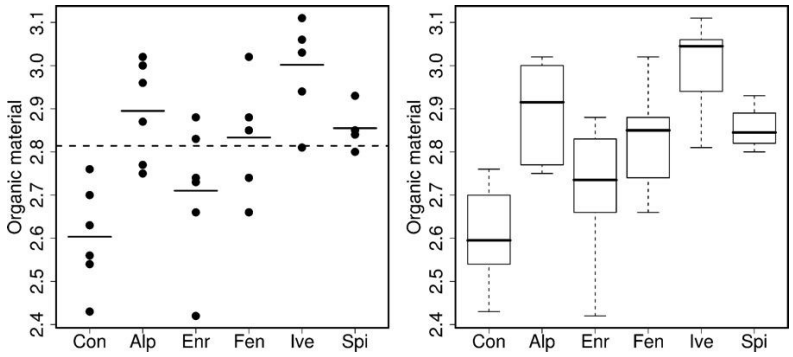
**Comparison of three or more groups.** The purpose of an analysis is often to compare different groups of data. Suppose, for example, that a meat scientist wants to examine the effect of three different storage conditions on the tenderness of meat. For that purpose 24 pieces of meat have been collected and allocated into three storage (or treatment) groups, each of size eight. In each group all eight pieces of meat are stored under the same conditions, and after some time the tenderness of each piece of meat is measured. The main question is whether the different storage conditions affect the tenderness: are the observed differences between the groups due to a real effect, or due to **random variation**? The term "**one-way analysis of variance**" (or **one-way ANOVA**) is used if there are three or more groups.

**Example: Dung decomposition.** An experiment with dung from heifers was carried out in order to explore the influence of antibiotics on the decomposition of dung organic material. As part of the experiment, 36 heifers were divided into six groups. All heifers were fed a standard feed, and antibiotics of different types (alpha-Cypermethrin, Enrofloxacin, Fenbendazole, Ivermectin, Spiramycin) were added to the feed for heifers in five of the groups. No antibiotics were added for heifers in the remaining group (the control group). For each heifer, a bag of dung was dug into the soil, and after eight weeks the amount of organic material was measured for each bag. The primary interest of the antibiotics study was to investigate **if there are differences in the amount of organic material among the antibiotics groups.**




The graph at the left is a strip chart of data points with group sample means (solid line segments) and the total mean of all observations (dashed line). The right is a usual visualization of parallel boxplots.

| Antibiotics | $n_j$ | $\bar{y}_j$ | $s_j$ | $s_j^2$ |
|---|---|---|---|---|
| Control | 6 | 2.603 | 0.119 | 0.0141 |
| $\alpha$-Cypermethrin | 6 | 2.895 | 0.117 | 0.0136 |
| Enrofloxacin | 6 | 2.710 | 0.162 | 0.0262 |
| Fenbendazole | 6 | 2.833 | 0.124 | 0.0153 |
| Ivermectin | 6 | 3.002 | 0.109 | 0.0120 |
| Spiramycin | 4 | 2.855 | 0.054 | 0.0030 |

The sample means and the sample standard deviations are computed for each group separately. We can make similar observations as we did in the boxplots. On average the amount of organic material is lower for the control group than for the antibiotics groups, and except for the spiramycin group the standard deviations are roughly the same in all groups.

The amount of organic material appears to be lower for the control group compared to any of the five types of antibiotics, suggesting that decomposition is generally inhibited by antibiotics. However, there is variation from group to group (**between-group variation**) as well as a relatively large variation within each group (**within-group variation**). The within-group variation seems to be roughly the same for all types, except perhaps for spiramycin, but that is hard to evaluate because there are fewer observations in that group. Analysis of variance (ANOVA) will test the equality of group population means by analyzing variances.

**Group means and SD's.** Consider the situation with n observations split into k groups. Label the groups 1 through $k$. Let $g(i)$ denote the group for observation $i$. Then $g(i)$ has one of the values $1, \ldots, k$. The sample mean $\bar{y}_j$ and sample variance $s_j^2$ in group $j$ are given by

$$\begin{aligned} \bar{y}_j &= \frac{\sum_{i:g(i)=j} y_i}{n_j} \\ s_j^2 &= \frac{\sum_{i:g(i)=j} (y_i - \bar{y}_j)^2}{n_j - 1} \end{aligned} \tag{6.1}$$

