Health Sciences M.Sc. Programme

Applied Biostatistics

Week 6: Comparing means

Comparing the means of large samples using the Normal Distribution

In many studies we are much more interested in the difference between two means than in their absolute value. This is usually straightforward if the means are estimated from two independent samples. It can be more difficult if the samples are matched or are the same.

When samples are large we can assume that sample means are observations from a Normal Distribution, and that the calculated standard errors are good estimates of the standard deviations of these Normal Distributions. We can use this to find confidence intervals.

For an example, in a study of respiratory symptoms in school children, we wanted to know whether children reported by their parents to have respiratory symptoms had worse lung function than children who were not reported to have symptoms. 92 children were reported to have cough during the day or at night, and their mean PEFR was 313.6 litre/min with standard deviation 55.2 litre/min. We thus have two large samples, and can apply the large sample Normal method or z method.

The difference between the two groups = 294.3 – 313.6 = –18.8. The standard error of the difference = 6.11 litre/min. (We omit the details of calculations and will use a computer program to do them.)

Because the sample is large, the difference between the means can be assumed to come from a Normal Distribution, and the estimated standard error to be a good estimate of the standard deviation of this distribution. The 95% confidence limits for the difference are thus $18.8 - 1.96 \times 6.11$ and $18.8 + 1.96 \times 6.11$, i.e. 6.8 and 30.8 1/min. This confidence interval does not include zero, so we have good evidence that there is a difference between mean lung function in these two groups of children. The difference itself is not very well estimated, however. It could be anything from 7 to 31 litre/min lower in children with the symptoms.

We can also carry out a significance test directly. The null hypothesis is that the means in the two populations are equal and the alternative hypothesis is that they are different. There are two assumptions: that the observations and groups are independent and that the samples are large enough for the standard errors to be well estimated. My rule of thumb is at least 50 in each group. Independence means that we should not have, for example, a group of 100 observations where there are 10 subjects with 10 observations on each. 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. To find a test statistic, we take the observed difference divided by its standard error. This will be from a Normal distribution, because the sample is large, with standard deviation 1.0, because we have divided by the standard error, and will have mean zero if the null hypothesis is true. Hence is would be from a Standard Normal distribution if the null hypothesis were true. The test statistic is –18.8/6.11 = –3.1. P = 0.002. If the null hypothesis were true, the data which we have observed would be unlikely. We can conclude that there is good evidence that children reported to have cough during the day or at night have lower PEFR than other children.

Table 1. 24 hour energy expenditure (MJ) in groups of lean and obese women (Prentice *et al.*, 1986, quoted in Altman 1991)

	Lean		Obese	
	6.13	8.08	8.79	9.97
	7.05	8.09	9.19	11.51
	7.48	8.11	9.21	11.85
	7.48	8.40	9.68	12.79
	7.53	10.15	9.69	
	7.58	10.88		
	7.90			
Number	13		9	
Mean	8.066		10.298	
Standard deviation	1.238		1.398	

In this case, we have two ways of interpreting the same calculation: as a confidence interval estimate or as a significance test. The confidence interval is usually superior, because we not only demonstrate the existence of a difference but also have some idea of its size. This is of particular value when the difference is not significant. For example, in the same study only 27 children were reported to have phlegm during the day or at night. These had mean PEFR of 298.0 litre/min and standard deviation 53.9 litre/min, hence a standard error for the mean of 10.4 litre/min. This is greater than the standard error for the mean for those with cough, because the sample size is smaller. The 1708 children not reported to have this symptom had mean 312.6 litre/min and standard deviation 55.4 litre/min, giving standard error 1.3 litre/min. Hence the difference between the means was –14.6, with standard error = 10.5. The test statistic is –14.6/10.5 = –1.4. This has a probability P = 0.16, and so the data are consistent with the null hypothesis. However, the 95% confidence interval for the difference is $-14.6 - 1.96 \times 10.5$ to $-14.6 + 1.96 \times 10.5 = -35$ to +6 litre/min. We see that the difference could be just as great as for cough. Because the size of the smaller sample is not so great, the test is less likely to detect a difference for the phlegm comparison than for the cough comparison.

