11. Analysis of variance and linear regression

Analysis of variance (ANOVA)

The influence of a factor on a particular property of the observed process/product can be tested with a test called analysis of variance. In the test, equality of means of r populations with $r \geq 2$, corresponding to r levels (treatments) of the factor, is tested. It is presumed that the populations are normally distributed and have similar variances.

The null hypothesis states that means of all populations are equal: $H_0(m_1 = m_2 = \cdots = m_r)$, while the alternative states that at least one mean is different: $H_1\big(m_i \neq m_j, \text{ for at least one pair } (i,j)\big)$. To test the H_0 a sample of n_i measurements is needed from each of the r populations, that is a total of r samples with a total of $n = \sum_{i=1}^{r} n_i$ measurements.

The test is based on a comparison of variation of the measured variable between treatments with variation of the variable within treatments. The **variation between treatments** is expressed by the mean square of differences S_1^2 between treatment means and the grand mean, while the **variation within treatments** is expressed by the mean square of differences S_2^2 between the treatment measurements and the treatment means. If the variation between treatments significantly exceeds the variation within treatments, H_0 is rejected. In the test the test statistic F is used:

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

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

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

In the calculation, the following formulas are used:

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

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

$$m = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i m_i, \qquad q_2 = q - q_1.$$

Linear regression

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

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

Possible values of r are limited to $[-1,1]$. The values of $|r| \approx 1$ show strong linear interdependence of X and Y, while values of $|r| < 0.5$ show that the linear regression is not appropriate for describing the interdependence of X and Y, as they may be independent variables or their interdependence is non-linear. Linear regression makes sense only if $r \geq 0.75$.

Based on a sample of measurements $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, the estimator $\hat{r}$ of the correlation coefficient r is calculated by replacing the covariance and the variances in the above equation with the corresponding sample quantities:

$$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \;\rightarrow\; \widehat{\mathrm{Cov}}[X,Y] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n^2}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\,,$$

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - E[X]^2 \;\rightarrow\; \widehat{\mathrm{Var}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$

$$\mathrm{Var}[Y] = \mathrm{E}[Y^2] - E[Y]^2 \;\rightarrow\; \widehat{\mathrm{Var}}[Y] = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2.$$

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

$$Y = aX + b$$

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

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

The estimators $\hat{a}$ and $\hat{b}$ of the coefficients a and b are determined based on the sample of measurements by the above equations using the corresponding sample quantities as in determination of the correlation coefficient.

Linear regression can also be used when the expected interdependence between the variables X and Y is not linear. In some cases the interdependence may be linearized by applying a logarithm or introducing a new variable:

Original interdependence		Linearized interdependence
$Y = be^{aX}$	$\rightarrow$	$Z = \ln b + aX,\ \text{where } Z = \ln Y,$
$Y = aX^2 + b$	$\rightarrow$	$Y = aZ + b,\ \text{where } Z = X^2.$

11. Analysis of variance and linear regression - problems

1. Compressor is fastened to its housing by springs. In a study the influence of the spring stiffness on the transmission of the vibration between the compressor and the housing is examined. The force amplitudes (in N), measured on the housing for three different types of springs, are gathered in the table. Can we claim that the force amplitude measured on the housing is the same regardless on the spring type? Which type of spring ensures minimum force amplitude on the housing?
   R: Yes. $f = 2.67; p = 0.110$. No type is significantly better than the other.

Spring A	12	17	11	14	10
Spring B	16	11	15	13	14
Spring C	12	11	10	9	12

2. In the process of welding in the protection atmosphere the effect of protective gas composition on the tensile strength of the weld is examined. Three different gas mixtures of Ar and CO_2 are tested. The measured tensile strengths of welds are gathered in the table. Can we claim that the protective gas composition significantly affects the tensile strength of the weld?
   R: Yes. $f = 5.89; p = 0.013$.

A	42	38	37	41	39	40	38
B	36	34	38	35	36		
C	38	39	37	35	36	40	

3. Displacement of a tool tip is measured as a function of the tool oil pressure. Measured data is in the table. Is linear function a reasonable description of dependence of the displacement on the pressure? Determine the coefficients of the corresponding linear function.
   R: Yes. $r = 1.0; y = 0.0594x - 0.238$.

Pressure [bar]	50	100	200	300	400	500
Displacement [µm]	2,6	5,7	11,8	17,7	23,5	29,4

4. Measured values of gas pressure at different values of its volume for a given mass of gas are presented in the table. Determine the constants κ and C in the function $pV^{\kappa} = C$.
   R: $r = -0.953; \kappa = 1.40; C = 0.135$.

V [m^3]	0,05	0,06	0,07	0,08	0,12	0,2
p [bar]	10,88	7,22	4,94	4,75	1,82	1,66

11. Analysis of variance and linear regression – additional problems

1. A passive reduction of the machine noise is attempted by modifying housing of the machine. Four different types of housing are tested. In the table the noise power levels (in dB), measured at a given distance from the machine, are gathered. Can we claim that all tested housings show the same noise damping ability? Which housing is best at damping noise and which is the worst?
   R: No. $f = 4.21$; $p = 0.023$; D damps best, C is the worst.

Housing A	32	36	30	31	34
Housing B	35	30	29	33	32
Housing C	33	37	38	35	34
Housing D	31	34	28	29	30

2. In a pharmaceutical factory various additives are tested to improve the fermentation process. The table gathers the fermentation yields measured for four different additives. Are the effects of the additives equivalent? Which additive provides the highest and which the lowest yield?
   R: No. $f = 3.91$; $p = 0.025$; A the highest, B the lowest.

A	16	19	17	14	21	22	20
B	14	13	12	15	16		
C	19	18	14	20	19	17	
D	16	17	18	21	15		

3. Fuel consumption of engines with various displacements is studied. The table shows the measured standard consumption, depending on the engine displacement. Does it make sense to describe the dependence of the fuel consumption on the engine displacement by a linear function? Determine the coefficients of such linear function. R: Yes. $r = 0.979$; $y = 2x + 3.85$.

Displacement [l]	1,2	1,5	1,6	1,8	2,0	2,5
Consumption [l/100 km]	6,3	6,6	7,0	7,6	8,1	8,7

4. The drag force on the body in the fluid is described by equation $F = CA\rho v^2/2$ where C is drag coefficient, A cross-sectional area of the body, ρ density and v velocity of the fluid. The table gathers the measured drag forces of a round plate with diameter of 50 mm for different air flow velocities. The air density was 1.29 kg/m^3. Determine the drag coefficient of the plate.
   R: $r = 0.945$; $C = 0.966$.

v [m/s]	5,0	7,0	9,0	10,0	12,0	15,0
F [mN]	68	58	75	98	141	305