Review

Statistics review 5: Comparison of means

Elise Whitley[1] and Jonathan Ball[2]

[1]Lecturer in Medical Statistics, University of Bristol, Bristol, UK
[2]Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 12 July 2002 *Critical Care* 2002, **6**:424-428
This article is online at http://ccforum.com/content/6/5/424
© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review introduces the commonly used t-test, used to compare a single mean with a hypothesized value, two means arising from paired data, or two means arising from unpaired data. The assumptions underlying these tests are also discussed.

Keywords comparison of two means, paired and unpaired data, t test

Previous reviews in this series have introduced the principals behind the calculation of confidence intervals and hypothesis testing. The present review covers the specific case of comparing means in rather more detail. Comparison of means arises in many different formats, and there are various methods available for dealing with each of these. Some of the simpler cases are covered in this review, namely comparison of a single observed mean with some hypothesized value, comparison of two means arising from paired data, and comparison of two means from unpaired data. All of these comparisons can be made using appropriate confidence intervals and t-tests as long as certain assumptions are met (see below). Future reviews will introduce techniques that can be used when the assumptions of the t-test are not valid or when the comparison is between three or more groups.

Of the three cases covered in this review, comparison of means from unpaired data is probably the most common. However, the single mean and paired data cases are introduced first because the t-test in these cases is more straightforward.

Comparison of a single mean with a hypothesized value

This situation is not very common in practice but on occasion it may be desirable to compare a mean value from a sample with some hypothesized value, perhaps from external standards. As an example, consider the data shown in Table 1. These are the haemoglobin concentrations of 15 UK adult males admitted into an intensive care unit (ICU). The population mean haemoglobin concentration in UK males is 15.0 g/dl. Is there any evidence that critical illness is associated with an acute anaemia?

The mean haemoglobin concentration of these men is 9.7 g/dl, which is lower than the population mean. However, in practice any sample of 15 men would be unlikely to have a mean haemoglobin of exactly 15.0 g/dl, so the question is whether this difference is likely to be a chance finding, due to random variation, or whether it is the result of some systematic difference between the men in the sample and those in the general population. The best way to determine which explanation is most likely is to calculate a confidence interval for the mean and to perform a hypothesis test.

The standard deviation (SD) of these data is 2.2 g/dl, and so a 95% confidence interval for the mean can be calculated using the standard error (SE) in the usual way. The SE in this case is $2.2/\sqrt{15} = 0.56$ and the corresponding 95% confidence interval is as follows.

$$9.7 \pm 2.14 \times 0.56 = 9.7 \pm 1.19 = (8.5, 10.9)$$

Note that the multiplier, in this case 2.14, comes from the t distribution because the sample size is small (for a fuller explanation of this calculation, see Statistics review 2 from this series). This confidence interval gives the range of likely values for the mean haemoglobin concentration in the population

ICU = intensive care unit; SD = standard deviation; SE = standard error.

Table 1

Haemoglobin concentrations (g/dl) for 15 UK males admitted into an intensive care unit

8.1	10.1	12.3
9.7	11.7	11.3
11.9	9.3	13.0
10.5	8.3	8.8
9.4	6.4	5.4

Table 2

Central venous oxygen saturation on admission and 6 h after admission to an intensive care unit

	Central venous oxygen saturation (%)		
Subject	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	−2.4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50.0	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56.0	59.5	3.5
10	65.3	59.8	−5.5
Mean	52.4	59.2	6.8

from which these men were drawn. In other words, assuming that this sample is representative, it is likely that the true mean haemoglobin in the population of adult male patients admitted to ICUs is between 8.5 and 10.9 g/dl. The haemoglobin concentration in the general population of adult men in the UK is well outside this range, and so the evidence suggests that men admitted to ICUs may genuinely have haemoglobin concentrations that are lower than the national average.