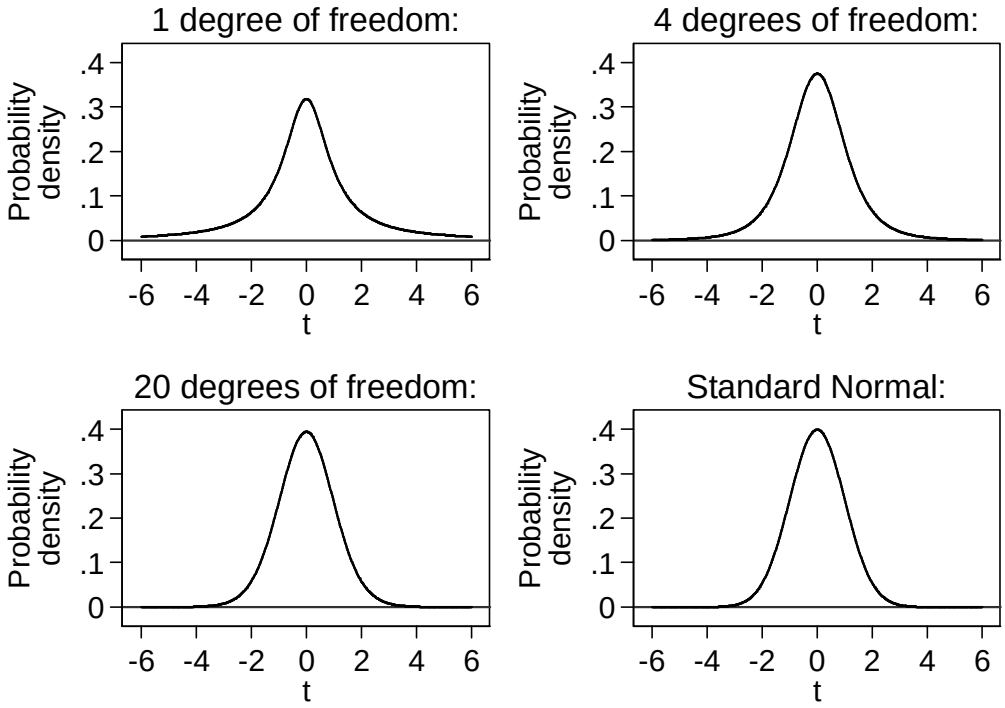
The two sample t method

Table 1 shows the 24 hour energy expenditure in two groups of women, classified as lean or obese. We shall estimate the difference between mean energy expenditure in the two populations. The samples are small and we cannot use the large sample Normal method. The standard error may not be sufficiently well estimated. The distribution of the standard error estimate depends on the distribution of the observations themselves and to allow for it we must make two assumptions about the data:

1. the observations come from Normal distributions,
2. the distributions in the two populations have the same variance. (N.B. The populations, not the samples from them, have the same variance.)

In addition, as before we must assume that the observations are independent.

Figure 1. Some members of the Student's t distribution family, showing the convergence to the Standard Normal distribution



As the two populations are assumed to have the same variance, we only need one variance estimate, which we call the common variance. To estimate this, we find the sum of squares about the mean and the degrees of freedom for each group, add them to get the total sum of squares about the group means and the total degrees of freedom and divide one by the other. The sums of squares about the two sample means are 18.394 and 15.632. This gives the combined sum of squares about the sample means to be 18.394 + 15.632 = 34.026. The combined degrees of freedom are = 13 + 9 – 2 = 20. Hence the common variance estimate = 34.026/20 = 1.701 and the standard deviation is 1.304 MJ. The standard error of the difference between means is calculated using this and is = 0.566 MJ. The difference in mean (obese – lean) = 10.298 – 8.066 = 2.232 MJ.

If we had a large sample, we could find a 95% confidence interval for the difference by $2.232 - 1.96 \times 0.566$ to $2.232 + 1.96 \times 0.566$. But this would not take into account that the standard error may not be well estimated and so may be two small. This problem was solved by W G Gossett, who wrote under the pseudonym Student, as Student's t distribution. Student's t distribution is a family of distributions, like the Normal, and has one parameter. This is called the degrees of freedom and we find the correct member of this family from the degrees of freedom used to calculate the standard error. Figure 1 shows some members of the Student's t distribution family. As the degrees of freedom increases, it gets more like the Standard Normal distribution, as we would expect.

