Mathematical Statistics

Jiří Anděl

January 10, 2017

Contents

Preface

Mathematical Statistics I — Statistical Methods is the course in winter semester in the frame 4/2 hours. It is devoted to the students who intend to continue to study probability, mathematical statistics, and econometry.

The course contains also some practical statistical methods and numerical examples. However, their number is very limited, because they will be trained in a special course. We expect that the material will be complemented by programme `R`.

The end of a proof is denoted by $\square$ and the end of an example by $\diamondsuit$.

Prague January 11, 2017

Chapter 1

Random variables

1.1 Introduction

In our life we frequently meet random events. To work with them mathematically, it is necessary to build a model for them. All possible results of an experiment constitute *a space of elementary events* Ω. Elements of the space Ω are denoted by ω. If the space Ω is too rich, we cannot deal with individual *elementary events* ω separately. A useful mathematical model uses only such sets of elementary events which form σ-algebra. Remember that σ-algebra is a non-empty system of subsets such that

(i) If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$.
(ii) If $A_1 \in \mathcal{A},\ A_2 \in \mathcal{A}, \ldots$, then $\cup A_i \in \mathcal{A}$.

Here $\Omega \setminus A = A^c$ is the complement of the set A. From the definition of σ-algebra it is possible to derive its further properties, especially:

(iii) $\emptyset \in \mathcal{A},\ \Omega \in \mathcal{A}$.
(iv) If $A_1 \in \mathcal{A},\ A_2 \in \mathcal{A}, \ldots$, then also $\cap A_i \in \mathcal{A}$.

The pair $(\Omega, \mathcal{A})$ is called *measurable space*. Assume now, that to every set $A \in \mathcal{A}$ it is possible to define a number $\mathsf{P}(A)$ such that the following conditions are fulfilled:

(a) $\mathsf{P}(A) \geq 0$ for every $A \in \mathcal{A}$.
(b) If $A_1, A_2, \ldots$ are disjoint sets belonging to $\mathcal{A}$, then $\mathsf{P}(\cup A_i) = \sum \mathsf{P}(A_i)$.
(c) $\mathsf{P}(\Omega) = 1$.

We get function P defined on $\mathcal{A}$, which is called *probability measure* (shortly *probability*). Conditions (a)–(c) can be formulated in words that probability is σ-additive normed set function. The subsets of the space Ω belonging to $\mathcal{A}$ are called *random events*.

We can ask why probability is introduced in such a complicated way. Why probability is not defined for every subset of the space Ω? Generally, it is not possible and this answer is related to the existence of non-measurable sets.

It is worth to notice that this definition of probability contains no instruction how to calculate it. This problem is, however, solved in other courses of mathematical statistics.

Recall that Borel σ-algebra $\mathcal{B}_n$ of subsets of n-dimensional Euclidean space $\mathbb{R}_n$ is the minimal σ-algebra, which contains all open intervals. Instead of $\mathcal{B}_1$ we usually write $\mathcal{B}$ and instead of $\mathbb{R}_1$ we write $\mathbb{R}$.

We say that X is *a measurable mapping* of the space $(\Omega, \mathcal{A}, \mathsf{P})$ into $(\mathbb{R}, \mathcal{B})$, if

$$\{\omega : X(\omega) \in B\} \in \mathcal{A} \quad \text{for every set} \quad B \in \mathcal{B}.$$

A measurable mapping X is called *random variable*. The random variable X is a function defined on the space Ω and its values $X(\omega)$ are real numbers. Of course, the definition implies that the function must be measurable.

The distribution function F of the random variable X is defined by the formula

$$F(x) = \mathsf{P}(X \leq x), \qquad x \in \mathbb{R}.$$

More precisely, instead of $\mathsf{P}(X \leq x)$ we should write $\mathsf{P}\{\omega : X(\omega) \leq x\}$, but the argument ω is usually dropped. On the measurable space $(\mathbb{R}, \mathcal{B})$ there exists a probability measure corresponding to the distribution function P and it is called *the distribution of the random variable* X.