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

**Within-group variation. Within-group variation** refers to the variation in each of the groups. It is illustrated by the vertical deviations (residuals) between the observations and their corresponding group means. The **residual sum of squares** is given by

$$\mathrm{SS}_e = \sum_{i=1}^{n} (y_i - \bar{y}_{g(i)})^2$$

It describes the within-group variation since it measures squared deviations between the observations and the group means. The residual degree of freedom is $df = n - k$. Thus, **residual variance**, also known as **residual mean squares (MS)**, becomes

$$\mathrm{MS}_e = \frac{\mathrm{SS}_e}{n - k}$$

**Residual variance.** The residual mean square $\mathrm{MS}_e$ is the estimate of common population variance $\sigma^2$. It is also called **residual variance**, and denoted by $s^2$. It can be computed as a weighted average of the group variance estimates, $s_j^2$, as follows

$$s^2 = \mathrm{MS}_e = \frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{n - k} \tag{6.2}$$

where $n$ is the total size of data. Note that the group variance $s_j^2$ is assigned the weight $n_j - 1$, the denominator in (6.1). The summation of (6.2) becomes the residual sum of squares $\mathrm{SS}_e$.

**Between-group variation. Between-group variation** refers to differences between the groups; for example, deviation between the different treatments in the antibiotics example.

$$\mathrm{SS}_{grp} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2; \quad \mathrm{MS}_{grp} = \frac{\mathrm{SS}_{grp}}{k - 1}$$

As illustrated in the strip chart, it is represented as deviation between the group means $\bar{y}_j$ (horizontal line segments) and the overall mean:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The group means $\bar{y}_i$'s act as our "observations." Thus, $df = k - 1$, and the "average" squared difference $\text{MS}_{grp}$ per group becomes the **mean squares** for the factor of interest (between groups).

**ANOVA model.** In the one-way layout with $k$ groups, the group means $\alpha_1, \ldots, \alpha_k$ are parameters, and we write the one-way ANOVA model

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

where $g(i) = j$ is the group "$j$" that corresponds to the measurement $y_i$. The "error" (or "residual") terms $e_1, \ldots, e_n$ are independent and $N(0, \sigma^2)$-distributed. In other words, it is assumed that there is a normal distribution for each group, and that group means $\alpha_j$'s are different from group to group but all groups share the same standard deviation (namely $\sigma$) representing "within-group variation." The parameter $\alpha_j$ represents the expected value (or the population average) in the $j$-th group.

**Null hypothesis for ANOVA model.** Consider the one-way ANOVA model with group mean $\alpha_j$ in the $j$-th group. As usual, $k$ denotes the number of groups. In a typical model, it tests the null hypothesis that $\mu = \mu_0$. However, in the ANOVA model we are interested in whether there is any difference between the groups. Thus, the null hypothesis of eqaul means is given by

$$H_0 : \alpha_1 = \cdots = \alpha_k$$

and the alternative is the opposite; namely, that at least two $\alpha$'s are different. Once the significant evidence is established, we are often interested in the group differences $\alpha_j - \alpha_l$ by comparing the $j$-th and the $l$-th group.

**Analysis of variance (ANOVA).** If there is no difference between any of the groups, then the group averages $\bar{y}_j$ will be of similar size and be similar to the overall mean $\bar{y}$. Hence, $\text{MS}_{grp}$ will be "small." On the other hand, if groups 1 and 2, say, are different, then the group averarages will be somewhat different; hence, $\text{MS}_{grp}$ will be "large." "Small" and "large" should be measured relative to the within-group variation $\text{MS}_e$, and $\text{MS}_{grp}$ is thus standardized with $\text{MS}_e$. Thus, we use

$$F_{\text{obs}} = \frac{\text{MS}_{grp}}{\text{MS}_e}$$

Large values of $F_{\text{obs}}$ are critical; that is, not in agreement with the assumption (null hypothesis) that there is no different between the groups.

**Analysis of variance Table (ANOVA Table).**

| Variation | SS | df | MS | $F_{\text{obs}}$ | $p$-value |
|---|---|---|---|---|---|
| Between groups | $\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$ | $k-1$ | $\frac{\text{SS}_{\text{grp}}}{\text{df}_{\text{grp}}}$ | $\frac{\text{MS}_{\text{grp}}}{\text{MS}_e}$ | $P(F \geq F_{\text{obs}})$ |
| Residual | $\sum_{i=1}^{n}\left(y_i - \bar{y}_{g(i)}\right)^2$ | $n-k$ | $\frac{\text{SS}_e}{\text{df}_e}$ | | |
| Total | $\sum_{i=1}^{n}(y_i - \bar{y})^2$ | $n-1$ | | | |

This disagreement is equivalent to $F_{\text{obs}}$ being larger, and the corresponding $p$-value are often inserted in an analysis of variance table. The $p$-value of being smaller than 0.05 indicates significance evidence toward the disagreement between groups.