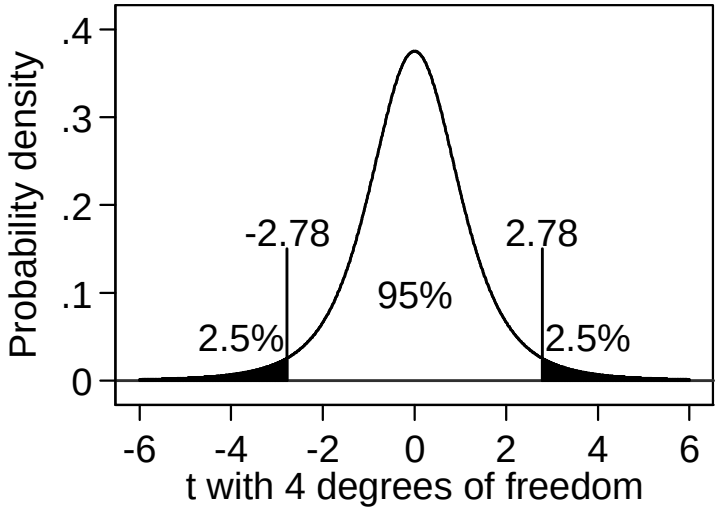
Like the Normal distribution, the probabilities from the t distribution cannot be calculated exactly. Table 2 shows some probability points from the t distribution for selected degrees of freedom. These are values such that the tabulated value is exceeded, in either direction, with the probability at the top of the column. Figure 2 shows the probability points for the t distribution with 4 degrees of freedom.

Table 2. Probability points of the t distribution

Degrees of freedom	Probability				Degrees of freedom	Probability			
	0.10	0.05	0.001	0.01		0.10	0.05	0.001	0.01
	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	4.01
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	1.72	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	1.89	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	3.05	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	2.14	2.98	4.14	120	1.66	1.98	2.62	3.37
15	1.75	2.13	2.95	4.07	∞	1.64	1.96	2.58	3.29

∞ = infinity, same as the Standard Normal distribution

Figure 2. 5% probability points for the t distribution with 4 degrees of freedom.



The value of the t Distribution for the 95% confidence interval is found from the 0.05 column and 20 d.f. row of Table 2 to be 2.09. Hence the 95% confidence interval is $2.232 - 2.09 \times 0.566$ to $2.232 + 2.09 \times 0.566 = 1.05$ to 3.42 litre/min. Hence the difference between the energy expenditure in obese women in this population is estimated to be between 1.05 and 3.42 MJ.

Figure 3. Scatter plot of the data of Table 1

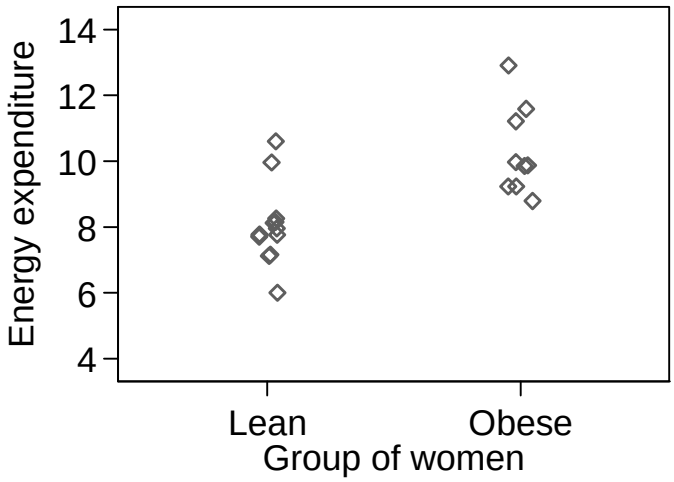
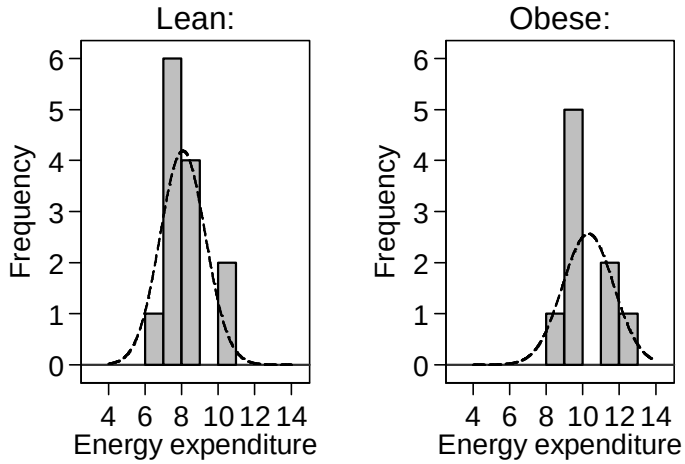


Figure 4. Histograms for the two groups of Table 1



To test the null hypothesis that the obese-lean difference is zero, the test statistic is difference over standard error: 2.232/0.566 =3.943. If the null hypothesis were true, this would be an observation from the t Distribution with 20 degrees of freedom. From Table 2, the probability of such an extreme value is less than 0.001. Computer calculation gives P=0.0008. Hence the data are not consistent with the null hypothesis and we can conclude that we have good evidence that energy expenditure is different for lean and obese women in this population.

