

# Inference about means

## Methods for inference about means

**Statistical inference** is the process of drawing conclusions from data, for example by confidence intervals and significance tests. In this lecture we shall look how we can draw conclusions from samples about the means of populations.

We shall first look at large samples, and at how we can make inferences about a single mean, means in paired data, and the difference between the means of two samples. For each of these we shall use a large sample Normal method or z method.

We shall then look at the same problems for small samples. For a single mean we shall describe the one sample t method, for paired data the paired t method, and for the means of two samples the two sample t method, also called the independent samples t method, or two group t method. For t methods there are strong assumptions about the distribution of the observations. I shall describe how we can use graphical methods to investigate these.

We shall not discuss what to do if we have means of more than two samples. The usual method for any size samples is one-way analysis of variance (anova), the assumptions of which are as for the two sample t method.

## The mean of a large sample

We can find confidence intervals and carry out significance tests for the means of large samples using the Normal distribution. We make use of two properties of large samples. First, the means of large samples drawn in the same way will follow a Normal distribution quite closely, as described in Week 2. Second, the standard deviation estimated from a large sample will be close to that for the whole population. This means that the standard error estimated from the sample will be a good estimate.

We find confidence intervals for means of large samples using the Normal distribution. We first estimate the standard error of the mean of the sample. This is easy to do from the standard deviation of the observations, it is the standard deviation divided by the square root of the sample size. Then the 95% confidence interval is the mean minus 1.96 standard errors to the mean plus 1.96 standard errors.

For example, Figure 1 shows the distribution of birthweight in 1749 singleton pregnancies to Caucasian mothers in South London. This is clearly negatively skew, unlike the distribution of birthweight for term births, which is approximately Normal. These birthweights have mean = 3296.0 g and standard deviation = 563.2 g. The standard error of the mean is 13.5 g. Because the sample is large, the mean birthweight will be from a Normal distribution with mean equal to the mean birthweight in the population and standard deviation very close to the estimated standard error of the mean, 13.5 g. Hence the 95% confidence interval for the population mean birthweight will be  $3296.0 - 1.96 \times 13.5$  g to  $3296.0 + 1.96 \times 13.5$  g, which gives 3270 g to 3322 g. Hence we estimate that the mean birthweight in this population to be between 3270 and 3322 g.

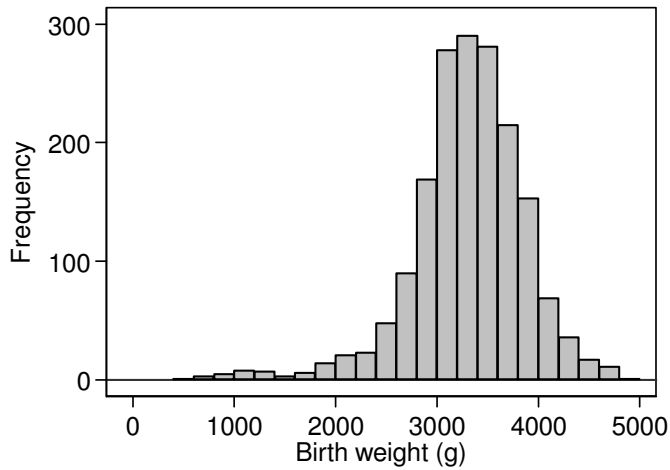


Figure 1. Birthweights of 1749 singleton births to Caucasian mothers in South London (data of Brooke *et al.*, 1989)

Table 1. Baseline depression score and fall after six weeks by treatment group for 525 patients with depression (Christensen *et al.*, 2004)

	Number	Baseline scores		Fall in scores	
		Mean	SD	Mean	SD
BluePages	165	21.1	10.4	3.9	9.1
MoodGYM	182	21.8	10.5	4.2	9.1
Controls	178	21.6	11.1	1.0	8.4