Some distributions have their fixed abbreviations. For example, *the normal distribution* with parameters μ and σ^2 is denoted by $\mathsf{N}(\mu, \sigma^2)$. The fact that the random variable X has the distribution $\mathsf{N}(\mu, \sigma^2)$ is written as $X \sim \mathsf{N}(\mu, \sigma^2)$. Analogous abbreviation is used also in other cases.

The mean value of the random variable X is

$$\mathsf{E}X = \int_\Omega X(\omega) \,\mathrm{d}\mathsf{P}(\omega), \tag{1.1}$$

if this integral exists. The denotation E is abbreviation of the word expectation. Formula (1.1) is starting point for many theoretical considerations but it is not suitable for practical calculations. The reason is that the measure P is only rarely known. It is a good luck that $\mathsf{E}X$ can be also expressed using integral where the measure is described only by distribution function.

Theorem 1.1 *We have*

$$\mathsf{E}X = \int_{-\infty}^{\infty} x \,\mathrm{d}F(x). \tag{1.2}$$

If g is such a function that $\mathsf{E}g(X)$ exists, then we have

$$\mathsf{E}g(X) = \int_{-\infty}^{\infty} g(x) \,\mathrm{d}F(x). \tag{1.3}$$

Proof. See Anděl (2007). □.

Formula (1.3) is very important. A direct calculation of $\mathsf{E}g(X)$ using only (1.2) would need the following steps. We would define $Y = g(X)$, then from the known distribution function F of the variable X we would have to calculate the distribution function G of the variable Y and finally using(1.2) we would get $\mathsf{E}Y = \int_{-\infty}^{\infty} y \,\mathrm{d}G(y)$. However, calculation of G is usually very complicated. Thus theorem that integral $\int y \,\mathrm{d}G(y)$ is equal to considerably simpler $\int g(x) \,\mathrm{d}F(x)$ is really very important.

Practically we use only two families of distribution functions. If there exists such a function f, that for every real x

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \tag{1.4}$$

holds, then we say that F corresponds to a *continuous distribution*. The function f is called *density of the distribution*. Since F is nondecreasing, the density f is nonnegative almost everywhere. Using (1.4) we see that $f(x) = F'(x)$ almost everywhere.

The second case arises when F je *a step funcion*. It means that there exists maximally counted set of numbers $x_1, x_2, \ldots$, in which F has steps $p_1, p_2, \ldots$. Otherwise F is constant. So we have $\sum p_i = 1$. In this case we say that F correspons to a *discrete distribution*.

Formula (1.3) implies that for the continuous distribution we have

$$\mathsf{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$$

and for the discrete distribution we have

$$\mathsf{E}g(X) = \sum_i g(x_i)\,p_i.$$

Notice that (1.3) enables to calculate probability $\mathsf{P}(X \in B)$ that X belongs to set $B \in \mathcal{B}$. If we take $g = \chi_B$ (the characteristic function of the set B), then from(1.3) we get

$$\mathsf{P}(X \in B) = \int_B \mathrm{d}F(x).$$

In the continuous case we have

$$\mathsf{P}(X \in B) = \int_B f(x)\,\mathrm{d}x$$

and in the discrete case

$$\mathsf{P}(X \in B) = \sum_{\{i:\, x_i \in B\}} p_i.$$

1.2 Moments

Let X be a random variable. *Moment* of the k-th order is

$$\mu'_k = \mathsf{E}X^k, \qquad k = 0, 1, \ldots \tag{1.5}$$

and *central moment* of the k-th order is

$$\mu_k = \mathsf{E}(X - \mathsf{E}X)^k, \qquad k = 0, 1, \ldots. \tag{1.6}$$

It is clear that the moments can be defined only in the cases when the integrals (1.5) and (1.6) exist. Further we define *absolute moments*

$$\mu_k^{\mathrm{abs}} = \mathsf{E}|X|^k, \qquad k \ \geq 0.$$

Here $\mu_1' = \mathsf{E}X$ is expectation and it is shortly denoted by μ. We have $\mu_0' = 1$, $\mu_0 = 1$, $\mu_1 = 0$. An important moment is μ_2. It is called *variance* and denoted by σ^2. From formula (1.6) we get that

$$\sigma^2 = \mathsf{E}X^2 - (\mathsf{E}X)^2.$$

If we want to emphasize that it is the *variance* of the random variable X, we write σ_X^2 or $\mathsf{var}\, X$ instead of σ^2. The parameter σ is called *standard deviation*.