All this is only valid if three assumptions hold: that the observations are independent, which they are, that the distribution of energy expenditure follows a Normal distribution in each population and that the variances are the same in each population. We can check the second and third graphically. Figure 3 shows a scatter plot of the data. The variability looks quite uniform, though some skewness is apparent. We can also check the distribution by histograms, as in Figure 4. The samples are two small to judge the shape of the distribution from these histograms. We can improve things a bit by putting both histograms onto the same graph. To do this we calculate the **within group residuals,** found by subtracting the group mean from each observation. Each group then has mean zero and if the assumption of uniform variance is true should have the same distribution. We can combine them and plot one histogram, as in Figure 5.

Figure 5 Histogram of residuals for Table 1

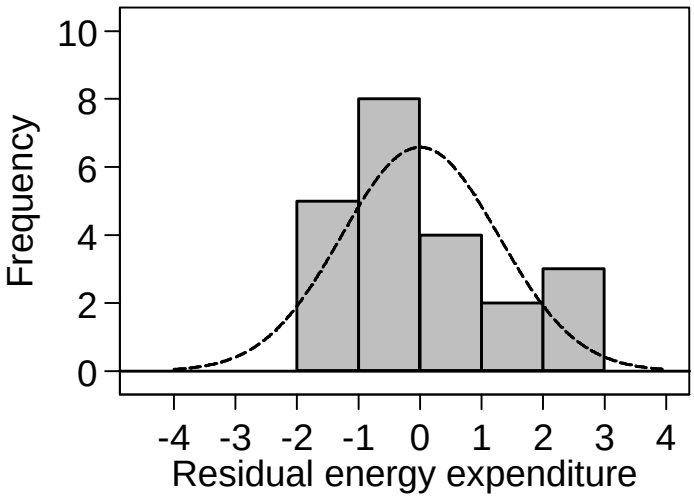
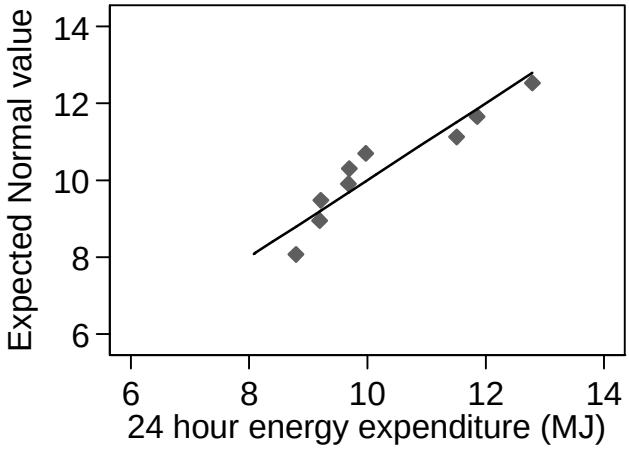


Figure 6. Normal plot for the energy expenditure in 9 obese women



This definitely looks skew.

Normal plot

Another way to check the approximation to a Normal distribution is the **Normal plot**, or **Normal quantile plot** (Q-Q plot in SPSS). I shall illustrate this first using the obese women only. We first sort the data into ascending order:

8.79 9.19 9.21 9.68 9.69 9.97 11.51 11.85 12.79

What would we expect the value of the first, second, third, etc. observations to be if they were from a Normal distribution? It is possible to calculate what the average values for the 1st, 2nd, 3rd, etc., observations of a Standard Normal sample with 9 observations are:

–1.28 –0.84 –0.52 –0.25 0.00 0.25 0.52 0.84 1.28

(As usual, the computer program is going to all this for us.) To get the expected values for a Normal sample with the same mean and standard deviation as energy expenditure, we multiply by standard deviation = 1.398 and add mean = 10.298:

8.51 9.12 9.57 9.95 10.30 10.65 11.02 11.47 12.09

Figure 7. Normal plot for the residuals for Table 1

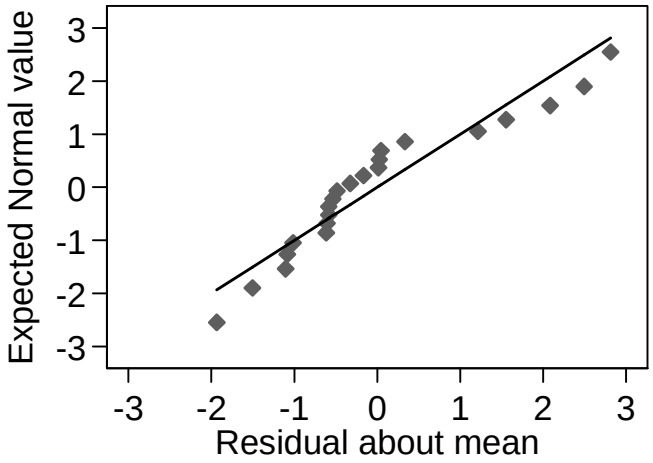
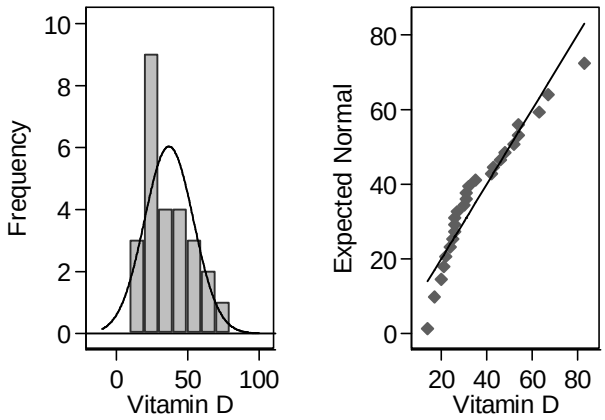


Figure 8. Histogram and Normal plot for Vitamin D levels measured in the blood of 26 healthy men

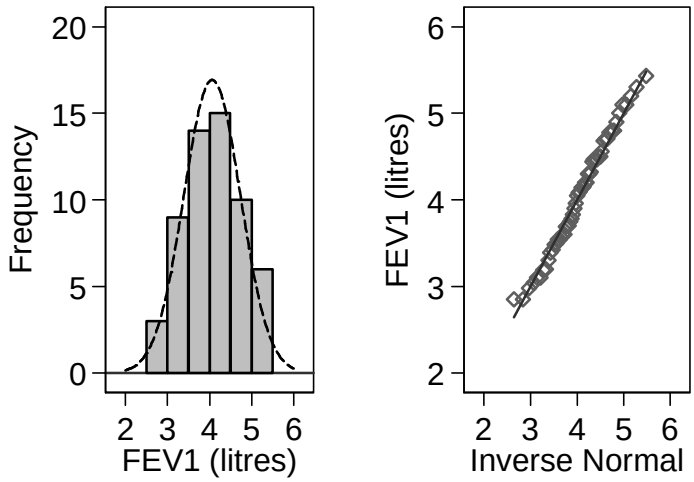


E.g. –1.28×1.398 + 10.298 = 8.51. We then plot the observed values against those we expect if the data followed a Normal distribution, as in Figure 6. The straight line on the graph is the line of equality, where the observed energy expenditure and that expected for a Normal distribution would be identical. In this case, the points do not lie close to this straight line, but appear to have a curve. This indicates skewness.

We can do this for all the data, using the residuals, as in Figure 7. The curvature and hence the skewness is apparent.

There are several variations on the Normal plot. Sometimes we use a Standard Normal distribution rather than one with mean and standard deviation equal to those of the variable. Sometimes we plot the expected Normal value on the horizontal axis and the observed value on the vertical (Stata does this). In this case, an upwards curvature indicates positive skewness.

Figure 9. Histogram and Normal plot for FEV1 for 57 male medical students



Skew distributions produce a clear bend or curve in Normal plot, distributions close to the Normal produce a straight line. Figure 8 shows a Normal plot for some observations of vitsmin D concentrations in blood, which have a distribution which is clearly positively skew. Figure 9 shows a Normal plot for the FEV1 data, which have a distribution close to the Normal and produce a straight line.

Effect of deviations from assumptions in the two sample t method

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 two-sample 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. This means that P values are too large and confidence intervals too wide.

We can often correct skewness by analysing a mathematical function of the data, rather than the observations themselves. We call this a transformation. The main ones used are the logarithm, the square root, and the reciprocal. Another approach is to use a method which does not require these assumptions about the data, such as the Mann Whitney U test. This only gives us a P value, not a confidence interval, unless we can make assumptions almost as strong as for the two sample t method.