The alternative hypothesis is formulated as

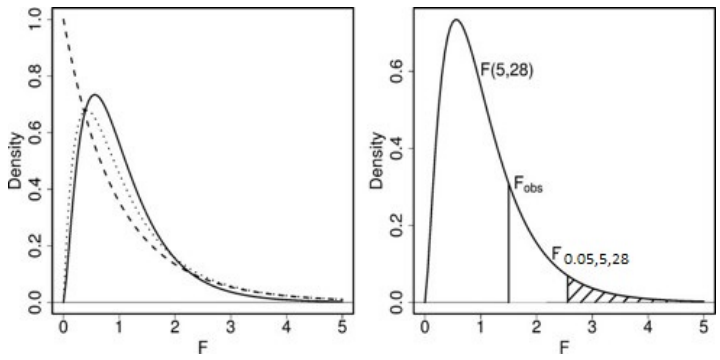
$$H_A : \alpha_j \neq \alpha_l \text{ for some pair } (j, l)$$

If $H_0$ is true then $\text{MS}_{grp}$ will be "small." On the other hand, if $H_0$ is false then $\text{MS}_{grp}$ will be "large." "Small" and "large" should be measured relative to the within-group variation, and $\text{MS}_{grp}$ is thus standardized with $\text{MS}_e$.

$$F_{\text{obs}} = \frac{\text{MS}_{grp}}{\text{MS}_e}$$

becomes the test statistic and note that large values of $F_{\text{obs}}$ are critical; that is, not in agreement with the null hypothesis $H_0$.

**F-test.**




If $H_0$ is true, then $F_{\text{obs}}$ comes from a so-called **F-distribution** with $(k-1, n-k)$ degrees of freedom. In the left the densities for the $F$-distribution represented by solid, dashed, and dotted line are respectively shown for three different pairs, $(5, 28)$, $(2, 27)$, and $(4, 19)$, of degrees of freedom.

Notice that $F_{k-1,n-k}$-distribution has a pair of degrees of freedom (not just a single value) and that the relevant degrees of freedom are the same as those used for computation of $\text{MS}_{grp}$ and $\text{MS}_e$. Since only large values of $F_{\text{obs}}$ are critical, we reject $H_0$ on the 5% significance level if

$$F_{\text{obs}} \geq F_{0.05,k-1,n-k}$$

where $F_{0.05,k-1,n-k}$ denotes the critical value of $F_{k-1,n-k}$-distribution. Equivalently the hypothesis is rejected if the p-value

$$P(F \geq F_{\text{obs}})$$

is 0.05 or smaller. Here $F$ follows $F_{k-1,n-k}$-distribution.

Example: Dung decomposition. The values from dung decomposition study are summarized in the ANOVA table below.

| Variation | SS | df | MS | F | p-value |
|---|---|---|---|---|---|
| Between types | 0.5908 | 5 | 0.1182 | 7.97 | <0.0001 |
| Residual | 0.415 | 28 | 0.0148 | | |

We obtain $F_{\text{obs}} = 7.97 > F_{0.05,5,28} = 2.56$, and conclude that there is strong evidence of group differences. Equivalently we can observe the $p$-value less than 0.0001; thus, the result is highly significant. Subsequently, we need to quantify the conclusion further: Which groups are different and how large are the differences?

**Standard errors in ANOVA model.** Recall the one-way ANOVA model

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

where $g(i)$ denotes the group corresponding to the $i$-th observation and $e_1, \ldots, e_n$ are independent and $N(0, \sigma^2)$-distributed. Then the estimate for the group means $\alpha_1, \ldots, \alpha_k$ are simply the group averages:

$$\hat{\alpha}_j = \frac{\sum_{i:g(i)=j} y_i}{n_j}$$

and the corresponding standard errors are given by

$$\text{SE}(\hat{\alpha}_j) = \frac{s}{\sqrt{n_j}}$$

It suggests that mean parameters for groups with many observations (large $n_j$) are estimated with greater precision than mean parameters with few observations.

**Standard errors in contrasts.** In the ANOVA setup the residual variance $s^2$ is given by

$$s^2 = \text{MS}_e = \frac{\text{SS}_e}{n-k} \tag{6.3}$$

which we call the *pooled variance estimate*. In the one-way ANOVA case we are very often interested in the differences or contrasts between group levels rather than the levels themselves. Hence, we are interested in quantities $\alpha_j - \alpha_l$ for two groups $j$ and $l$. Then the estimate is simply the difference between the two estimates, and the corresponding standard error is given by

$$\text{SE}(\hat{\alpha}_j - \hat{\alpha}_l) = s\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}$$

The formulas above are particularly useful for pairwise comparisons.