Using binomial theorem we get from (1.6)

$$\mu_k = \sum_{i=0}^{k} \binom{k}{i} (-1)^i \mu_{k-i}' \mu^i.$$

Thus

$$\begin{aligned} \mu_3 &= \mu_3' - 3\mu_2'\mu + 2\mu^3, \\ \mu_4 &= \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4. \end{aligned}$$

If $\sigma^2 > 0$ and if the moments μ_3 and μ_4 exist we define *kurtosis* α_3 and *skewness* α_4 using formulas

$$\alpha_3 = \frac{\mu_3}{\sigma^3}, \qquad \alpha_4 = \frac{\mu_4}{\sigma^4}.$$

Theorem 1.2 *Define $Y = a + bX$. If $\mathsf{E}X$ exists, then $\mathsf{E}Y = a + b\mathsf{E}X$. If $\mathsf{E}X^2 < \infty$ then $\mathsf{var}\, Y = b^2\, \mathsf{var}\ X$.*

Proof. Formulas follow by easy insertion. □

Especially for $b = 1$ we have $\mathsf{var}(a + X) = \mathsf{var}\, X$. We see that variance is shift - invariant.

Theorem 1.3 *Kurtosis and skewness fulfill the inequality*

$$\alpha_4 - \alpha_3^2 \geq 1.$$

Proof. Let $\mathsf{E}X^4 < \infty$. Define $Y = X - \mathsf{E}X$. Then $\mathsf{E}Y = 0$ and $\mathsf{E}Y^k = \mathsf{E}(X - \mathsf{E}X)^k$ for every k. Choose real numbers a, b, c and introduce the vector $\boldsymbol{u} = (a, b, c)'$. Then we have

$$\begin{aligned} 0 \leq \mathsf{E}(a + bY + cY^2)^2 &= \mathsf{E}(a^2 + 2abY + 2acY^2 + b^2Y^2 + 2bcY^3 + c^2Y^4) \\ &= a^2 + 2ac\mu_2 + b^2\mu_2 + 2bc\mu_3 + c^2\mu_4 = \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}, \end{aligned}$$

where

$$\boldsymbol{M} = \begin{pmatrix} 1 & 0 & \mu_2 \\ 0 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}.$$

The matrix $\boldsymbol{M}$ is symmetric. The inequality $\boldsymbol{u}'\boldsymbol{M}\boldsymbol{u} \geq 0$ holding for every vector $\boldsymbol{u}$ proves that $\boldsymbol{M}$ is positive semidefinite. Its determinant is thus nonnegative. This implies

$$0 \leq |\boldsymbol{M}| = \mu_2\mu_4 - \mu_3^2 - \mu_2^3,$$

which gives

$$\mu_2\mu_4 - \mu_3^2 \geq \mu_2^3.$$

We divide the last inequality by μ_2^3, and this gives the assertion of the theorem. □

Some distributions have moments of all orders, e.g. *rectangular distribution* $\mathsf{R}(0,1)$ with the density

$$f(x) = \begin{cases} 1 & \text{for } x \in (0,1), \\ 0 & \text{for } x \notin (0,1). \end{cases}$$

We shall quite often work with densities that are nonvanishing only on a set M. Then we write only the formula for density $f(x)$ for $x \in M$. If we do not introduce formula for $x \notin M$, then we understand $f(x) = 0$ for $x \notin M$. There are distributions where even the moment of the first order does not exist. This is the case of *Cauchy distribution* $\mathsf{C}(a,b)$ with the density

$$f(x) = \frac{1}{\pi}\frac{b}{b^2 + (x-a)^2}, \qquad x \in \mathbb{R}, \quad b > 0.$$

Let us mention that this is not the worst case. If we consider absolute moments μ_k^{abs} for all nonnegative k (not only for nonnegative integers), the distribution $\mathsf{C}(a,b)$ would have finite absolute moments μ_k^{abs} for $k \in [0,1)$. But the distribution with the density

$$f(x) = \frac{1}{2|x| \ln^2 |x|} \qquad \text{for } |x| > \mathrm{e}$$

has no finite absolute moments of the k-th order for $k > 0$. (See Stoops, Barr 1971.)