If we cannot assume uniform variance, the effect is usually small if the two populations are from a Normal Distribution. Unequal variance is often associated with skewness in the data, in which case a transformation designed to correct one fault often tends to correct the other as well.

If distributions are Normal, can use the Satterthwaite correction to the degrees of freedom. If variances are unequal, we cannot estimate a common variance. Instead we use the large sample form of the standard error of the difference between means. We replace the t value for confidence intervals by t with fewer degrees of freedom. This adjusted degrees of freedom depends on the relative sizes of the variances. The larger variance dominates and if one is much larger than the other the degrees of freedom for that group are the only degrees of freedom.

For the energy expenditure example: degrees of freedom = 20 (= 13 + 9 – 2). Satterthwaite's degrees of freedom = 15.9187. We round this down to 15 to use the t table. For this example, equal variances gives 95% CI = 1.05 to 3.42 MJ, P=0.0008, unequal variances gives 95% CI = 1.00 to 3.46 MJ. P=0.0014. Satterthwaite’s method is an approximation for use in

unusual circumstances. The equal variance method is the standard t test. SPSS always gives both the standard and the Satterthwaite method.

For large samples, we can ignore the assumptions about the distribution and the Satterthwaite method becomes the large sample comparison of two means. SPSS does not have a program for the large sample comparison, we must use the t test as an approximation.

Table 3. PEFR (litre/min) measured by Wright meter and mini meter, female subjects

Subject	Wright PEFR	Mini PEFR	Difference
1	490	525	–35
2	397	415	–18
3	512	508	4
4	401	444	–43
5	470	500	–30
6	415	460	–45
7	431	390	41
8	429	432	–3
9	420	420	0
10	275	227	48
11	165	268	–103
12	421	443	–22
Mean			–17.2
Standard deviation			40.3
Standard error of mean			11.6

The one sample t method

We can use the t Distribution to find confidence intervals for means estimated from a small sample from a Normal Distribution. We do not usually have small samples in sample surveys, but we often do find them in clinical studies. For example, we can use the t Distribution to find confidence intervals for the size of difference between measurements obtained from subjects under two conditions.

Consider the data of Table 3. These are results from a comparison of two instruments for measuring PEFR, a Wright Peak Flow Meter and a Mini Peak Flow Meter. The subjects were family and colleagues, and so not a random sample. Each gave two readings on each instrument in random order. Table 3 shows the second reading on each. We shall measure the amount of bias between the instruments, the amount by which one tends to read above the other. The first step is to find the differences (Wright – mini). We then find the mean difference and its standard error in the usual way.

Figure 10. Histogram and Normal plot of differences for the data of Table 3

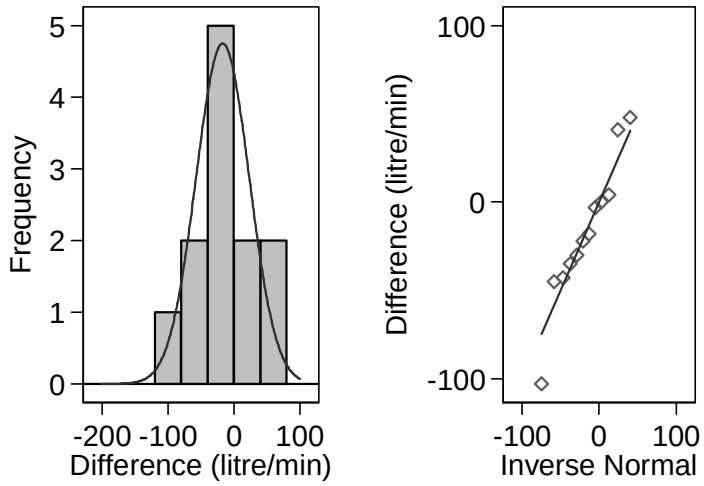
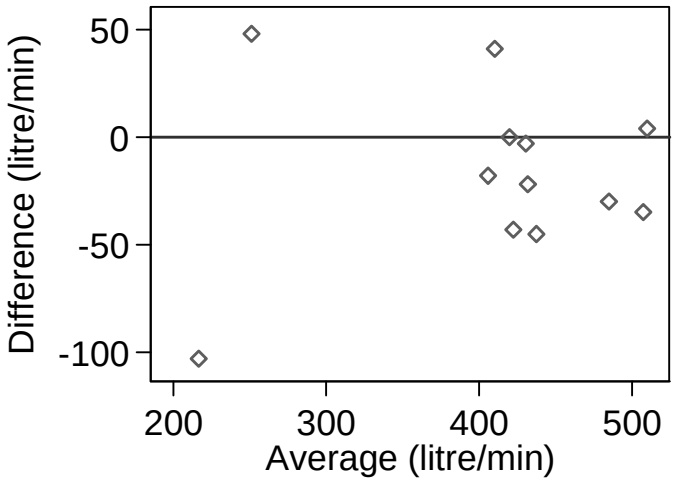


