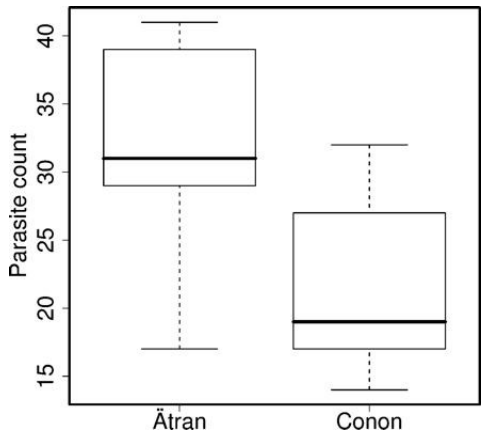
Comparison of two groups. We often want to compare two populations on the basis of experiment. For example, a researcher wants to test the effect of new drug on blood pressure. In the experiment an improvement may have been often due to the **placebo effect** when a participant (a subject) of the treatment believes that he or she has been given an effective treatment.

- Data must be collected from **control group** in which the subjects are given a placebo, and from **treatment group** in which the subjects are treated with the new drug.
- The design of experiment should consider the **randomization** by assigning the subjects between the control and the treatment group randomly.
- A **double-blind experiment** by concealing the nature of treatment from the subjects and the person taking measurements.

Example: Parasite counts for salmons. An experiment with two difference salmon stocks, from River Conon in Scotland and from River Atran in Sweden, was carried out as follows. Thirteen fish from each stock were infected and after four weeks the number of a certain type of parasites was counted for each of the 26 fish with the following results.

Stock	No. of parasites												
Ätran	31	31	32	22	41	31	29	40	41	39	36	17	29
Conon	18	26	16	20	14	28	18	27	17	32	19	17	28

The purpose of the study was to investigate if the number of parasites during an infection is the same for the two salmon stocks.



Parallel boxplots for the two samples are shown as above. The observed parasite counts are generally higher for the Atran group compared to the Conon group, indicating that Atran salmon are more susceptible to parasites.

Unpaired samples and independence.

- It is important to distinguish "paired" samples from "unpaired" samples (that is, independent groups) because different methods of analysis are appropriate for independent groups. For unpaired samples, we impose an assumption of independence between all observations. This means that the observations do not share information.
- This setup with independent samples corresponds to a one-way analysis. It is called "**one-way analysis**" because different sources of variation are compared and "one-way" because only **one factor**, treatment or control, is varied in the experiment.

The purpose of the statistical analysis is to clarify whether the observed difference of two groups is caused by an actual difference or by random (sampling) variation.

Paired samples and dependence. "Paired" samples occur, for example, if two measurements are collected for each subject in the sample under different circumstances (treatments), or if measurements are taken on pairs of related observational units such as twins. In studies with two drugs under investigation, for example, it is common that the subjects try one drug in one period and the other drug in another period; thus, they are "dependent." As a consequence, the difference between two groups is confused with the group variation, making the "one-way analysis" inappropriate for paired data. In the paired samples, rather than using the original measurements, we will use the **pairwise differences**, i.e., compute the difference in two measurements within a pair. This gives us a single sample consisting of differences. If two measurements within the pair do not change we would expect the difference to vary around zero (the null hypothesis).

Summary of grouped data. It is very important to distinguish two independent samples from paired samples because different analysis methods are appropriate.

- **Two independent samples** where the samples correspond to two different groups or treatments and can be assumed to be independent.
- **Paired samples** where the observations consist of pairs of measurements, with the observations in a pair corresponding to two different groups or treatments.

For unpaired samples like the example of parasite counts for salmon we impose an assumption of independence between all observations. Loosely speaking, independence means that the observations do not share information.

Group mean and variation. Consider the situation with n observations split into two groups. Label the group 1 and group 2. Let $g(i)$ denote the group for observation i. Then $g(i)$ has one of the "levels," 1 or 2. The sample mean $\bar{y}_j$ and sample variance s_j^2 in group $j = 1$ or 2 are given by

$$\begin{aligned} \bar{y}_j &= \frac{\sum_{i:g(i)=j} y_i}{n_j} \\ s_j^2 &= \frac{\sum_{i:g(i)=j} (y_i - \bar{y}_j)^2}{n_j - 1} \end{aligned} \tag{5.1}$$

where n_j is the size of group $j = 1$ or 2, and the sum $\sum_{i:g(i)=j}$ means the sum over all observations that belong to group j.

