11 Introduction to correlation analysis

Processing the data we are often interested in the relationship between two variables. Then if they are not independent we are interested in the strength of the relationship. The relationship between two sets of interval-scaled or ratio-scaled random variables is processed by correlation analysis. (Regression analysis and correlation analysis are focused on similar tasks. In case of regression analysis there is processed one dependent variable and one ore more independent variables; correlation analysis is aimed to measure the strength of two equivalent variables.)

In this chapter only linear relationship is treated and the bivariate normal distribution is assumed.

Remark 11.1
Let us recall the definition and properties of the correlation coefficient.

$$R(X,Y) = \begin{cases} E\left(\frac{X-E(X)}{\sqrt{D(X)}} \cdot \frac{Y-E(Y)}{\sqrt{D(Y)}}\right) & \text{for } \sqrt{D(X)}\sqrt{D(Y)} > 0 \\ 0 & \text{otherwise} \end{cases}$$

The properties:

1. $R(X,Y) = \begin{cases} \frac{C(X,Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}} & \text{for } \sqrt{D(X)}\sqrt{D(Y)} > 0 \\ 0 & \text{otherwise} \end{cases}$
2. $R(X,X) = \begin{cases} 1 & \text{pro } D(X) \neq 0 \\ 0 & \text{jinak} \end{cases}$
3. $R(X,Y) = R(Y,X)$
4. $-1 \leq R(X,Y) \leq 1$
5. $R(X,Y) = 1$, then constants $a, b \in \mathbf{R}, b > 0$ exists such that $P(Y = a + bX) = 1$,
   $R(X,Y) = -1$, then constants $a, b \in \mathbf{R}, b < 0$ exists such that $P(Y = a + bX) = 1$,
6. $R(a + bX, c + dY) = \text{sgn}(bd)R(X,Y)$
7. If the random variables X, Y are independet then $R(X,Y) = 0$.
   (The reverse implication does not hold in general!)

It is obvious that if the relationship between two variables is linear, the correlation coefficient is a perfect indicator of the strength of this relationship. As the value of $|R(X,Y)|$ approaches to 1, the relationship between X, Y is stronger. The positive values of the correlation coefficient are related to the positive slope of positive linear dependence. The negative values of the correlation coefficient are related to the negative slope of negative linear dependence. If the random variables are independent then the correlation coefficient is equal to zero. [It can be zero in case of some non-linear dependence as well!!]

The population correlation coefficient is usually unknown since the distribution of the random vector (X,Y) is usually unknown. But it can be estimated by sample correlation coefficient.

Definition 11.2
Let the random sample $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ follows a bivariate distribution Let M_1, M_2 be sample means, $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2$; $S_2^2 \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2$ be sample variances and
$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ be a sample covariance.
Then

$$R_{12} = \frac{S_{12}}{S_1 \cdot S_2} \text{ for } S_1 \cdot S_2 > 0$$

is called sample correlation coefficient. If S_1 or S_2 are equal to zero then correlation coefficientis not defined.

Remark 11.3
The sample correlation coefficient R_{12} is not an unbiased estimator of population correlation coefficient $R(X,Y)$, but for $n > 30$ rhe bias is negligible. The properties of the sample correlation coefficient R_{12} are paralel to the population correlation coefficient $R(X,Y)$.

In the following text the bivariate normal distribution of a random sample $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ will be assumed.

Theorem 11.4
Let the random vector X, Y follows bivariate normal distribution. Then the random variables X and Y are independent if and only if the correlation coefficient $\varrho = R(X,Y) = 0$. [In case of bivariate normal distribution the independence and non-correlation is equivalent.]

Theorem 11.5
Let $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ be a random sample from bivariate normal distribution. and let $\varrho = 0$. Then the statistic

$$T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$$

follows the student $t-$distribution with $(n-2)$ degrees of freedom. This T statistic is instrumental towards hypothesis about independence of random variables X, Y.

Theorem 11.6
Considering the random sample from bivariate normal distribution, at the significance level α the null hypothesis $H_0: \ \varrho = 0$is rejected in favour of alternative hypothesis H_1, if the test statistic $T = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$ falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

for two-tailed test $\quad H_1: \ \varrho \neq 0 \quad W = (-\infty, \ -t_{1-\alpha/2}(n-2)\rangle \bigcup \langle t_{1-\alpha/2}(n-2), \ \infty)$
for left-tailed test $\quad H_1: \ \varrho < 0 \quad W = (-\infty, \ -t_{1-\alpha}(n-2)\rangle$
for right-tailed test $\quad H_1: \ \varrho > 0 \quad W = \langle t_{1-\alpha}(n-2), \ \infty)$

Example 11.7
The score of two subjects of eight randomly drawn students are recorded.

1	2	3	4	5	6	7	8
80	50	36	58	42	60	56	68
65	60	35	39	48	44	48	61

At the significance level 0.05 carry out the test that the results in considered two subjects are not positively correlated.

Solution
seminar session

Through the hypothesis $H_0: \ \varrho = 0$ the independence of two normal variables was tested. Now we are interested in the strength of linear relationship. The test statistic of following test about correlation coefficient ϱ is made through use of a particular function of sample correlation coefficient R_{12} given by the following theorem.