In connection with moments some important problems have been formulated. We can ask, for example, if a distribution having all moments $\mu'_k\,(k = 0, 1, \dots)$ finite is uniquely determined. The answer is: sometimes yes, sometimes not. Consider the density of the *log-normal distribution*

$$p_1(x) = \frac{1}{\sqrt{2\pi}x} \exp\left\{-\frac{1}{2}\ln^2 x\right\} \quad \text{for } x > 0$$

and the density defined by the formula

$$p_2(x) = [1 + a\sin(2\pi \ln x)]p_1(x) \quad \text{for } x > 0,$$

where $a \in [-1, 1]$ is an arbitrary number. Both the densities have the same series of the moments $\mu'_0, \mu'_1, \dots$. This situation is quite unpleasant since the lognormal distribution is frequently used in theory as well as in practical applications. We have shown that this distribution cannot be characterized by its moments.

However, many distributions are determined uniquely by their moments. In many cases the following theorem is used.

Theorem 1.4 *Let $\mu'_1, \mu'_2, \dots$ be a sequence of moments. If the series*

$$\sum_{k=1}^{\infty} \frac{\mu'_k}{k!} t^k$$

for some $t > 0$ converges absolutely, then the given series of the moments determines the distribution function uniquely.

Proof. See Rao (1978). □

The condition introduced in Theorem 1.4 implies that the *normal distribution* $\mathsf{N}(\mu, \sigma^2)$ with the density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad x \in \mathbb{R}, \quad \sigma > 0,$$

is determined uniquely by its moments. We can see it on the distribution $\mathsf{N}(0, \sigma^2)$, the moments of which are

$$\mu'_{2k-1} = 0, \qquad \mu'_{2k} = \frac{(2k)!\,\sigma^{2k}}{k!\,2^k} \qquad \text{for } k = 1, 2, \ldots .$$

Since the distribution $\mathsf{N}(0, \sigma^2)$ has vanishing expectation, the moments μ'_k are the same as μ_k.

If $\mu = 0, \sigma^2 = 1$, we have *the standard normal distribution*. Its density is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

and the distribution function

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)\,\mathrm{d}u.$$

The function φ is even, because $\varphi(-x) = \varphi(x)$. It implies that $\Phi(-x) = 1 - \Phi(x)$.

We say that a random variable X has a *degenerated distribution* if X is equal to a constant μ with probability 1. It is clear that in this case $\mathsf{E}X = \mu$, $\mathsf{var}\, X = 0$. Thus this degenerated distribution is considered as $\mathsf{N}(\mu, 0)$.

Not every sequence of numbers is a *sequence of moments*. For example, there exists no distribution with the moments $\mu'_1 = 2, \quad \mu'_2 = 1$. For such a distribution it would hold

$$\sigma^2 = \mu'_2 - (\mu'_1)^2 = 1 - 2^2 = -3 < 0.$$

This is not possible, because the variance is nonnegative (see its definition). Let us investigate the problem when a finite series of numbers $\mu'_0 = 1,\ \mu'_1, \ldots, \mu'_{2n}$ is a series of moments. Consider an arbitrary vector $\boldsymbol{c} = (c_0, \ldots, c_n)'$ with real components. It is easy to verify that

$$0 \le \mathsf{E}\left(\sum_{j=0}^{n} c_j X^j\right)^2 = \mathsf{E}\sum_{j=0}^{n}\sum_{k=0}^{n} c_j c_k X^{j+k} = \sum_{j=0}^{n}\sum_{k=0}^{n} c_j c_k \mu'_{j+k} = \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c},$$

where

$$\boldsymbol{A} = \begin{pmatrix} \mu'_0 & \mu'_1 & \cdots & \mu'_n \\ \mu'_1 & \mu'_2 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \mu'_n & \mu'_{n+1} & \ldots & \mu'_{2n} \end{pmatrix}.$$

It means that the matrix $\boldsymbol{A}$ must be positively semidefinite. It can be proved that it is also a sufficient condition (see Krejn, Nudelman 1973). In the mentioned book it is

proved that an infinite series of numbers $\mu'_0 = 1,\ \mu'_1, \mu'_2, \dots$ is a series of of moments if and only if the infinite quadratic form

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} c_j c_k \mu'_{j+k}$$

is positive definite.