Figure 11. Plot of difference against mean for the data of Table 3



To find the 95% confidence interval for the mean difference we must suppose that the differences follow a Normal Distribution. To calculate the interval, we first require the relevant point of the t Distribution from Table 2. There are 12 differences and hence 12 – 1 = 11 degrees of freedom associated with the variance. We want a probability of 0.95 of being closer to zero than *t*, so we go to Table 2 with probability = 1 – 0.95 = 0.05. Using the 11 d.f. row, we get $t = 2.20$. Hence the difference between a sample mean and the population mean is less than 2.20 standard errors with probability 0.95, and the 95% confidence interval is $-17.2 - 2.20 \times 11.6$ to $-17.2 + 2.20 \times 11.6 = -42.7$ to 8.3 litre/min. In the large sample case, we would use the Normal Distribution instead of the t Distribution, putting 1.96 instead of 2.20. We would not then need the differences themselves to follow a Normal Distribution.

On the basis of these data, the mini meter could tend to over-read by as much as 43 litre/min, or to under-read by as much as 8 litre/min. An error of 43 litre/min is quite substantial, and we may have a problem. We would need a much larger sample to obtain a more precise estimate if we thought this were required.

We can also use the t Distribution to test the null hypothesis that the mean difference is zero. If the null hypothesis were true, and the differences follow a Normal Distribution, the test statistic mean/standard error would be from a t Distribution with 11 degrees of freedom. For the example, we have –17.2/11.6 = –1.48. If we go to the 11 d.f. row of Table 2, we find that the probability of such an extreme value arising is greater than 0.10, the 0.10 point of the distribution being 1.80. Using a computer we would find P = 0.17. The data are consistent with the null hypothesis and we have failed to demonstrate the existence of a bias. Note that the confidence interval is more informative than the significance test.

We could also use the sign test to test the null hypothesis of no bias. This gives us 3 positives out of 11 differences (one difference, being zero, gives no useful information) which gives a two sided probability of 0.23. This is greater than the t probability, but fairly similar. The t test gives the smaller probability because, provided the assumption of a Normal distribution is true, the t test is the most powerful test.

The validity of the paired t method depends on three assumptions:

1. Observations are independent
2. The differences are from a Normal Distribution, which we can check with a histogram or a Normal plot,
3. The mean and SD of the differences are constant, i.e. unrelated to magnitude., which we can check with plot of difference against average.

The subjects are 12 different women and so are independent. Figure 10 shows the histogram and Normal plot. There is no obvious skewness. Figure 11 shows a plot of the difference against the subject mean. If the difference depends on magnitude, then we should be careful of drawing any conclusion about the mean difference. We may want to investigate this further, perhaps by looking at the ratio instead of the difference, or estimating the difference as a function of the mean of the two measurements. In this case the difference between the two readings does not appear to be related to the level of PEFR and we need not be concerned about this.

Deviations from the assumptions of the paired t methods

The most likely effect of skewness is that we lose power and may fail to detect differences which exist or have confidence intervals which are too wide. We are unlikely to get spurious significant differences. This means that we need not worry about small departures from the Normal. If there is an obvious departure from the Normal, we should try to transform the data to the Normal before we apply the t Distribution. An alternative is to use a method which does not require these assumptions about the distribution, such as the sign test. There is also a test called the Wilcoxon matched pairs signed rank test (horrible name) which requires quite strong assumptions about the data.

Deviations from the assumptions of t methods

Sometimes we have data which do not follow a Normal distribution or do not have uniform variance. There are two general strategies:

1. Transformations: we find a mathematical function of the data which does have an approximately Normal distribution and uniform variance.
2. Non-parametric methods: we use a method which does not require data to follow a Normal distribution or to have uniform variance.

Figure 12. Serum triglyceride measured in the cord blood of 282 babies, with corresponding Normal distribution curve.