Difference between two groups. The statistical model for the comparison of two groups is given by

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \ldots, n$$

where $g(i)$ is either 1 or 2, and $e_1, \ldots, e_n$ are from $N(0, \sigma^2)$. We are interested in the difference or the contrast $\alpha_1 - \alpha_2$ between two groups rather than the parameters α_1 and α_2 of group. For example, how much larger is the expected response in the treated group compared to the control group? Hence, we are interested in quantity of difference between "group 1" and "group 2." Naturally, the estimate of contrast is simply the difference between the two estimates.

$$\hat{\alpha}_1 - \hat{\alpha}_2 = \bar{y}_1 - \bar{y}_2$$

Pooled variance and standard error. The overall variance s^2 can be computed as a weighted average of the group variance estimates, s_j^2, as follows.

$$s^2 = \frac{\sum_i (y_i - \hat{\alpha}_{g(i)})^2}{n-2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n-2} \tag{5.2}$$

where $n = n_1 + n_2$ is the total size. Note that the group variance s_j^2 is assigned the weight $(n_j - 1)$, the denominator in (5.1). The summation of (5.2) is called the **pooled variance**. Finally we get the corresponding standard error for the estimate of contrast.

$$\text{SE}(\hat{\alpha}_1 - \hat{\alpha}_2) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example: Parasite counts for salmons. The sample mean $\bar{y}_j$ and sample standard deviation s_j are computed for group $j = 1$ (Atran group) and $j = 2$ (Conon group).

$$\bar{y}_1 = 32.23, \qquad s_1 = 7.28$$
$$\bar{y}_2 = 21.54, \qquad s_2 = 5.81$$

The summary statistics and the boxplots tell the same story. The observed parasite counts are generally higher for the Atran group compared to the Conon group, indicating that Atran salmons are more susceptible to parasites. The result might be different, though, if we repeated the experiment and used new samples of salmon. We need to clarify whether the observed difference between $\bar{y}_1$ and $\bar{y}_2$ is caused by an actual difference between the stocks or by random (sampling) variation.

The statistical model for the salmon data is given by

$$y_i = \alpha_{g(i)} + e_i, \quad i = 1, \ldots, 26$$

where $g(i)$ is either $j = 1$ or 2 (either "Atran" or "Conon"), and $e_1, \ldots, e_{26}$ are from $N(0, \sigma^2)$. In other words, all observations are independent, Atran observations are from $N(\alpha_1, \sigma^2)$, and Conon observations are from $N(\alpha_2, \sigma^2)$. We already computed the group means $\bar{y}_j$ and group

standard deviations s_j. With the above notation we obtain the pooled variance s^2 and the standard deviation s.

$$s^2 = \frac{12s_1^2 + 12s_2^2}{24} = 43.4$$
$$s = 6.59$$

The difference in parasite counts is estimated by

$$\hat{\alpha}_1 - \hat{\alpha}_2 = 32.23 - 21.54 = 10.69$$

With a standard error of

$$\text{SE}(\hat{\alpha}_1 - \hat{\alpha}_2) = s\sqrt{\frac{2}{13}} = 2.58$$

the 95% confidence interval for $\alpha_1 - \alpha_2$ is calculated as

$$10.69 \pm (2.064)(2.58) = (5.36,\ 16.02)$$

where the critical value in t_{24}-distribution is $t_{0.025,24} = 2.064$. In particular, we see that zero is not included in the confidence interval, so the data are not in accordance with "difference of zero" between the stock means. In other words, the data suggests that Atran salmon are more susceptible than Conon salmon to parasites during an infection.

Hypothesis test for two groups. The salmon data with two samples corresponding to two different salmon stocks, Atran or Conon, are obtained. If $\alpha_1 = \alpha_2$ then there is no difference between the stocks when it comes to parasites during infections. Hence, the null hypothesis is "$H_0 : \alpha_1 = \alpha_2$." If we define $\theta = \alpha_1 - \alpha_2$ then the hypothesis can be written as "$H_0 : \theta = 0$."

$$\hat{\theta} = \hat{\alpha}_1 - \hat{\alpha}_2 = 32.23 - 21.54 = 10.69$$
$$\text{SE}(\hat{\theta}) = s\sqrt{\frac{2}{13}} = 2.58$$

The t-test statistic is therefore

$$T_{\text{obs}} = \frac{\hat{\theta} - 0}{\text{SE}(\hat{\theta})} = \frac{10.69}{2.58} = 4.14$$

The corresponding p-value is calculated as

$$p\text{-value} = (2)P(T \geq 4.14) = 0.00037$$

where T follows a t-distribution with 24 degree of freedom. Hence, if there is no difference between the two salmon stocks then the observed value 4.14 of T_{obs} is very unlikely. We firmly reject the null hypothesis of the differnece being zero. Since we observe that the estimate of contrast (10.69) is positive, we conclude that Atran salmons are more susceptible to parasites than Conon salmons. Notice how we reached the same conclusion from the confidence interval, since it did not include zero.