The following formulas can be used for calculating moments (see Shao 2005).

Theorem 1.5 *Let F be a distribution function and a a real number. Then $\int [F(x+a) - F(x)]\,\mathrm{d}x = a$.*

Dkaz. Assume first that $a > 0$. Then using Fubini theorem we get

$$\int [F(x+a)-F(x)]\,\mathrm{d}x = \int \left[\int_x^{x+a} \mathrm{d}F(y)\right] \mathrm{d}x = \int \left(\int_{y-a}^{y} \mathrm{d}x\right) \mathrm{d}F(y) = \int a\,\mathrm{d}F(y) = a.$$

The proof is analogous for $a < 0$. □

Theorem 1.6 *Let X be a nonnegative random variable with the distribution function F and with a finite expectation. Then*

$$\mathsf{E}X = \int_0^{\infty} [1 - F(x)]\,\mathrm{d}x.$$

Proof. We have

$$\int_0^{\infty} [1 - F(x)]\,\mathrm{d}x = \int_0^{\infty} \left[\int_x^{\infty} \mathrm{d}F(y)\right] \mathrm{d}x = \int_0^{\infty} \left[\int_0^{y} \mathrm{d}x\right] \mathrm{d}F(y) = \int_0^{\infty} y\,\mathrm{d}F(y)$$

$$= \mathsf{E}X. \qquad \square$$

Theorem 1.7 *Let X be a random variable with distribution function F and with a finite expectation. Then*

$$\mathsf{E}X = \int_0^{\infty} [1 - F(x)]\,\mathrm{d}x - \int_{-\infty}^{0} F(x)\,\mathrm{d}x.$$

Proof. It was proved that

$$\int_0^{\infty} [1 - F(x)]\,\mathrm{d}x = \int_0^{\infty} y\,\mathrm{d}F(y).$$

Further we have

$$\int_{-\infty}^{0} F(x)\,\mathrm{d}x = \int_{-\infty}^{0} \left[\int_{-\infty}^{x} \mathrm{d}F(y)\right] \mathrm{d}x = \int_{-\infty}^{0} \left[\int_{y}^{0} \mathrm{d}x\right] \mathrm{d}F(y)$$

$$= \int_{-\infty}^{0} (-y)\,\mathrm{d}F(y) = -\int_{-\infty}^{0} y\,\mathrm{d}F(y).$$

Thus

$$\mathsf{E}X = \int_{-\infty}^{\infty} y\,\mathrm{d}F(y) = \int_{-\infty}^{0} y\,\mathrm{d}F(y) + \int_0^{\infty} y\,\mathrm{d}F(y)$$

$$= -\int_{-\infty}^{0} F(x)\,\mathrm{d}x + \int_0^{\infty} [1 - F(x)]\,\mathrm{d}x. \qquad \square$$

Theorem 1.8 *Let X be a nonnegative random variable with the distribution function F. Let $\mathsf{E}X^2 < \infty$. Then*

$$\mathsf{E}X^2 = 2\int_0^\infty x[1 - F(x)]\,\mathrm{d}x.$$

Proof. We have

$$2\int_0^\infty x[1 - F(x)]\,\mathrm{d}x = 2\int_0^\infty x\left[\int_x^\infty \mathrm{d}F(y)\right]\mathrm{d}x = 2\int_0^\infty \left[\int_0^y x\,\mathrm{d}x\right]\mathrm{d}F(y)$$

$$= \int_0^\infty y^2\,\mathrm{d}F(y) = \mathsf{E}X^2. \qquad \square$$

1.3 Quantile function

If a distribution function is given it is often necessary to find its inverse function $F^{-1}(u)$. It is simple when $F(x)$ is increasing and continuous. In the general case so called *quantile function* F^{-1} is defined by the formula

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \qquad 0 < u < 1.$$

Values of the function $F^{-1}(u)$ are called *quantiles*. E.g. α-quantile is the value $F^{-1}(\alpha)$.