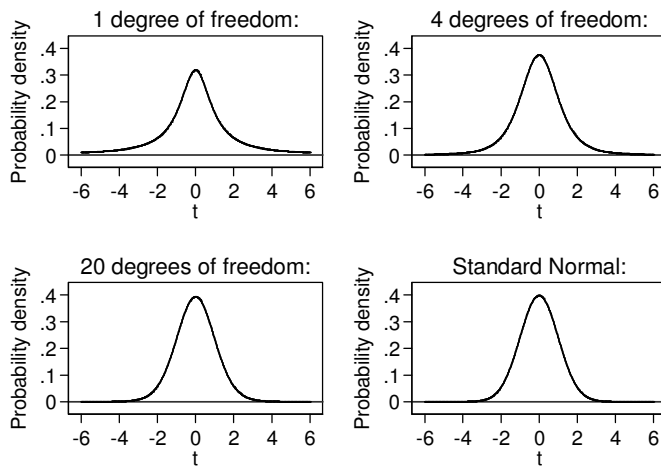


Figure 2. Student's t distribution with 1, 4, and 20 degrees of freedom, with the Standard Normal distribution

The only assumptions we have to make about the data for this method are:

- The observations are independent. We should not have, for example, a group of 100 observations where there are 10 subjects with 10 observations on each.
- The sample is large enough for the standard errors to be well estimated. My rule of thumb is 100 for one group.

## Differences between means in paired large samples

In healthcare research, we more often want to compare groups of subjects than use a single sample to estimate the mean in the population. For example, Christensen *et al.* (2004) compared interventions for depression delivered using the internet. They recruited 525 people with symptoms of depression identified in a survey. These were randomly allocated to a website, BluePages, offering information about depression ( $n = 166$ ) or a cognitive behaviour therapy website, MoodGYM, ( $n = 182$ ), or a control intervention using an attention placebo ( $n = 178$ ). The main outcome measure was the Center for Epidemiologic Studies depression scale. This consists of 20 questions scored 0 (not depressed) to 3 (depressed) and summed, giving a score between 0 and 60. The means and standard deviations of their depression scores before randomisation and of their falls in depression score are shown in Table 1. All the subjects were depressed and so should have initial scores well above zero and the standard deviations are approximately half the mean, so we can deduce that the depression scores had a positively skew distribution.

Ninety (17%) of subjects did not return post-intervention questionnaires and the authors assumed that their scores were unchanged. This means that whatever the original distribution of the scores, the differences must have a large spike of at least 90 observations at zero. They could not have a Normal distribution.

Because the three treatment groups are all fairly large samples, we can estimate a confidence interval for the mean fall for each of them. For BluePages, 165 subjects had mean fall in depression score = 3.9 with standard deviation = 9.1. The standard error of the mean is 0.71 and so the 95% confidence interval for the mean fall is  $3.9 - 1.96 \times 0.71$  to  $3.9 + 1.96 \times 0.71$ , which is 2.5 to 5.3 points on the depression scale. This is our interval estimate for the mean fall in depression score assuming non-responders do not change.

We can also test the null hypothesis that the mean change is zero against the alternative hypothesis that there is a change, in either direction. Because the sample is large, the mean will be from a Normal distribution with standard deviation equal to the standard error of the mean, 0.71. Hence the observed sample mean minus the unknown population mean then divided by the standard error will be an observation from the Standard Normal distribution. The null hypothesis says that the population mean is zero, so if this is true then sample mean over standard error will be from a Standard Normal distribution. For the BluePages group this is  $z = 3.9/0.71 = 5.49$ . The probability of getting such an extreme value from a Standard Normal distribution is very small indeed, 0.00000004. We cannot trust the approximation to the Normal distribution will be close enough for such very small probabilities to be reliable, so we usually quote this as  $P < 0.0001$ . This is the large sample Normal test for a single mean, also called the  $z$  test for a single mean.

The assumptions required for this method are:

- The observations are independent.
- The sample is large enough. My rule of thumb is 100 for one group.
- The mean and standard deviation of differences are constant, i.e. not related to the size of the variable.

We can check the last by plotting the difference against the average of the two measurements for the subject. I shall describe this in detail later under paired t test.

## Comparing the means of two independent large samples

We can also find a confidence interval for difference between the means of two independent samples. For example, we shall compare the mean fall in score for BluePages with MoodGYM. The difference between the means, BluePages minus MoodGYM, =  $-0.3$ . We can find the standard error for the difference by squaring the standard error of each mean, adding, and taking the square root. ***This only works when the groups are independent.*** If we were to do it for paired data like the before and after measurements above, the standard error might be much too large. For BluePages and MoodGYM, we have

$$\sqrt{0.71^2 + 0.67^2} = 0.98$$

The 95% CI is then given by  $-0.3 - 1.96 \times 0.98$  to  $-0.3 + 1.96 \times 0.98 = -2.2$  to  $+1.6$ .