Exploration of how likely it is that this difference is due to chance requires a hypothesis test, in this case the one sample t-test. The t-test formally examines how far the estimated mean haemoglobin of men admitted to ICU, in this case 9.7 g/dl, lies from the hypothesized value of 15.0 g/dl. The null hypothesis is that the mean haemoglobin concentration of men admitted to ICU is the same as the standard for the adult male UK population, and so the further away the sample mean is from this hypothesized value, the less likely it is that the difference arose by chance.

The t statistic, from which a *P* value is derived, is as follows.

$$t = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{SE of sample mean}} \qquad (1)$$

In other words, t is the number of SEs that separate the sample mean from the hypothesized value. The associated *P* value is obtained by comparison with the t distribution introduced in Statistics review 2, with larger t statistics (regardless of sign) corresponding to smaller *P* values. As previously described, the shape of the t distribution is determined by the degrees of freedom, which, in the case of the one sample t-test, is equal to the sample size minus 1.

The t statistic for the haemoglobin example is as follows.

$$t = \frac{9.7 - 15.0}{0.56} = \frac{-5.3}{0.56} = -9.54$$

In other words, the observed mean haemoglobin concentration is 9.54 SEs below the hypothesized mean. Tabulated values indicate how likely this is to occur in practice, and for a sample size of 15 (corresponding to 14 degrees of freedom) the *P* value is less than 0.0001. In other words, it is extremely unlikely that the mean haemoglobin in this sample would differ from that in the general population to this extent by chance alone. This may indicate that there is a genuine difference in haemoglobin concentrations in men admitted to the ICU, but as always it is vital that this result be interpreted in context. For example, it is important to know how this sample of men was selected and whether they are representative of all UK men admitted to ICUs.

Note that the *P* value gives no indication of the size of any difference; it merely indicates the probability that the difference arose by chance. In order to assess the magnitude of any difference, it is essential also to have the confidence interval calculated above.

Comparison of two means arising from paired data

A special case of the one sample t-test arises when paired data are used. Paired data arise in a number of different situations, such as in a matched case–control study in which individual cases and controls are matched to each other, or in a repeat measures study in which some measurement is made on the same set of individuals on more than one occasion (generally under different circumstances). For example, Table 2 shows central venous oxygen saturation in 10 patients on admission and 6 hours after admission to an ICU.

The mean admission central venous oxygen saturation was 52.4% as compared with a mean of 59.2% after 6 hours, corresponding to an increase of 6.8%. Again, the question is whether this difference is likely to reflect a genuine effect of admission and treatment or whether it is simply due to chance. In other words, the null hypothesis is that the mean central venous oxygen saturation on admission is the same as the mean saturation after 6 hours. However, because the data are paired, the two sets of observations are not independent of each other, and it is important to account for this pairing in the analysis. The way to do this is to concentrate on the differences between the pairs of measurements rather than on the measurements themselves.

The differences between the admission and post-admission central venous oxygen saturations are given in the rightmost column of Table 2, and the mean of these differences is 6.8%. In these terms, the null hypothesis is that the mean of the differences in central venous oxygen saturation is zero. The appropriate t-test therefore compares the observed mean of the differences with a hypothesized value of 0. In other words, the paired t-test is simply a special case of the single sample t-test described above.

The t statistic for the paired t-test is as follows.

$$t = \frac{\text{sample mean of differences} - 0}{\text{SE of sample mean of differences}}$$

$$= \frac{\text{sample mean of differences}}{\text{SE of sample mean of differences}} \qquad (2)$$

The SD of the differences in the current example is 7.5, and this corresponds to a SE of $7.5/\sqrt{10} = 2.4$. The t statistic is therefore $t = 6.8/2.4 = 2.87$, and this corresponds to a *P* value of 0.02 (based on a t distribution with $10 - 1 = 9$ degrees of freedom). In other words, there is some evidence to suggest that admission to ICU and subsequent treatment may increase central venous oxygen saturation beyond the level expected by chance.