An important application of the quantile function is its role in construction of different distributions. It will be demonstrated in theorem 1.11.

The density of the rectangular distribution $\mathsf{R}(0,1)$ was introduced in section 1.2. Its distribution function is

$$R(u) = \begin{cases} 0 & \text{for } u \leq 0, \\ u & \text{for } 0 < u \leq 1, \\ 1 & \text{for } 1 < u. \end{cases} \tag{1.7}$$

The following theorem will be used in some proofs of forthcoming theorems.

Theorem 1.9 *Let a random variable X have the distribution function F. Then the relation $\mathsf{P}\{F(X) \leq F(x)\} = F(x)$ holds for every real x.*

Proof. We have

$$\{F(X) \leq F(x)\} = [\{F(X) \leq F(x)\} \cap \{X \leq x\}] \cup [\{F(X) \leq F(x)\} \cap \{X > x\}].$$

Since

$$\{X \leq x\} \subset \{F(X) \leq F(x)\}, \qquad \{X > x\} \cap \{F(X) < F(x)\} = \emptyset,$$

we get

$$\{F(X) \leq F(x)\} = \{X \leq x\} \cup [\{X > x\} \cap \{F(X) = F(x)\}]. \tag{1.8}$$

Notice that

$$\mathsf{P}[\{X > x\} \cap \{F(X) = F(x)\}] = 0,$$

since in this case X must be in an interval where F is constant. If we calculate probabilities of the left and right hand side of the formula (1.8), we get the assertion. $\square$

Theorem 1.10 *Let X have a continuous distribution function F. Then the random variable $U = F(X)$ has the distribution* $\mathsf{R}(0,1)$.

Proof. We prove that $\mathsf{P}(U \le u) = R(u)$, where the function $R(u)$ is defined in (1.7). If $u < 0$ or $1 \le u$, the assertion is obvious. Let $0 < u < 1$. Since F is continuous, for a given u there exists such an x that $F(x) = u$. Using theorem 1.9 we obtain $\mathsf{P}(U \le u) = \mathsf{P}\{F(X) \le F(x)\} = u$. □

If we assume in theorem 1.10 that F is increasing then the proof without using theorem 1.9 would follow from

$$\mathsf{P}(U \le u) = \mathsf{P}[F(X) \le u] = \mathsf{P}[X \le F^{-1}(u)] = F[F^{-1}(u)] = u.$$

Theorem 1.10 says that the rectangular distribution can be obtained using transformation $U = F(X)$ from any other distribution with continuous distribution function.

Theorem 1.11 *Let $U \sim \mathsf{R}(0,1)$ and let F be a distribution function. Then the random variable $X = F^{-1}(U)$ has the distribution function F.*

Proof. Let $u \in (0,1)$ and let x be a number such that $0 < F(x) < 1$. First we show thet the inequality $F(x) \ge u$ is fulfilled if and only if $x \ge F^{-1}(u)$. Assume that

$$x \ge F^{-1}(u) = \inf\{y : F(y) \ge u\}.$$

Since F is nondecreasing and continuous from the right the set $\{y : F(y) \ge u\}$ is an interval containing its left point. Thus we must have $F(x) \ge u$. Now assume that $F(x) \ge u$. But then $x \ge \inf\{y : F(y) \ge u\} = F^{-1}(u)$. Together we have $\mathsf{P}\{F^{-1}(U) \le x\} = \mathsf{P}\{U \le F(x)\} = F(x)$, which we wanted to prove. □

If in theorem 1.11 we add the assumption that F is continuous and increasing, the assertion would simply follow from the fact that

$$\mathsf{P}(X \le x) = \mathsf{P}[F^{-1}(U) \le x] = \mathsf{P}[U \le F(x)] = F(x).$$

The variables with the distribution $\mathsf{R}(0,1)$ can be easily generated by computers. Thus theorem 1.11 describes a procedure how to obtain variables with an arbitrary given distribution function F. In most cases the calculation if the quantile function is complicated and so to generate large number of such variables can be time consuming. Often instead of direct application of theorem 1.11 special procedures are used.