We can also do a test of the null hypothesis that in the population the difference between the means is zero against the alternative hypothesis that the difference in the population is not zero. As for the paired example above, because we have a large sample the observed difference minus the population difference then divided by the estimated standard error of the difference should be an observation from a Standard Normal distribution. If the null hypothesis were true, the population difference would be zero. The test statistic is observed difference divided by its standard error,  $z = -0.3/0.98 = -0.31$ . The probability of an observation from the Standard Normal distribution being as far from its expected value, zero, as  $-0.31$  is  $P=0.76$ . Hence the difference is not significant. We can tell this from the 95% confidence interval, also, as this includes zero, the null hypothesis value for the difference. This is the large sample Normal distribution test or  $z$  test for the means of two independent groups.

We can carry out the same calculations for the comparison of each active intervention with control. For BluePages, the difference between mean falls is  $3.9 - 1.0 = 2.9$  and the standard error of the difference is 0.95. Hence the 95% confidence interval is  $2.9 - 1.96 \times 0.95$  to  $2.9 + 1.96 \times 0.95$  which gives 1.0 to 4.8. The test of significance gives  $z = 2.9/0.95 = 3.05$ ,  $P = 0.002$ . For MoodGYM minus control the observed difference is 3.2 with standard error 0.92, the 95% confidence interval being 1.4 to 5.0 and  $z = 3.48$ ,  $P = 0.0005$ .

If we want to test the overall null hypothesis that the three treatments would produce the same mean fall in the population, we could do this by applying the Bonferroni correction to these three P values. Multiplying by 3 would give the smallest P value =  $0.0005 \times 3 = 0.0015$ , which is still highly significant. Christensen *et al.* (2004) did not do the analysis exactly as we have here. They used an analysis of variance method, which I shall omit, to compare all three groups simultaneously.

The large sample Normal method for comparing two means requires two assumptions about the data.

- The observations and groups are independent. We should not have links between observations in the two groups, such as a matched study where each subject in one group is matched, e.g. by age and sex, with a subject in the other group.
- The samples are large enough for the standard errors to be well estimated and for the means to be observations from Normal distributions. My rule of thumb is that for a single sample there should be at least 100 observations and for two samples at least 50 in each.

Some computer programs do not do large sample  $z$  tests directly. You have to use the command for a one sample or paired  $t$  test, or for a two-sample  $t$  test with unequal variances. I describe these below. For large samples, they give the same answers as the  $z$  tests.

## The $t$ distribution

When samples are small, we cannot apply the large sample Normal distribution methods safely. This problem was tackled by a statistician who published under the pseudonym Student, because his employers would not allow him to publish the results of his work. The probability distribution which he discovered is known as Student's  $t$  distribution as a result and the methods which use it as Student's  $t$  tests.

We have seen that when the sample is large, the observed sample mean minus the population mean divided by the standard error follows the Standard Normal distribution. When the sample is small this is not so. The distribution followed depends on the distribution of the observations themselves, unlike the large sample case where this is irrelevant. We have to assume that the data themselves come from a population which follows a Normal distribution. We have seen that some naturally occurring variables do this and some do not. We shall see in Week 5 that many variables which do not follow a Normal distribution can be made to do so by changing the way in which we look at them, using a transformation such as the logarithm. When the observations come from a population which follows a Normal distribution, then the sample mean minus the population mean divided by the standard error of the mean follows **Student's  $t$  distribution**, or simply the  **$t$  distribution**. Student's  $t$  distribution may be defined as the distribution which this ratio would follow.

Like the Normal distribution, Student's  $t$  distribution is a family of distributions rather than just one. This family has only has one parameter, the number which tells us with which member of the family of  $t$  distributions we are dealing. This is called the **degrees of freedom**. We have already used this term in the calculation of variances and standard deviations. The degrees of freedom of the  $t$  distribution is equal to the degrees of freedom of the standard deviation used in the calculation of the standard error.