**Pairwise comparisons.** Sometimes interest is in particular groups from the experiment, and we want to compare group "$j$" and group "$l$," say. Still, the analysis is carried out using all data since this makes the estimate of the standard deviation more precise. In a sense we "borrow information" from all observations when we estimate the residual variance by $s^2$ in (6.3, even though we use only the data from the two groups in question to estimate the mean difference. The 95% confidence interval for $\alpha_j - \alpha_l$ is calculated as

$$(\hat{\alpha}_j - \hat{\alpha}_l) \pm (t_{0.025,n-k})s\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}$$

Here $s$ is the residual standard deviation from (6.3), and the critical value is obtained from to the $t_{n-k}$-distribution. Simultaneous calculations of confidence interval for different pairs of groups should not be done uncritically, though, due to the multiple testing problem.

The null hypothesis becomes

$$H_0 : \alpha_j - \alpha_l = 0$$

and by rejecting $H_0$ we consider the two-sided alternative

$$H_A : \alpha_j - \alpha_l \neq 0$$

The difference is significant on the 5% significance level if and only if

$$|\hat{\alpha}_j - \hat{\alpha}_l| \geq (t_{0.025,n-k})s\sqrt{\frac{1}{n_j} + \frac{1}{n_l}}$$

The right-hand side of this equation is called the margin of error for the difference between group "$j$" and group "$l$." We can compare differences of two groups, and see if there are significant differences.

**Example: Dung decomposition.** Recall that the residual standard deviation is obtained by

$$s = \sqrt{0.01482} = 0.1217$$

The margin of error for 95% confidence interval of difference between Control and Spiramycin is given by

$$(2.048)(0.1217)\sqrt{\frac{1}{6} + \frac{1}{4}} = 0.161$$

where $t_{0.025,28} = 2.048$ is the critical value from $t_{28}$-distribution. Similarly we can obtain the margin of error for all other comparisons by

$$(2.048)(0.1217)\sqrt{\frac{1}{6} + \frac{1}{6}} = 0.144$$

For the Spiramycin group, we find that

$$\hat{\alpha}_{\text{spiramycin}} - \hat{\alpha}_{\text{control}} = 0.252 > 0.161,$$

so the group is significantly different from the control group. On the other hand, there is no significant difference between the enrofloxacin group and the control group since

$$\hat{\alpha}_{\text{enroflox}} - \hat{\alpha}_{\text{control}} = 0.107 < 0.144.$$

Using the same arguments for the remaining three antibiotic types, we conclude that the amount of organic material is significantly lower for the control groups than for all other groups, except the enrofloxacin group.

**Example: Dung decomposition.** In the one-way ANOVA analysis, we want to test the hypothesis of an overall effect. The test is reported by the summary from the aov() function:

```
model <- aov(org ~ type)
summary(model)
```

The output contains one line per source of variation, and for each source it lists the degrees of freedom, the SS-value, and the MS-value. Moreover, the F test for the effect of type is carried out: the value of F test and the associated $p$-value are reported. Notice how the degrees of freedom for the test, here $(5, 28)$, also appear in the output.

| Antibiotics | $n_j$ | $\hat{\alpha}_j$ | $\mathrm{SE}(\hat{\alpha}_j)$ | $\hat{\alpha}_j - \hat{\alpha}_{\text{control}}$ | $\mathrm{SE}(\hat{\alpha}_j - \hat{\alpha}_{\text{control}})$ |
|---|---|---|---|---|---|
| Control | 6 | 2.603 | 0.0497 | — | — |
| $\alpha$-Cypermethrin | 6 | 2.895 | 0.0497 | 0.2917 | 0.0703 |
| Enrofloxacin | 6 | 2.710 | 0.0497 | 0.1067 | 0.0703 |
| Fenbendazole | 6 | 2.833 | 0.0497 | 0.2300 | 0.0703 |
| Ivermectin | 6 | 3.002 | 0.0497 | 0.3983 | 0.0703 |
| Spiramycin | 4 | 2.855 | 0.0609 | 0.2517 | 0.0786 |

We use the antibiotics data for illustration of multiple comparison, and reproduce the table above. Since the vector "`type`" contains text values, R automatically uses it as a factor. The `lm()` and `summary()` calls are used as follows:

```
outcome <- lm(org ~ type)
summary(outcome)
```

**Multiple comparisons.** When comparing the mean differences simultaneously a simple comparison using t-distribution will inflate the probability of declaring a significant difference when it is not in fact significant. To compare pairwise differences for multiple pairs of groups, we can use the `TukeyHSD()`.

```
TukeyHSD(outcome, conf.level=0.95)
```

For each pair of groups it provides the difference "`diff`" in the observed means, the confidence interval of the lower end point "`lwr`" and the upper end point "`upr`", and "`p adj`" giving the p-value after adjustment for the multiple comparisons.