Let us remark that theorem 1.9 and proofs of the theorems 1.10 and 1.11 follow the paper Angus (1994).

At the end of this chapter we introduce a definition of median which belongs to the distribution with the distribution function F. *Median* $\tilde{\mu}$ is a number which satisfies

$$F(\tilde{\mu}^-) \le \frac{1}{2}, \qquad F(\tilde{\mu}) \ge \frac{1}{2}.$$

The symbol $F(\tilde{\mu}^-)$ denotes the limit from the left hand side. The median always exists but it is not defined uniquely. Similarly the number $\tilde{\mu}_p$ (called p-th *percentile*) for $0 < p < 1$ is defined by

$$F(\tilde{\mu}_p^-) \le p, \qquad F(\tilde{\mu}_p) \ge p.$$

Theorem 1.12 *Let a distribution with the distribution function F have expectation μ and a finite standard deviation σ. Then we have for (an arbitrary) median $\tilde{\mu}$*

$$|\mu - \tilde{\mu}| \leq \sigma$$

and for (an arbitrary) percentile $\tilde{\mu}_p$ we have

$$|\mu - \tilde{\mu}_p| \leq \sigma \max\left(\sqrt{\frac{1-p}{p}}, \sqrt{\frac{p}{1-p}}\right).$$

Proof. See O'Cinneide (1990). □

Chapter 2

Random vectors

2.1 Introduction

Let random variables $X_1, \ldots, X_n$ be defined on the same probability space $(\Omega, \mathcal{A}, \mathsf{P})$. Then $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is called *random vector*. If the random variables X_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, m)$ are defined on the same probability space, then $\boldsymbol{X} = (X_{ij})$ is *random matrix*. The expectation of the random vector is the vector $\mathsf{E}\boldsymbol{X} = (\mathsf{E}X_1, \ldots, \mathsf{E}X_n)'$; similarly we define expectation of the random matrix by $\mathsf{E}\boldsymbol{X} = (\mathsf{E}X_{ij})$. The function

$$F(x_1, \ldots, x_n) = \mathsf{P}(X_1 \le x_1, \ldots, X_n \le x_n)$$

is called *distribution function of the random vector* $\boldsymbol{X}$. Sometimes it is called *simultaneous distribution function* of the random variables $X_1, \ldots, X_n$. If there exists a function f such that

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n)\, \mathrm{d}u_1 \ldots \mathrm{d}u_n, \tag{2.1}$$

then we say that the vector $\boldsymbol{X}$ has *continuous distribution* and that f is its *simultaneous density*. In view of (2.1) we have

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n} \qquad \text{almost everywhere.} \tag{2.2}$$

If $g(x_1, \ldots, x_n)$ is borel measurable function, then it can be proved that

$$\mathsf{E}g(X_1, \ldots, X_n) = \int \cdots \int g(x_1, \ldots, x_n)\, \mathrm{d}F(x_1, \ldots, x_n). \tag{2.3}$$

If only some of the variables $X_1, \ldots, X_n$ are investigated then their distribution function is called *marginal* and eventual corresponding density is also *marginal*.

If there exists the simultaneous density then all the marginal densities also exist; the reversed assertion does not hold.

To illustrate the situation, we shall deal for a while only with two random variables X_1 and X_2. Their simultaneous density is $F(x_1, x_2) = \mathsf{P}(X_1 \le x_1, X_2 \le x_2)$. It is obvious that the marginal distribution function $F_1(x_1) = \mathsf{P}(X_1 \le x_1)$ of the random variable X_1 satisfies

$$F_1(x_1) = F(x_1, \infty),$$

where

$$F(x_1, \infty) = \lim_{x_2 \to \infty} F(x_1, x_2).$$

If X_1, X_2 have the simultaneous density $f(x_1, x_2)$, then using (2.1) we have

$$F(x_1, \infty) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(u_1, u_2)\, \mathrm{d}u_1\, \mathrm{d}u_2.$$

