming the null hypothesis}\\ \text{is true}\end{array}\right)$$

If the *P* value is low, the researcher must logically conclude *either* that something rather surprising has happened (under the null hypothesis), *or* that the null hypothesis is not true. Formally, when the *P* value is low enough, we decide to reject the null hypothesis, i.e. the notion that the groups came from the same population, and conclude that the groups are 'statistically significantly different'. Traditionally, $P < 0.05$, or $< 5\%$, has been recommended as a threshold at which one may reject the null hypothesis, but there is nothing magical (or indeed particularly scientific) about 0.05. Sometimes a lower threshold is used, reflecting a greater desire for conservatism; in other words only to claim a difference is real when the sample data are especially strong. The use of formal levels at which the null hypothesis is rejected or accepted has diminished in recent years with a greater emphasis placed on the interpretation of findings in their scientific context, and in particular on the *clinical* or *substantive* importance of the findings.

COMPARING TWO MEANS: THE *t*-TEST

Before considering further issues in the interpretation of hypothesis tests, we describe and illustrate the *t*-test for comparing two means, using the data shown in Fig. 1. The computations required are reasonably straightforward, as shown in Table 1.

The data from Panel A in Fig. 1, shown in Table 1, produce a *t*-statistic of –0.69 with a corresponding *P* value of 0.50. This means that a result on this test could be expected to be as extreme as this on 50% of occasions when there was no true

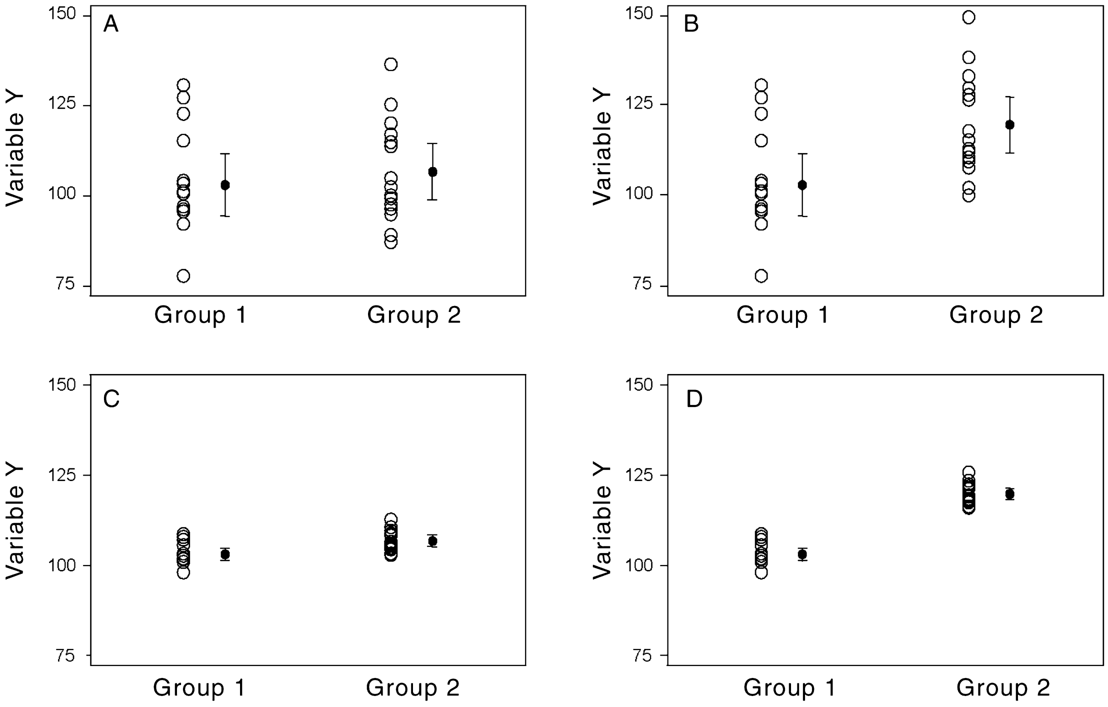

Fig. 1 Comparison of variable Y between two groups, with raw data, mean and 95% confidence interval shown for each group within each panel (based on sample of size 15 within each group). Panel A: Mean difference small, variation large. Panel B: Mean difference large, variation large. Panel C: Mean difference small, variation small. Panel D: Mean difference large, variation small.