Table 2. Two tailed probability points of the t Distribution

D.f.	Probability				D.f.	Probability			
	(10%)	(5%)	(1%)	(0.1%)		(10%)	(5%)	(1%)	(0.1%)
1	6.31	12.70	63.66	636.62	16	1.75	2.12	2.92	4.02
2	2.92	4.30	9.93	31.60	17	1.74	2.11	2.90	3.97
3	2.35	3.18	5.84	12.92	18	1.73	2.10	2.88	3.92
4	2.13	2.78	4.60	8.61	19	1.73	2.09	2.86	3.88
5	2.02	2.57	4.03	6.87	20	1.73	2.09	2.85	3.85
6	1.94	2.45	3.71	5.96	21	1.72	2.08	2.83	3.82
7	1.90	2.36	3.50	5.41	22	1.72	2.07	2.82	3.79
8	1.86	2.31	3.36	5.04	23	1.71	2.07	2.81	3.77
9	1.83	2.26	3.25	4.78	24	1.71	2.06	2.80	3.75
10	1.81	2.23	3.17	4.59	25	1.71	2.06	2.79	3.73
11	1.80	2.20	3.11	4.44	30	1.70	2.04	2.75	3.65
12	1.78	2.18	3.06	4.32	40	1.68	2.02	2.70	3.55
13	1.77	2.16	3.01	4.22	60	1.67	2.00	2.66	3.46
14	1.76	2.15	2.98	4.14	120	1.66	1.98	2.62	3.37
15	1.75	2.13	2.95	4.07	∞	1.65	1.96	2.58	3.29

D.f. = Degrees of freedom

∞ = infinity, same as the Standard Normal Distribution

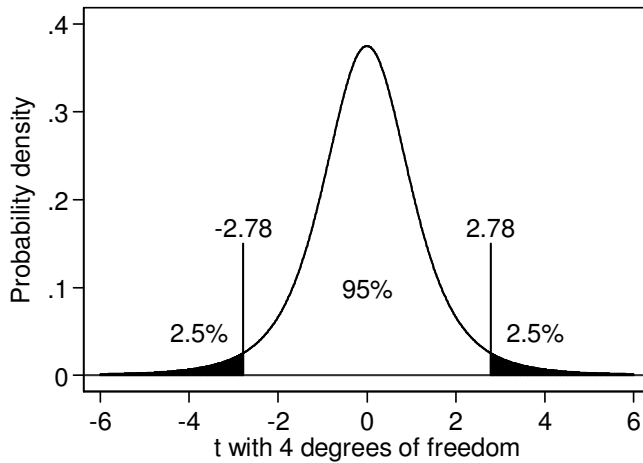


Figure 3. 5% probability points of the t distribution with 4 degrees of freedom

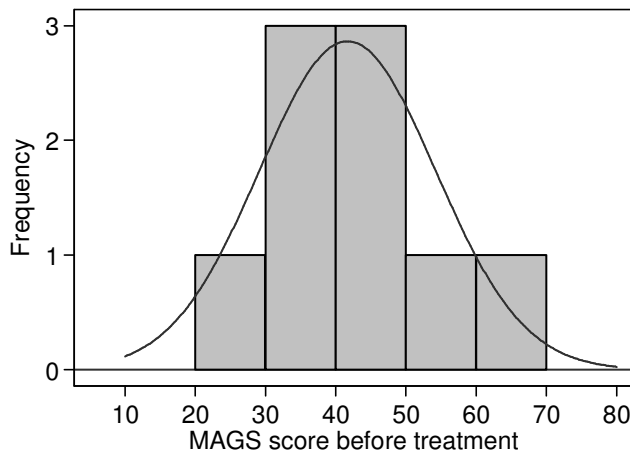


Figure 4. Histogram for the MAGS score before treatment, with Normal distribution of same mean and variance

Figure 2 shows some members of the Student's t distribution family. When the degrees of freedom are small, corresponding to small samples, the t distribution has much longer tails than the Normal. This reflects the greater uncertainty in the standard error of the mean. As the degrees of freedom and hence the related sample size gets bigger, the t distribution gets closer and closer to the Standard Normal distribution. The t distribution reaches the Normal distribution in theory when the sample is infinitely large. In practice, it is difficult to tell the Normal and t distributions apart at about 30 degrees of freedom.