However, the *P* value in isolation gives no information about the likely size of any effect. As indicated above, this is rectified by calculating a 95% confidence interval from the mean and SE of the differences. In this case the 95% confidence interval is as follows.

$$6.8 \pm 2.26 \times 2.4 = 6.8 \pm 5.34 = (1.4, 12.2)$$

This indicates that the true increase in central venous oxygen saturation due to ICU admission and treatment in the population is probably between 1.4% and 12.2%. The decision as to whether this difference is likely to be important in practice should be based on the statistical evidence in combination with other relevant clinical factors. However, it is worth noting that the confidence interval excludes 0 (the expected difference if the null hypothesis were true); thus, although the increase may be small (1.4%), it is unlikely that the effect is to decrease saturation.

Table 3

Mean and standard deviation of mean arterial pressure

	Mean arterial pressure (mmHg)	
	Standard therapy	Early goal-directed therapy
Number of patients	119	117
Mean	81	95
Standard deviation	18	19

Comparison of two means arising from unpaired data

The most common comparison is probably that of two means arising from unpaired data (i.e. comparison of data from two independent groups). For example, consider the results from a recently published trial that compared early goal-directed therapy with standard therapy in the treatment of severe sepsis and septic shock [1]. A total of 263 patients were randomized and 236 completed 6 hours of treatment. The mean arterial pressures after 6 hours of treatment in the standard and early goal-directed therapy groups are shown in Table 3.

Note that the authors of this study also collected information on baseline mean arterial pressure and examined the 6-hour pressures in the context of these (using a method known as analysis of covariance) [1]. In practice this is a more appropriate analysis, but for illustrative purposes the focus here is on 6-hour mean arterial pressures only.

It appears that the mean arterial pressure was 14 mmHg higher in the early goal-directed therapy group. The 95% confidence intervals for the mean arterial pressure in the two groups are as follows.

$$\text{Standard therapy: } 81 \pm 1.96 \times \frac{18}{\sqrt{119}} = 81 \pm 3.23 = (77.8, 84.2)$$

$$\text{Early goal-directed therapy: } 95 \pm 1.96 \times \frac{19}{\sqrt{117}} = 95 \pm 3.44 = (91.6, 98.4)$$

There is no overlap between the two confidence intervals and, because these are the ranges in which the true population values are likely to lie, this supports the notion that there may be a difference between the two groups. However, it is more useful to estimate the size of any difference directly, and this can be done in the usual way. The only difference is in the calculation of the SE.

In the paired case attention is focused on the mean of the differences; in the unpaired case interest is in the difference of the means. Because the sample sizes in the unpaired case may be (and indeed usually are) different, the combined SE takes this into account and gives more weight to the larger sample size because this is likely to be more reliable. The pooled SD for the difference in means is calculated as follows:

$$\text{SD}_{\text{difference}} = \sqrt{\frac{(n_1 - 1) \times \text{SD}_1^2 + (n_2 - 1) \times \text{SD}_2^2}{(n_1 + n_2 - 2)}} \qquad (3)$$

where SD_1 and SD_2 are the SDs in the two groups and n_1 and n_2 are the two sample sizes. The pooled SE for the difference in means is then as follows.

$$\text{SE}_{\text{difference}} = \text{SD}_{\text{difference}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (4)$$

This SE for the difference in means can now be used to calculate a confidence interval for the difference in means and to perform an unpaired t-test, as above.

The pooled SD in the early goal-directed therapy trial example is:

$$\text{SD}_{\text{difference}} = \sqrt{\frac{(119 - 1) \times 18^2 + (117 - 1) \times 19^2}{(119 + 117 - 2)}}$$

$$= \sqrt{\frac{38{,}232 + 41{,}876}{234}} = \sqrt{342.34} = 18.50$$

and the corresponding pooled SE is:

$$\text{SE}_{\text{difference}} = 18.50 \times \sqrt{\frac{1}{119} + \frac{1}{117}} = 18.50 \times \sqrt{0.008 + 0.009}$$

$$= 18.50 \times 0.13 = 2.41$$

The difference in mean arterial pressure between the early goal-directed and standard therapy groups is 14 mmHg, with a corresponding 95% confidence interval of $14 \pm 1.96 \times 2.41 = (9.3, 18.7)$ mmHg. If there were no difference in the mean arterial pressures of patients randomized to early goal-directed and standard therapy then the difference in means would be close to 0. However, the confidence interval excludes this value and suggests that the true difference is likely to be between 9.3 and 18.7 mmHg.