Table 1 Summary statistics and t-test calculations for the data in Figure 1

		Panel A		Panel B		Panel C		Panel D	
Summary statistics		Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
Mean	$\bar{y}_1, \bar{y}_2$	103.0	106.7	103.0	119.7	103.0	106.7	103.0	119.7
SD	s_1, s_2	15.7	14.1	15.7	14.1	3.1	2.8	3.1	2.8
Sample size	n_1, n_2	15	15	15	15	15	15	15	15
Two-sample *t*-test calculations									
Mean difference	$\bar{y}_1 - \bar{y}_2$	– 3.73		– 16.7		– 3.73		– 16.7	
Pooled SD	$s = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-1}}$	14.89		14.89		2.98		2.98	
SE of mean difference	$\text{SE} = s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$	5.44		5.44		1.09		1.09	
Test statistic	$\dfrac{\bar{y}_1 - \bar{y}_2}{\text{SE}}$	– 0.69 –		3.08		– 3.44		– 15.4	
Degrees of of freedom (d.f.)	$n_1 + n_2 - 2$	28		28		28		28	
P value	Prob $(t \le -\dfrac{\bar{y}_1 - \bar{y}_2}{\text{SE}}$ or $t \ge \dfrac{\bar{y}_1 - \bar{y}_2}{\text{SE}})$	0.50		0.005		0.002		< 0.0001	

SD, standard deviation; SE, standard error.

difference. Because the *P* value is greater than 0.05, we cannot claim statistical significance, and would conclude that the data provide no evidence for a real difference in scores between these two groups. In contrast, the data from the other three panels all produce higher *t*-statistics and lower *P* values, which are all statistically significant at the usual level of 0.05. Note that the difference between means in Panel B is highly significant despite the fact that the values in the two groups overlap substantially. It is important to remember that the hypotheses being tested relate to the population means and not to the individual values themselves. Finally, as an aside, the data in all four panels were simulated from distributions for all of which there was a true difference between the two groups. The reason that the data and corresponding *t*-test for Panel A show no evidence of this is that the sample size was relatively small, given that the true mean difference (five units) was small relative to the variability (true SD = 15).

With modern computers, we usually do not have to calculate a *t*-test for our own data, since statistical packages are far quicker and more accurate. However, occasionally we do not have the raw data, and we only have the means and SD themselves, and perhaps we need to check the accuracy of someone else's computations. Another reason for looking at least briefly at the computational formulae for the *t*-test is that they provide some insight into what is being done. In particular, it can be seen (Table 1) that the test is based on calculating a test statistic, called the *t*-statistic, which has the signal-to-noise ratio form of difference between means divided by the standard error of that difference.

INTERPRETATION OF HYPOTHESIS TESTS

It is important to remember that the *P* value from a *t*-test used to compare two groups only tells us if the mean difference between the groups is *statistically significant* or not. We still have the problem of assessing if any difference is *clinically significant* (or important). One way to compare the means of two groups, to convey both statistical and clinical significance, is to present the mean difference between the groups along with its 95% CI, and we will discuss this further in a subsequent article. As previously discussed, the *P* value is determined not only by the actual size of the observed difference but also by the sample size: small differences become statistically significant as sample size increases. This is a general feature of all hypothesis tests. Beware the very large study that reports an unimportant clinical difference as statistically significant, and, equally, the very small study that reports a potentially important clinical difference as not statistically significant.

More generally, the reader may observe that the logic of hypothesis testing is somewhat convoluted, and indeed it has long been the subject of controversy in the statistical literature.

Goodman has recently provided an excellent review for a clinical readership.[3] We do not have space here to cover all of the issues that have been raised, but it is important to note just a few.

A key concern is the extent to which one can interpret a P value as quantifying *evidence* for or against a research hypothesis. Many clinicians would informally and intuitively expect a P value to mean something like 'the probability that two or more groups come from the same population.' Not only is this an incorrect interpretation (see the earlier definition), but it may actually lead to erroneous conclusions. It can be shown, for example, that under most circumstances a P value of 0.05 actually represents much weaker evidence against the null hypothesis than a naive interpretation, along the lines of 'a 5% chance that the two groups are the same' would suggest.[4] A key problem in talking about evidence is that no study or P value can be interpreted in isolation.[5] The final interpretation of a study's findings must depend on what is known about the subject matter from other studies, from relevant biological understanding, and so on. Goodman shows that a result with $P = 0.05$ in fact means that the data reduce the chance that the null hypothesis is true to no less than 15% of whatever prior probability might be attached to it, while a P value approaching 0.01 is required to achieve a reduction to less than 5%.[4]

Further interpretational issues with hypothesis tests relate to long-run error rates and the concepts of Type I and Type II errors. We will return to these issues in a later article when we discuss sample size and power calculations.