Like the Normal, the t distribution has no simple formulae for its probabilities. Instead we used numerical approximations to calculate the number which replaces 1.96 in confidence interval calculations and the P values in significance tests. If we do these calculations using one of the many computer programs available, the program will calculate these for us. For the purposes of illustration, I shall also give a short table of the distribution for different degrees of freedom (Table 2). For each of the degrees of freedom given, this gives the value which will be exceeded, in either positive or negative direction, with the given probability. For example, Figure 3 shows the 5% two sided probability points of the t distribution with 4 degrees of freedom.

We can use Student's t distribution to replace the Normal distribution in confidence interval and significance tests for small samples. To do this we must be able to assume that the observations themselves come from a Normal distribution, plus other assumptions for different applications as described below.

### **The one sample t method**

We can use the t distribution to carry out all the analyses of means of small samples which we did above using the Normal distribution for large samples. We seldom want to estimate the mean of a population from the mean of a small sample, but we shall start with this as it is the easiest.

For our example, we shall use data from nine patients with chronic non-healing wounds (Shukla *et al.*, 2004). Biopsies were assessed using the microscopic angiogenesis grading system (MAGS) score, which provides an index of how well small blood vessels are developing and hence of epithelial regeneration. High scores are good. The nine observations were 20, 31, 34, 39, 43, 45, 49, 51, and 63.

We can use these measurements to estimate the mean MAGS score in non-healing patients. The mean score before treatment is 41.7 and the standard deviation is 12.5 with 8 degrees of freedom. The standard error of the mean is 4.2. If we had a large sample, we could estimate a 95% confidence interval for the mean by subtracting and adding 1.96 standard errors:  $41.7 - 1.96 \times 4.2$  to  $41.7 + 1.96 \times 4.2$ . But we have only 9 observations, so this would not be valid. Instead we use the t distribution with 8 degrees of freedom. From Table 2, the 5% point of the t distribution with 8 degrees of freedom is 2.31, so the confidence interval for the mean MAGS score is  $41.7 - 2.31 \times 4.2$  to  $41.7 + 2.31 \times 4.2 = 32.0$  to 51.4.

This is only valid provided we can assume the observations come from a Normal distribution. We may know from our experience of the measurement that this variable usually follows a Normal distribution, but we always like to check that our sample is compatible. I describe how to do this in the next section.

Table 3. MAGS score before and after treatment with topical placental extract in 9 patients with non-healing wounds (Shukla *et al.*, 2004)

MAGS score before	MAGS score after	Difference, MAGS before minus MAGS after	Average of MAGS before and MAGS after
20	32	12	26.0
31	47	16	39.0
34	43	9	38.5
39	43	4	41.0
43	55	12	49.0
45	52	7	48.5
49	61	12	55.0
51	55	4	53.0
63	71	8	67.0

### Checking the assumption of a Normal distribution

When I introduced the Normal distribution, I showed histograms of several large samples and superimposed Normal distribution curves on them to show whether the Normal distribution fitted the data. For small samples, it can be difficult to judge from a histogram whether the Normal distribution is a good fit. Figure 4 shows a histogram for the MAGS score before treatment. We cannot really say whether the distribution and the data have the same shape.

There is a better graphical method to examine the fit of a Normal distribution to a set of data, the Normal quantile plot or Normal plot for short. We shall not cover this in the IAPT course, but should you need to do statistical analysis, you can find details in my other statistics courses.

### The paired t method

The paired t method is the version of the one sample t method usually seen in research publications. Here we have paired observations, such as the same subject before and after an intervention, the same subject receiving two different interventions as in a cross-over trial, or matched case and control in a case-control study. Table 3 shows fuller data from Shukla *et al.* (2004). In this trial, patients with chronic non-healing wounds were randomised to receive topical placental extract or to control. The data in Table 3 show the MAGS score before and after treatment in a group of 9 patients in the active treatment group. We want to know whether we have evidence that mean MAGS score changed and what the average score might be. I have calculated the difference between the MAGS score after treatment and the MAGS score before treatment, i.e. the increase in the MAGS score.

