The analysis of variance

Equality of means of $r \geq 2$ normally distributed populations with similar variances can be tested with a test called the analysis of variance. In this test, the null hypothesis states that means of all populations are equal and the alternative hypothesis states that at least one mean is different from the others:

$$\begin{aligned} &\mathrm{H}_0 : m_1 = m_2 = \cdots = m_r \,, \\ &\mathrm{H}_1 : m_i \neq m_j \,, \quad \text{for at least one pair } (i,j) \,. \end{aligned} \tag{1}$$

To test the null hypothesis, one sample (usually called *treatment*) from each of the r populations is needed. If the size of the i-th treatment is n_i, then the total size is $n = \sum_{i=1}^{r} n_i$.

The test is based on the comparison of the sum of square differences between treatment means and the grand mean (*treatment sum of squares*) q_1 with the sum of square differences of observations within a treatment from the treatment mean (*error sum of squares*) q_2. Using the above sums of squares we express the *mean square of treatments* S_1^2 which describes the variability between the treatments and the *error mean square* S_2^2 which describes the variability within treatments. If the mean square of the treatments is significantly larger than the error mean square, H_0 is rejected. The following test statistics is used:

$$F = \frac{S_1^2}{S_2^2} \,. \tag{2}$$

The test statistics has F distribution with $(r-1, n-r)$ degrees of freedom. The rejection region is $(f_{r-1,n-r;\alpha}, \infty)$. Quantities needed for the calculation of the test statistics are entered into the table:

Source of variation	Sum of squares	Degrees of freedom	Mean square	Test statistics
Between treatments	q_1	$r-1$	$S_1^2 = \dfrac{q_1}{r-1}$	$F = \dfrac{S_1^2}{S_2^2}$
Within treatments	q_2	$n-r$	$S_2^2 = \dfrac{q_2}{n-r}$	
Total	q	$n-1$		

In the calculation we use the following formulas:

$$q = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij}^2 \right) - n m^2 \,, \qquad m = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} \,, \tag{3}$$

$$q_1 = \left(\sum_{i=1}^{r} n_i m_i^2 \right) - n m^2 \,, \qquad m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \,, \tag{4}$$

$$q_2 = q - q_1 \,, \qquad n = \sum_{i=1}^{r} n_i \,. \tag{5}$$

Linear regression

The *correlation coefficient* r describes the suitability of the use of linear regression to describe the interdependence of random variables X and Y:

$$r = \frac{\mathrm{Cov}\,[X,Y]}{\sqrt{\mathrm{Var}\,[X]\,\mathrm{Var}\,[Y]}}\,. \tag{6}$$

Possible values of r are limited to $[-1,1]$. The values of $|r| \approx 1$ show strong linear interdependence of X and Y, while the value of $|r| < 0.5$ shows that the linear regression is inappropriate for describing the interdependence of X and Y, as they may be independent variables or their interdependence is non-linear. Linear regression makes sense if $r \geq 0.75$. For a sample of measurements $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ the correlation coefficient is calculated by:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}\,. \tag{7}$$

In linear regression, the mathematical model of linear dependence is considered:

$$Y = aX + b\,. \tag{8}$$

The goal of linear regression is to find the values of the coefficients a and b which minimize the sum of the squares of the deviations of the observations from the regression line. The solution is:

$$a = \frac{\mathrm{Cov}\,[X,Y]}{\mathrm{Var}\,[X]}\,, \qquad\qquad b = \mathrm{E}\,[Y] - a\mathrm{E}\,[X]\ . \tag{9}$$

The estimators $\hat{a}$ and $\hat{b}$ of the coeficients a and b are determined from the sample of measurements by:

$$\hat{a} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\,, \qquad\qquad \hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{10}$$

The linear regression can also be used when the expected interdependence between the two variables is not linear. In such cases the interdependence may be linearized by applying logarithm or introducing a new variable:

$$\begin{aligned} &Y = b\,\mathrm{e}^{aX} \quad \rightarrow \quad \ln Y = \ln b + aX\,, \\ &Y = aX^2 + b \quad \rightarrow \quad Y = aZ + b\,, \quad \text{where } X^2 = Z\,. \end{aligned} \tag{11}$$