ONE-TAILED AND TWO-TAILED TESTS

All the P values calculated in Table 1 are *two-tailed* (or 'two-sided'), which means that their calculations included the probability of extreme values occurring in either direction (i.e. higher or lower) from zero, the value expected under the null hypothesis. Extremes in both tails of the distribution were included because the alternative hypothesis, H_1, specified only that the means of the two groups were different, not that they differed in a particular direction. If the investigator can confidently rule out, on *a priori* grounds, that the difference between two comparison groups could go in one direction, then it is justified to specify a one-sided alternative hypothesis and use a *one-tailed* test. Assuming the observed sample difference is in the same direction as specified by the alternative hypothesis, the one-tailed P value is exactly half the two-tailed value, because of the symmetry of the normal and t distributions.

The question of whether and when it is legitimate to use a one-tailed test (thereby, in general, producing a smaller P value) has aroused some controversy. But given the simple relationship between the two P values, and the move away from automated interpretation of tests using arbitrary levels such as 0.05 or 0.01, the issue is not really a major one. We follow the common convention of recommending that two-tailed tests be used unless there is a very strong case for the one-tailed version.[6]

We illustrate this by returning to the example of verbal IQ at age 5 years in the very low birthweight cohort. The standard two-sample t-test calculation produces a t-statistic of –2.03 with a corresponding P value of 0.04, for comparing the mean verbal IQ of children with birthweight < 1000 g with that of children with birthweight 1000–1500 g. This means that there is marginally significant evidence for a true difference using the 0.05 significance level. If it could be confidently assumed *a priori* that the true mean verbal IQ in the lower birthweight group could not possibly be greater than in the higher birthweight group, a one-tailed test would be justified, and this would give a P value of 0.02. However, there is no reason to think that the verbal IQ in the heavier children should necessarily be higher than in the lower birthweight children, and hence a one-tailed test could not be justified. The conventional advice is that such *a priori* assumptions are usually difficult to sustain (surprising reversals of effect do occur), and that the implicit conservatism of the two-tailed result is to be preferred. We will also see that this approach maintains a closer connection between the interpretation of hypothesis tests and CIs.

MORE ON THE *t*-TEST: RELATING THE *t*-STATISTIC TO THE NORMAL DISTRIBUTION

We saw in the last article[1] that when the sample size is at least moderate, the sampling distribution of a sample mean is *normal*, centred around the population mean and with standard deviation equal to the *standard error of the mean* (SEM). In the context of hypothesis testing, where the parameter of interest is a difference, the population mean under the null hypothesis is zero. These facts imply that, if the null hypothesis is true, when we divide the sample mean by its SE we obtain a statistic that has (approximately) a standard normal distribution (mean 0, SD 1). We can then readily determine, for example, that a value of 3 would be surprising and give rise to a very small P value for the null hypothesis that the true mean is zero. The ratio of an estimate to its SE is often referred to as a *z-statistic*, since 'Z' is often used to refer to the standard normal distribution. This terminology is analogous to use of the *z-score* to refer to the standardized version of an individual value of a variable, such as height for age and gender, for example.

With small samples, the ratio of a mean difference to its SE is called a t-statistic, since its distribution is not in fact normal but the closely related t, assuming that the variable being measured itself follows a normal distribution. The t distribution has a very similar, symmetric, shape to the normal, but it is more spread out. This extra spread arises for t-statistics because of the fact that the standard deviation of the measurements (used in calculating the SE) has to be estimated from the data. The standard t-distribution (scaled like the standard normal) has one extra parameter to define it: the so-called *degrees of freedom (*d.f.*).* The d.f. appropriate for any t-statistic essentially reflect the sample size and thus the uncertainty of estimation of the standard deviation. As the sample size, and so d.f.*,* increases, the t-distribution quite rapidly approaches the standard normal distribution.

The t-statistic for comparing two means is the observed difference between the means, standardized by the (estimated) standard error of that difference. The appropriate d.f. for the standard two-sample (independent groups) t-test is $n_1 + n_2–2$. Since the t-distribution is similar to the normal, as long as d.f. are not too small, we can make informal interpretations of the statistical significance of t statistics by referring them to the cut-offs of ± 2 ($P \approx 0.05$), ± 2.6 ($P \approx 0.01$) and ± 3.3 ($P \approx 0.001$).

THE PAIRED *t*-TEST

A different version of the t-test is required if the data in two groups are not independent, but represent paired values.

Table 2 Summary statistics and calculations for paired *t*-test on the data in Figure 2

Summary statistics		
Mean difference	$\frac{\sum \text{differences}}{n}$ or $\bar{y}_1 - \bar{y}_2$	–3.73
SD of difference	s_D	6.01
Sample size	n	15
Paired *t*-test calculations		
SE of mean difference	$SE = \frac{s_D}{\sqrt{n}}$	1.55
Test statistic	$\frac{\bar{y}_1 - \bar{y}_2}{SE}$	–2.40
Degrees of freedom (d.f.)	$n - 1$	14
P value	Prob ($t \leq -\frac{\bar{y}_1 - \bar{y}_2}{SE}$ or $t \geq \frac{\bar{y}_1 - \bar{y}_2}{SE}$)	0.03

(where the symbol 'Σ' means 'the sum of')

A typical example is a before–after study in a group of patients where an observation is made of a continuous variable, something is done to the patient (they might be given a drug, for example), and the variable is then re-measured. It is natural to expect that observations taken on the same patient before and after the drug administration will be more alike than observations taken on different patients. The two-sample *t*-test described above does not take any account of such pairing and so is an inappropriate method for paired data. However, the appropriate *paired t-test* is similar in its logical structure.