The authors of the paper did not do any further analysis of these data, as they were all positive differences and the MAGS score clearly increases following treatment. We shall use them to estimate the mean increase in MAGS score. The mean and standard deviation of the increase in MAGS score are 9.33 and 4.03 respectively. We have 9 observations so the number of degrees of freedom for the calculation of the standard deviation is  $9 - 1 = 8$ .



The standard error of the mean difference is 1.34. To estimate the 95% confidence interval for the mean from this small sample, we use the 5% point of the t distribution with 8 degrees of freedom. From the 8 degrees of freedom row in Table 2 this is 2.31. The 95% confidence interval is therefore the mean minus or plus 2.31 standard errors,  $9.33 - 2.31 \times 1.34$  to  $9.33 + 2.31 \times 1.34$ , which gives us 6.2 to 12.4.

We can also test the null hypothesis that in the population the mean increase is zero. The test statistic is the mean divided by its standard error. This is  $9.33/1.34 = 6.96$ . If we look in the 8 degrees of freedom row in Table 2, we see that this is larger than the largest number there, 5.04, which corresponds to a probability of 0.001. Hence we could say  $P < 0.001$ . In practice, we would do this using a computer program, which gives us  $P = 0.0001$ . The difference is highly significant.

There are several assumptions which we must make about the data for the paired t method test to be valid:

- the observations must be independent, apart from the pairing,
- the differences must follow a Normal distribution,
- the mean and standard deviation of the differences must be constant, i.e. not related to the size of the measurement.

The first of these, independence, depends on the design. It is met for the MAGS data, because the pairs of data come from nine different subjects. The second can be tested graphically. The third, that the mean and the variability are not related to the magnitude, can also be investigated graphically. We do a scatter plot of the difference against the average of the two observations, as in Figure 5. We do this because the average of the two measurements is the best estimate we have of the subject's true MAGS score over the period. Using only one of the measurements, either before or after, on the horizontal axis tends to produce spurious relationships between difference and magnitude. For the MAGS data, Figure 5 shows little evidence that either the mean difference or the variability of the differences is related to the magnitude of MAGS score for the subject.

### **The two sample t method**

This is also called the unpaired t method or unpaired t test, the two group t method, or Student's two sample t test. It enables us to estimate the difference between means or test the null hypothesis of no difference in the population, even when the samples are small.

Our example is a comparison of capillary density between patients with diabetic foot ulcers and a group of non-ulcerated controls (Table 4). The data are shown graphically in Figure 6. The samples are small, only 23 ulcer patients and 19 controls, so we cannot use the large sample Normal method. The standard error will not be sufficiently well estimated.

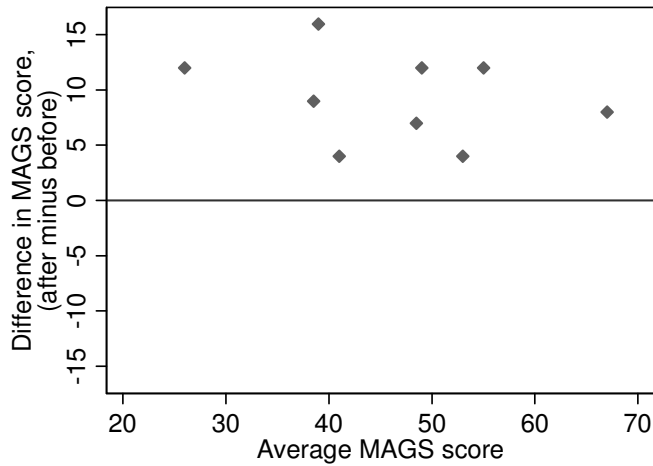


Figure 5. Difference versus mean plot for the increases in MAGS score.