Theorem 11.8
Let $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ be a random sample from bivariate normal distribution with correlation coefficient $R(X,Y) = \varrho$. The statistic

$$Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$$

is called *Fisher R_{12}-to-z transformation* and its approximate expected value and variance follows:
$E(Z) = \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)}$
$D(Z) = \frac{1}{n-3}$.
Then standardized statistic $U = \frac{Z-E(Z)}{\sqrt{D(Z)}} \approx N(0,1)$.

Theorem 11.9
Let $\binom{X_1}{Y_1}, \ldots, \binom{X_n}{Y_n}$ be a random sample from bivariate normal distribution with correlation coefficient $R(X,Y) = \varrho$. Let R_{12} be the sample correlation coefficient, let $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ be Fisher R_{12}-to-z transformation and let $c \in (-1,1)$ be a given constant.
At the significance level α the null hypothesis $H_0: \varrho = c$ is rejected in favour of alternative hypothesis H_1, if the test statistic

$$U = \frac{Z - \frac{1}{2}\ln\frac{1+c}{1-c} - \frac{c}{2(n-1)}}{\sqrt{\frac{1}{n-3}}}$$

falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

for two-tailed test	$H_1: \varrho \neq c$	$W = (-\infty, -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2}, \infty)$
for left-tailed test	$H_1: \varrho < c$	$W = (-\infty, -u_{1-\alpha}\rangle$
for right-tailed test	$H_1: \varrho > c$	$W = \langle u_{1-\alpha}, \infty)$

Example 11.10
A ferrum content was determined in an iron ore sample of size 600 by two analytic methods, where the sample correlation coefficient was $R_{12} = 0,85$. A technical literature states that the correlation coefficient between considered methods is $\varrho = 0,9$. At the significance level 0.05 carry out a test $H_0 : \varrho = 0,9$ against $H_1 : \varrho \neq 0,9$.

Solution
seminar session
The statistic U can be used to find confidence intervals for ϱ. First the limits for the constant $\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho}$ are derived, then these limits are transformed to the limits for ϱ using hyperbolic tangent.

Theorem 11.11
Let the assumptions from 11.9 hold. Then the $100(1-\alpha)\%$ confidence interval
•for the expression $\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho}$ has the form:
$\frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} \in \left(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\ ,\ Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$ with approximate probability $1-\alpha$.
•for the parameter ϱ has the form:
$\varrho \in \left(\mathrm{tgh}(Z - \frac{u_{1-\alpha/2}}{\sqrt{n-3}})\ ,\ \mathrm{tgh}(Z + \frac{u_{1-\alpha/2}}{\sqrt{n-3}})\right)$ with approximate probability $1-\alpha$.

Remark 11.12
$\mathrm{tgh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ for $x \in \mathbf{R}$.

Example 11.13
An officer of human resources department of particular firm is interested in a relationship between a number of absence days due to illness per year (variable Y) and age of employee (variable X). Therefore the data about 10 employees were drawn randomly.

1	2	3	4	5	6	7	8	9	10
27	61	37	23	46	58	29	36	64	40
15	6	10	18	9	7	14	11	5	8

Under the assumption that $\binom{X}{Y}$ follows bivariate normal distribution do following tasks:

a) Calculate sample correlation coefficient.

b) At the significance level 0.05 carry out a test that X and Y are independent.

c) Determine the 95% confidence interval for correlation coefficient ϱ.

Solution
seminar session

Remark 11.14
We may have two sample correlation coefficients R_{12}, R_{12}^* corresponding to two independent bivariate normal distributions. The question to be asked is: "Do both of these sample correlation coefficients represent population having the same true value of correlation coefficient $\varrho = \varrho^*$? The following theorem deals this question.

Theorem 11.15
Two independent bivariate normal samples of sizes n and n^* with correlation coefficients $\varrho,\ \varrho^*$ are given. Let R_{12}, R_{12}^* be sample correlation coefficients and Z, Z^* are corresponding Fisher transformations.
At the significance level α the null hypothesis $H_0 :\ \varrho = \varrho^*$ is rejected in favour of alternative hypothesis H_1, if the test statistic

$$U = \frac{Z - Z^*}{\sqrt{\frac{1}{n-3} + \frac{1}{n^*-3}}}$$

falls within the critical region W. According to the form of the alternative hypothesis the list of corresponding critical regions follows :

for two-tailed test	$H_1 :\ \varrho \neq \varrho^*$	$W = (-\infty,\ -u_{1-\alpha/2}\rangle \bigcup \langle u_{1-\alpha/2},\ \infty)$
for left-tailed test	$H_1 :\ \varrho < \varrho^*$	$W = (-\infty,\ -u_{1-\alpha}\rangle$
for right-tailed test	$H_1 :\ \varrho > \varrho^*$	$W = \langle u_{1-\alpha},\ \infty)$

Example 11.16
A medical research observed the concentration of substances A and B in urine of patients with particular kidney illness. In a sample of 100 healthy individuals the sample correlation coefficient between concentration of A and B was 0,65. In a sample of 142 individuals with mentioned kidney illness the sample correlation coefficient was 0,37. At the significance level 0.05 test the hypothesis that the true correlation coefficients are equal.

Solution
seminar session