As an example, we can reconsider the data in Panel A of Fig. 1, this time assuming they arose as paired observations in a cross-over trial, linked as shown in Fig. 2. In most, but not all, pairs there is an increase between Time 1 and Time 2. The calculations required for the paired *t*-test are shown in Table 2. These are actually simpler than for the two-sample test and proceed from first calculating the *paired difference* values, and then calculating their mean. Note that the mean of the difference values is the same as the difference of the two separate means, but the SE calculation for the paired comparison is different. It is worth noting that the *t*-statistic is now statistically significant ($P = 0.03$), whereas using the independent groups calculation (Table 1; Panel A) it was not. The SE of the mean difference ('noise') was substantially reduced in the paired comparison because the large variation *between* individuals was removed by focusing on the *within*-individual changes. Clinicians should take advantage of situations where data can be paired so that statistical answers to clinical questions can be achieved more efficiently, i.e. with fewer patients.

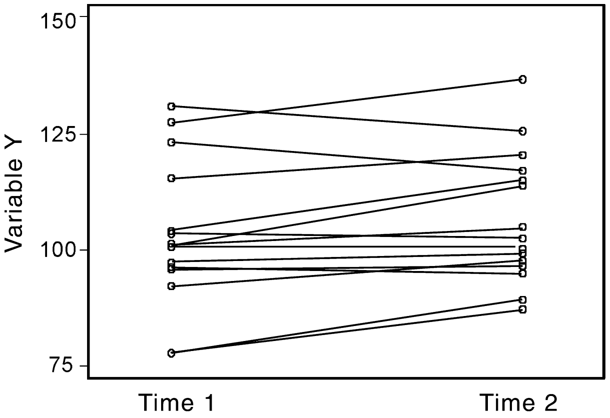

Fig. 2 Comparison of variable Y between two time points, with paired data.

WHEN IS THE *t*-TEST VALID?

For normally distributed variables, or for variables that can be transformed into a normal distribution by taking logarithms, for example, the difference between group means is of major interest and the *t*-test is clearly applicable. If the data are not normally distributed and cannot be made so, it may be preferable to make comparisons using other parameters. With positively skewed data, medians are often used. Tests based on medians are called 'non-parametric tests' and these will be discussed later in the statistics series.

One practical concern that researchers often face is to determine whether their data are sufficiently close to normally distributed for the *t*-test to be valid. It is important to remember the role of sample size in this: with large samples, the distribution of the data values themselves becomes less and less relevant to the statistical properties of test statistics (because of the Central Limit Theorem). With small samples it is also important to be aware that the assumption of normality of the data is probably less critical than the question of whether the variances in the two groups can be assumed to be equal. In fact the *t*-test is fairly *robust* (statistician's language for giving a valid answer) to moderate departures from its underlying assumptions of normally distributed data and equality of variance, except in the presence of very small or unequal sample sizes. If sample sizes differ non-trivially between the groups, it is advised to use a 'conservative' variation of the standard test that does not assume equal variances.[7,8] This method, using the so-called Satterthwaite approximation, is available in most statistical packages.

In the next statistics article, we will discuss the comparison of groups where the variable being considered is a binary indicator (such as 'dead/alive', or 'remission/no remission'), in particular describing the Chi-squared test for comparing proportions.

REFERENCES

1 Carlin JB, Doyle LW. Statistics for clinicians. 3: Basic concepts of statistical reasoning: Standard errors and confidence intervals. *J. Paediatr. Child Health* 2000; **36**: 502–5.

2 Doyle LW, Carlin JB. 2: Statistics for clinicians. I: Introduction. *J. Paediatr. Child Health* 2000; **36**: 74–5.

3 Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann. Intern. Med.* 1999; **130**: 995–1004.

4 Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* 1999; **130**: 1005–13.

5 Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987; **257**: 2459–63.

6 Motulsky H. *Intuitive Biostatistics.* Oxford University Press, New York, 1995.

7 Armitage P, Berry G. *Statistical Methods in Medical Research*, 3rd edn. Blackwell Scientific Publications, Oxford, 1994.

8 Moser BK, Stevens GR. Homogeneity of variance in the two-sample means test. *Am. Statistician* 1992; **46**: 19–21.