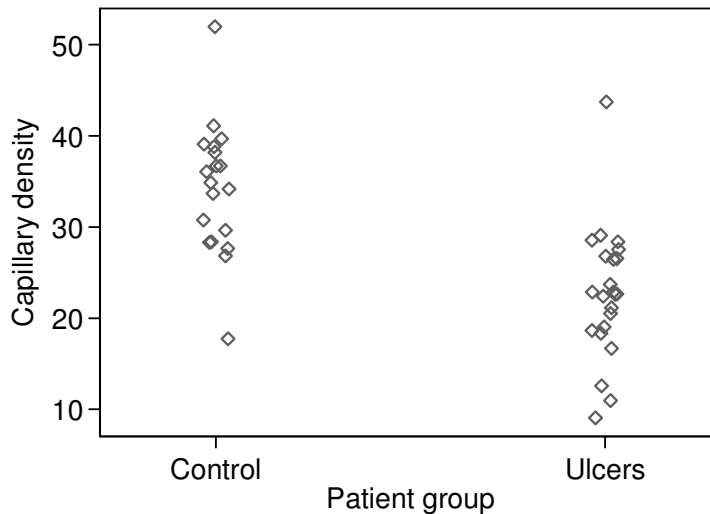


Figure 6. Scatter diagram showing capillary density in two groups of patients

For the two-sample t method, we must make three assumptions about the data:

- The observations and groups are independent.
- The observations come from Normal distributions,
- The distributions in the two populations have the same variance. (N.B. The populations, not the samples from them, have the same variance.)

If the distributions in the two populations have the same variance, we need only one estimate of variance. We call this the **common** or **pooled** variance estimate. It is a weighted average of the two sample variances, weighted by the degrees of freedom. The degrees of freedom for this common variance estimate are the number of observations minus 2. We then use this common estimate of variance to estimate the standard error of the difference between the means.

Table 4. Capillary density (per mm<sup>2</sup>) in the feet of ulcerated patients and a healthy control group (data supplied by Marc Lamah)

	Controls	Ulcerated patients
	17.5	9.0
	27.5	11.0
	27.0	12.5
	29.5	18.0
	27.0	18.0
	29.0	18.0
	34.5	18.5
	31.0	20.0
	35.5	20.0
	33.5	22.0
	35.5	22.5
	34.0	22.5
	36.5	23.0
	38.0	23.0
	40.0	24.0
	39.5	26.5
	40.0	26.5
	40.0	27.0
	52.0	27.5
		28.0
		28.5
		29.0
		44.5
Number	19	23
Mean	34.08	22.59
SD	7.29	7.31

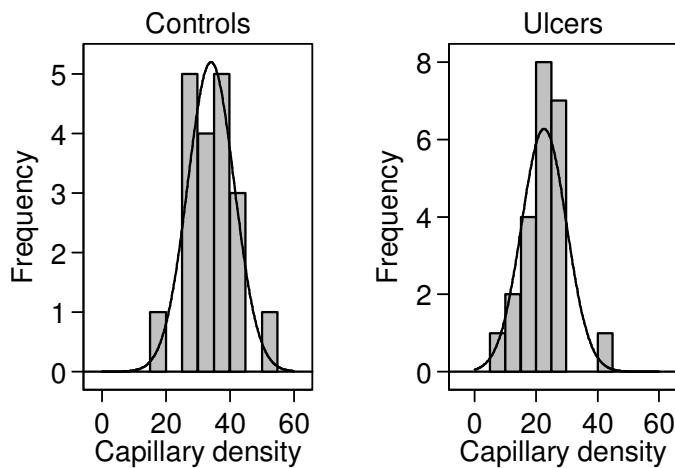


Figure 7. Histograms of capillary density in two groups of patients

For the capillary density example, the common variance = 53.31, SD = 7.30 capillaries/mm<sup>2</sup>, df = 19 + 23 - 2 = 40. The difference (control - ulcer) = 34.08 - 22.59 = 11.49 capillaries/mm<sup>2</sup>. The standard error of the difference = 2.26 capillaries/mm<sup>2</sup>. Then the 95% confidence interval for difference is given by 11.49 -  $t \times 2.26$  to 11.49 +  $t \times 2.26$ . Here  $t$  comes not from the Normal distribution but the  $t$  distribution with 40 degrees of freedom. It is the 5% point of the distribution, because 5% of observations will be further from zero than  $t$ , 95% will be closer to zero than  $t$ . From Table 2, for 40 degrees of freedom,  $t = 2.02$ . Hence the 95% CI is 11.49 - 2.02  $\times$  2.26 to 11.49 + 2.02  $\times$  2.26 = 6.92 to 16.07 capillaries/mm<sup>2</sup>.