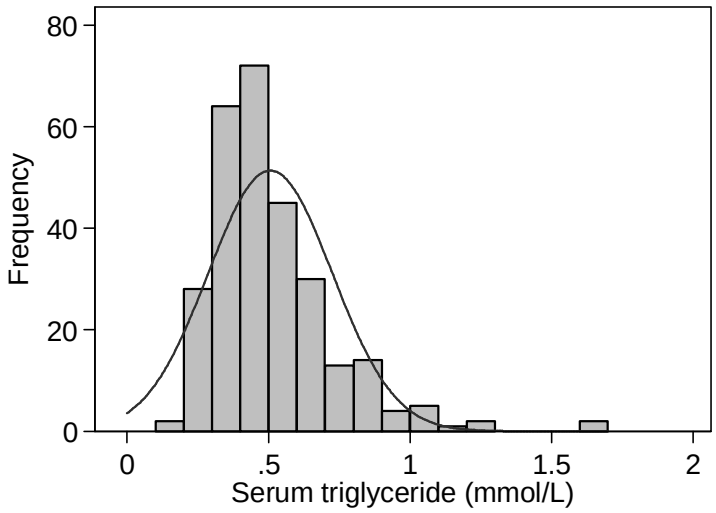
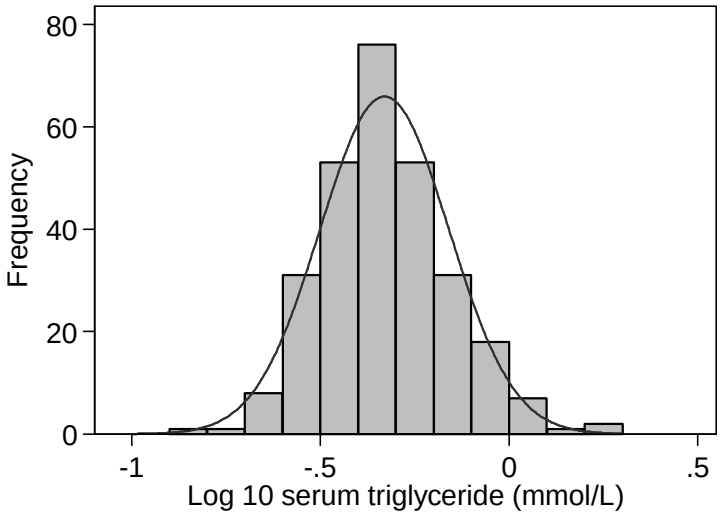


Figure 13. Log transformation to base 10 of serum triglyceride measured in the cord blood of 282 babies, with corresponding Normal distribution curve.



When we use a transformation we analyse a mathematical function of the data. This is often the logarithm, square root, or reciprocal (one over our number). For example, the serum triglyceride of the cord blood of 282 babies is shown in Figure 12. As we noted in the first lecture, this has a positively skew distribution and does not fit the Normal distribution well. The logarithm of these measurements in shown in Figure 13. The distribution is now more symmetrical and quite close to a Normal distribution.

Transformations work well for significance tests. For confidence intervals: the interval can be difficult to interpret. We cannot transform back to the original scale.

When no transformation is possible, we can use other methods known as non-parametric, because they do not need us to assume that data follow a particular family of distributions, such as Normal, where we have to estimate the parameters. Non-parametric methods do not require the assumptions that methods using the t distribution do. These methods provide only significance tests. To use

them to estimate confidence intervals we must make such strong assumptions that we could use a t method anyway.

For two groups we can use the Mann Whitney U test, which is also known as the Wilcoxon two sample test. For this we must assume that our observations are independent, as for the two-sample t method. We also assume observations can be ordered, i.e. we can say whether one observation is greater or smaller than another. If some observations have the same value and so cannot be ordered, we say they are tied. There is an approximation if observations are tied.

For pairs we can use the sign test, already discussed. We must assume pairs of observations independent, as for the paired t method. There is another test called the Wilcoxon matched pairs or signed rank test. Again, we assume the pairs of observations are independent, as for the paired t method. We must also assume differences between pairs are meaningful, so we need quantitative data. We must also assume that the distribution of differences is symmetrical. This is a strong assumption. We could not make it for the pronethalol data, for example. Again, there is an approximation if observations are tied, i.e. differences for two pairs cannot be separated.

Martin Bland,
7 February 2012

References

Altman DG. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Prentice AM, Black AE, Coward WA, Davies HL, Goldberg GR, Murgatroyd PR, Ashford J, Sawyer M, Whitehead RG. (1986) High-levels of energy-expenditure in obese women. *British Medical Journal* **292**, 983-987.