Samples with different SD's. Until now we have assumed that the standard deviation σ is the same for all observations. The assumption of **variance homogeneity** is essential, in particular for the computation of standard errors and confidence intervals. However, in the situation with two unpaired samples it is possible to handle the situation with different standard deviations. Assume that the observations are independent and come from two different groups, group 1 and group 2. Both the mean and standard deviation are allowed to vary between groups, so observations from group 1 are from $N(\alpha_1, \sigma_1^2)$ and observations from group 2 are from $N(\alpha_2, \sigma_2^2)$. Then the estimates of the means are unchanged and thus equal to the group sample means, that is,

$$\hat{\alpha}_1 = \bar{y}_1, \quad \hat{\alpha}_2 = \bar{y}_2.$$

Here the standard error for the difference estimate is given by

$$\mathrm{SE}(\hat{\alpha}_1 - \hat{\alpha}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In order to construct a confidence interval, we need to use t-distribution, but since we do not have a pooled standard deviation it is not obvious how many degrees of freedom to use. It turns out that it is appropriate to use

$$df = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\frac{\mathrm{SE}_1^4}{n_1 - 1} + \frac{\mathrm{SE}_2^4}{n_2 - 1}}$$

where $\mathrm{SE}_j^2 = s_j^2/n_j$ for $j = 1$ or 2. The value for df is not necessarily an integer. The corresponding $(1-\alpha)\%$ confidence interval becomes

$$(\hat{\alpha}_1 - \hat{\alpha}_2) \pm (t_{\alpha/2, df})\mathrm{SE}(\hat{\alpha}_1 - \hat{\alpha}_2)$$

Example: Parasite counts for salmons. We already computed a 95% confidence interval for the difference between expected parasite counts for Atran and Conon salmon under the assumption that the standard deviation is the same in both groups. If we are not willing to make this assumption, then we could compute the confidence interval based on the general formula. Here we get the standard error and the degrees of freedom

$$\mathrm{SE}(\hat{\alpha}_1 - \hat{\alpha}_2) = \sqrt{\frac{(7.28)^2}{13} + \frac{(5.81)^2}{13}} = 2.58, \quad df = 22.9$$

The critical value for t-distribution with 22.9 degrees of freedom is calculated as $t_{0.025,22.9} = 2.069$. Thus, the 95% confidence interval for the difference becomes

$$10.69 \pm (2.069)(2.58) = (5.35,\ 16.04)$$

almost the same as the previous calculation, where the standard deviation was assumed to be the same for the two stocks.

When to assume equal SD's?

- For the salmon data, the two confidence intervals (assuming equal standard deviations or not) were almost identical, because there are the same number of observations in the two groups and the sample standard deviations from two groups are close to each other.
- If the group standard deviations are close, we usually prefer to use the confidence interval based on the assumption of equal standard deviations, mainly because the estimate of the standard error is more precise, as it is based on all observations. The results are quite robust as long as the samples are roughly of the same size and not too small.
- Larger differences between the group standard deviations indicate that the assumption of equal standard deviations is not reasonable; hence we would rather use the confidence interval based on the general formula.

Example: Parasite counts for salmons. To calculate the summary statistics on observations for each group, we must use group mean and SD, which requires us to calculate the sum of column values specified for each group. We can use the `tapply()` function to apply a single function, `mean()` or `sd()`. In our example, we wish to use it as follows.

```
tapply(parasites, stock, mean)
tapply(parasites, stock, sd)
```

The first argument to `tapply()` is the column name for data values. The second argument is another column name which determines which group each value belongs to.

The `t.test()` function can be used for the analysis of two samples. The confidence intervals and t-test can be obtained by the following commands:

```
t.test(parasites ~ stock)
t.test(parasites ~ stock, var.equal=T)
```

In the command lines for `t.test()` above, "parasites" on the left-hand side of "~" is modeled as grouped data indicated by "stock" on the right-hand side. The default call `t.test(parasites ~ stock)` corresponds to the analysis of two independent samples with different standard deviations. If the standard deviations are assumed to be identical, then use the option "`var.equal=T`"

Example: Word count. When reading a dataset with the `read.csv()` function, R automatically looks for missing values that look like "NA". Sometimes a empty value ("") indicates that a value is `NA`. The na.strings option can be used to tell R which value must be treated as NA value.

```
Data <- read.csv(file.choose(),na.strings="")
attach(Data)
```

The `attach()` command makes it possible to use the variables "MWC" and "FWC" with reference to the data frame "Data". Then the following command produces parallel boxplots for MWC and FWC.

```
boxplot(MWC, FWC, col="green", ylab="Word counts")
```

The default call `t.test(FWC,MWC)` corresponds to the analysis of two independent samples, "FWC" and "MWC", with different standard deviations. If the standard deviations are assumed

to be identical, then use the option "`var.equal=T`". If the alternative hypothesis is that the first sample x is "greater" than the second group y, then use the option "`alternative="greater"`".

```
t.test(FWC, MWC, alternative="greater", var.equal=T)
```