21.6: Analysis of mean values

6.1 Confidence interval for a mean

If the outcome measure is taken as the mean ($\bar{x}$) of a sample of $n$ observations, for example, the weights of a sample of newborn infants, the standard error of the mean is given by $\sigma/\sqrt{n}$, where $\sigma$ is the standard deviation of the variable measured (for example, weights of newborn infants) in the population from which the sample of $n$ observations was taken. The 95% CI on the mean is given by $\bar{x} \pm 1.96(\sigma/\sqrt{n})$.

In general, $\sigma$ (the standard deviation in the population) will not be known but must be estimated, based on the $n$ observations in the sample. Thus, the estimate of $\sigma$ is subject to sampling error also, and this must be taken into account in the computation of the CI on the mean. This is done by using a multiplying factor in the CI calculation taken from tables of the t-distribution, rather than from tables of the ‘Normal’ distribution, on which Table 21.1 was based. The value of the multiplying factor will depend on the size of the sample from which the standard deviation was estimated. For example, for 95% CIs, appropriate multiplying factors for sample sizes of 10, 20, 50, and 100 are 2.26, 2.09, 2.01, and 1.98, respectively. (Note that, in using the tables, the values of $t$ are given for different ‘degrees of freedom’. In the situation considered here, the degrees of freedom correspond to the sample size minus one, i.e. $n - 1$.) If the sample size is 30 or more, little error is introduced by using the value of 1.96 derived from the normal distribution when calculating 95% CI, rather than the appropriate t-value.

If the estimate of the standard deviation, based on the sample, is $s$, the 95% CI on the mean is given by $\bar{x} \pm t(s/\sqrt{n})$. For example, if the mean birthweight of 25 infants was 3.10 kg and the standard deviation of the weights in the sample was 0.90 kg, the 95% CI would be given by $(3.10 \pm 2.06(0.90/\sqrt{25}))$, i.e. 2.73 – 3.47 kg, where the multiplying factor 2.06 is taken from a table of the t-distribution corresponding to 24 df.
### 6.2 Difference between two means

In a trial, it is very common to want to compare the means of observations in different groups, for example, to compare observations from an intervention group with those from a control group. Suppose that two groups are to be compared and the means are \( \bar{x}_1 \) and \( \bar{x}_2 \), respectively, and the corresponding standard deviations observed in the groups are \( s_1 \) and \( s_2 \). The standard error of the difference between the means is given by \( \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \), where \( s \) is the pooled estimate of the standard deviation, based on the observations from the two groups. \( s \) is estimated as:

\[
s = \sqrt{\left( n_1 - 1 \right) s_1^2 + \left( n_2 - 1 \right) s_2^2 \over n_1 + n_2 - 2 }
\]

The 95% CI for the difference between the means is given by:

\[
\left( \bar{x}_1 - \bar{x}_2 \right) \pm t s \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}
\]

where \( t \) is taken from a table of the t-distribution with \( (n_1 + n_2 - 2) \) df.

For example, suppose erythrocyte sedimentation rates (ESRs) were measured in an intervention group and in a control group, as shown in Table 21.7. The standard deviation \( s \) may be calculated as \( \sqrt{\left( 9 \times 2.412 + 11 \times 2.542 \right) \over 10 + 12 - 2} = 2.48 \), and the 95% CI on the difference is given by:

\[
\left( 9.7 - 6.5 \right) \pm \left( 2.09 \times 2.48 \right) \sqrt{\left( \frac{1}{10} + \frac{1}{12} \right)} = 3.2 \pm 2.2 = 1.0 - 5.4.
\]

To test the null hypothesis that there is no true difference in the mean ESRs between the two groups, a statistical test must be performed. A test statistic is calculated to assess the probability of the observed results (or more extreme) if there really is no difference between the two groups. The difference of the means divided by the standard error of the difference gives a value of a test statistic that may be looked up in tables of the t-distribution with \( (n_1 + n_2 - 2) \) df.

For the example in Table 21.7, the test statistic is \( \left( \frac{\left( \bar{x}_1 - \bar{x}_2 \right)}{s \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right)} } \right) = 3.01 \). The associated p-value is 0.0035, i.e. if there really is no effect of the intervention on ESRs, the chance of observing a difference in the means as large or larger than that in the study is 0.35% (i.e. not impossible, but rather unlikely!).

### 6.3 Analysis of more than two groups

If a study involves the comparison of observations in more than two groups, it is necessary to generalize the methods given in Section 6.2. This is straightforward but is beyond the scope of this book, and the reader is referred to standard statistical texts, such as that by Armitage and Berry (1987) or Kirkwood and Sterne (2003), for details. The relevant sections to which to refer are those on ‘one-way analysis of variance’. 
Of course, it is always possible to use the methods given in Section 6.2 to compare groups, just two at a time. This is a reasonable approach, but some caution must be exercised when interpreting the findings, as the chances of finding at least one pair to be significantly different (for example, p < 0.05) may be substantial, even if there are, in truth, no differences between the groups. To illustrate this, suppose six groups are being compared. In an analysis of variance, the question is asked: ‘Considered as a whole, is the variation between the means observed in the six groups more than might be expected to arise by chance if there were no differences in the true means?’ This question may be answered with one statistical test in an analysis of variance, and the null hypothesis may, or may not, be rejected on the basis of this one test. Suppose, however, it was decided to examine all possible pairs of comparisons of the groups. There are 15 possible pairs, and, if a $t$-test was done on each pair, there is a reasonable chance that at least one comparison would be found to be ‘p < 0.05’ by chance alone, because of the number of different tests that had been performed. There are ways of adjusting the significance levels to allow for this effect, and the reader is referred to standard texts again for a discussion of ‘the multiple comparison problem’.