We can also carry out a test of significance, testing the null hypothesis that in the population the difference between means = 0. We take the observed difference divided by its standard error and, if the null hypothesis were true, this would be an observation from the  $t$  distribution with 40 degrees of freedom. We have

$$\text{difference/SE} = 11.49/2.26 = 5.08.$$

From Table 2, the probability of such an extreme value is less than 0.001. If we use a good computer program, this will calculate the P value for us more accurately. In this case we get P = 0.0000, which we report as P < 0.0001.

We can check the assumption that capillary density follows a Normal distribution in each population by histograms. Figure 7 shows histograms for each group. There are not enough observations to judge whether the data follow Normal distributions. We can improve matters by combining the two groups. The distribution would be affected by any difference between the means, perhaps even becoming bimodal. We get round this by subtracting the group mean from each observation to give **residuals**. The residuals have mean = 0 in each group. We can then put them together to form a single distribution, as shown in Figure 8. This looks fairly symmetrical, but there are still only a few observations.

The other assumption is that the variances are the same in each population. For the capillary density, Table 4 shows that the standard deviations are similar, being 7.29 capillaries/mm<sup>2</sup> for the control sample and 7.31 capillaries/mm<sup>2</sup> for the ulcerated sample.

We can also test the equality of variances, either with an F test or Levene's test. However, tests have the unfortunately property that they miss large differences for small samples, when differences might matter, and find them for large samples, when they matter much less. It is usually preferable to judge whether the assumption of uniform variance is plausible from the scatter plot (Figure 6). For the capillary density, the spread of the two groups looks fairly similar.

Methods using the  $t$  distribution depend on some strong assumptions about the distributions from which the data come. In general, for two equal sized samples the  $t$  method is very resistant to deviations from Normality, though as the samples become less equal in size the approximation becomes less good. The most likely effect of skewness is that we lose power. P values are then too large and confidence intervals too wide.

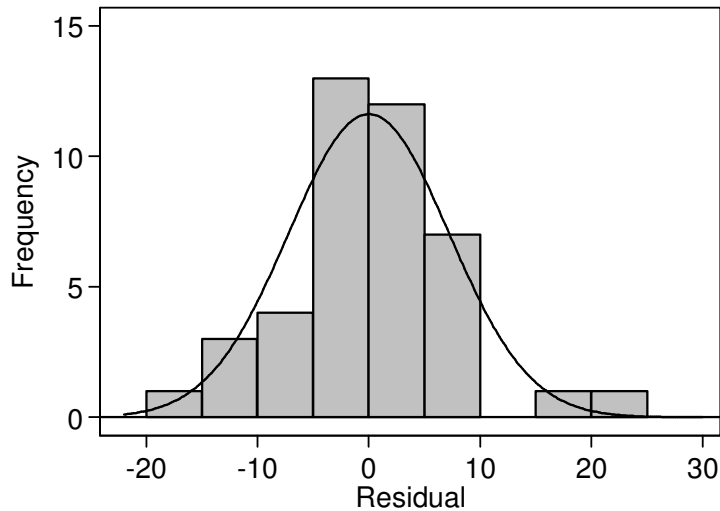


Figure 8. Distribution of residual capillary density, with corresponding Normal distribution curve

### References

Brooke OG, Anderson HR, Bland JM, Peacock JL, Stewart CM. (1989) Effects on birth weight of smoking, alcohol, caffeine, socioeconomic factors, and psychosocial stress. *British Medical Journal*, **298**, 795-801.

Christensen H, Griffiths KM, Jorm AF. (2004) Delivering interventions for depression by using the internet: randomised controlled trial. *British Medical Journal* **328**, 265-268.

Prentice AM, Black AE, Coward WA, Davies HL, Goldberg GR, Murgatroyd PR, Shukla VK, Rasheed MA, Kumar M, Gupta SK, Pandey SS. (2004) A trial to determine the role of placental extract in the treatment of chronic non-healing wounds. *Journal of Wound Care* **13**, 177-9.

Steenmoorle P, Julina GN. (2004) Can laboratory investigation help us to decide when to discontinue larval therapy? *Journal of Wound Care* **13**, 38-40.