To explore the likely role of chance in explaining this difference, an unpaired t-test can be performed. The null hypothesis in this case is that the means in the two populations are the same or, in other words, that the difference in the means is 0. As for the previous two cases, a t statistic is calculated.

$$t = \frac{\text{difference in sample means}}{\text{SE of difference in sample means}}$$

A *P* value may be obtained by comparison with the t distribution on $n_1 + n_2 - 2$ degrees of freedom. Again, the larger the t statistic, the smaller the *P* value will be.

In the early goal-directed therapy example $t = 14/2.41 = 5.81$, with a corresponding *P* value less than 0.0001. In other words, it is extremely unlikely that a difference in mean arterial pressure of this magnitude would be observed just by chance. This supports the notion that there may be a genuine difference between the two groups and, assuming that the randomization and conduct of the trial was appropriate, this suggests that early goal-directed therapy may be successful in raising mean arterial pressure by between 9.3 and 18.7 mmHg. As always, it is important to interpret this finding in the context of the study population and, in particular, to consider how readily the results may be generalized to the general population of patients with severe sepsis or septic shock.

Assumptions and limitations

In common with other statistical tests, the t-tests presented here require that certain assumptions be made regarding the format of the data. The one sample t-test requires that the data have an approximately Normal distribution, whereas the paired t-test requires that the distribution of the differences are approximately Normal. The unpaired t-test relies on the assumption that the data from the two samples are both Normally distributed, and has the additional requirement that the SDs from the two samples are approximately equal.

Formal statistical tests exist to examine whether a set of data are Normal or whether two SDs (or, equivalently, two variances) are equal [2], although results from these should always be interpreted in the context of the sample size and associated statistical power in the usual way. However, the t-test is known to be robust to modest departures from these assumptions, and so a more informal investigation of the data may often be sufficient in practice.

If assumptions of Normality are violated, then appropriate transformation of the data (as outlined in Statistics review 1) may be used before performing any calculations. Similarly, transformations may also be useful if the SDs are very different in the unpaired case [3]. However, it may not always be possible to get around these limitations; where this is the case, there are a series of alternative tests that can be used. Known as nonparametric tests, they require very few or very limited assumptions to be made about the format of the data, and can therefore be used in situations where classical methods, such as t-tests, may be inappropriate. These methods will be the subject of the next review, along with a discussion of the relative merits of parametric and nonparametric approaches.

Finally, the methods presented here are restricted to the case where comparison is to be made between one or two groups. This is probably the most common situation in practice but it is by no means uncommon to want to explore differences in means across three or more groups, for example lung function in nonsmokers, current smokers and ex-smokers. This requires an alternative approach that is known as analysis of variance (ANOVA), and will be the subject of a future review.

This article is the fifth in an ongoing, educational review series on medical statistics in critical care. Previous articles have covered 'presenting and summarizing data', 'samples and populations', 'hypotheses testing and *P* values' and 'sample size calculations'. Future topics to be covered include comparison of proportions, simple regression and analysis of survival data, to name but a few. If there is a medical statistics topic you would like explained, contact us on editorial@ccforum.com.

Competing interests

None declared.

References

1. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M: **Early goal-directed therapy in the treatment of severe sepsis and septic shock.** *N Engl J Med* 2001, **345**:1368-1377.
2. Altman DG: *Practical Statistics for Medical Research*. London, UK: Chapman & Hall, 1991.
3. Kirkwood BR: *Essentials of Medical Statistics*. Oxford, UK: Blackwell Science Ltd, 1988.