In view of (2.2) we have that the density $f_1(x_1)$ of the variable X_1 is

$$f_1(x_1) = \frac{\mathrm{d}F_1(x_1)}{\mathrm{d}x_1} = \frac{\mathrm{d}F(x_1, \infty)}{\mathrm{d}x_1} = \int_{-\infty}^{\infty} f(x_1, u_2)\, \mathrm{d}u_2.$$

Analogous formulas can be easily derived also for larger number of random variables. The result is formulated in the sentence that the marginal density can be obtained from the simultaneous density integrating superfluous variables.

2.2 Variance matrix

Let the variables $X_1, \ldots, X_n$ have finite second moments. *Covariance* $\mathsf{cov}(X_i, X_j)$ of random variables X_i, X_j is

$$\mathsf{cov}(X_i, X_j) = \mathsf{E}(X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j).$$

Equivalently, we obtain

$$\mathsf{cov}(X_i, X_j) = \mathsf{E}X_iX_j - \mathsf{E}X_i\, \mathsf{E}X_j.$$

This last expression is suitable for practical calculations. We can see that $\mathsf{cov}(X_i, X_i) = \mathsf{var}\, X_i$. Denote

$$F_{ij}(x_i, x_j) = \mathsf{P}(X_i \le x_i, X_j \le x_j)$$

the distribution function of the vector (X_i, X_j). Let $F_i(x) = \mathsf{P}(X_i \le x)$ for $i = 1, \ldots, n$ be the marginal distribution functions of the individual components of the vector $\boldsymbol{X}$. It follows from (2.3) that

$$\begin{aligned}\mathsf{cov}(X_i, X_j) &= \iint \left[x_i - \int x\, \mathrm{d}F_i(x)\right] \left[x_j - \int x\, \mathrm{d}F_j(x)\right] \mathrm{d}F_{ij}(x_i, x_j) \\ &= \iint x_i x_j\, \mathrm{d}F_{ij}(x_i, x_j) - \left[\int x\, \mathrm{d}F_i(x)\right] \left[\int x\, dF_j(x)\right].\end{aligned}$$

If a two-dimensional vector (X_i, X_j) has a continuous distribution with the density $f_{ij}(x_i, x_j)$then

$$\iint x_i x_j\, \mathrm{d}F_{ij}(x_i, x_j) = \iint x_i x_j f_{ij}(x_i, x_j)\, \mathrm{d}x_i\, \mathrm{d}x_j.$$

Instead of $\mathsf{cov}(X_i, X_j)$ it is written $\sigma_{X_iX_j}$ or shortly σ_{ij}. We know already that $\sigma_{ii} = \sigma_i^2$ is the variance of the variable X_i. If we write $\mathsf{cov}(X_i, X_j)$ as the elements of the matrix, we get the *variance matrix* $\boldsymbol{V} = (\sigma_{ij})$. The matrix $\boldsymbol{V}$ is often written as $\mathsf{var}\, \boldsymbol{X}$.

Theorem 2.1 *The variance matrix can be written in the form*

$$\boldsymbol{V} = \mathsf{E}(\boldsymbol{X} - \mathsf{E}\boldsymbol{X})(\boldsymbol{X} - \mathsf{E}\boldsymbol{X})' \tag{2.4}$$

and in the form

$$\boldsymbol{V} = \mathsf{E}\boldsymbol{X}\boldsymbol{X}' - (\mathsf{E}\boldsymbol{X})(\mathsf{E}\boldsymbol{X})'.$$

Proof. The first formula can be proved componentwise. The second formula follows from the first one by an easy derivation. □

Theorem 2.2 *Let $\boldsymbol{a}$ be an $m \times 1$ vector and $\boldsymbol{B}$ an $m \times n$ matrix. Define $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$. If $\mathsf{E}\boldsymbol{X}$ exists then $\mathsf{E}\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\mathsf{E}\boldsymbol{X}$. If $\boldsymbol{X}$ has components with finite second moments, then*

$$\mathsf{var}\,\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'. \tag{2.5}$$

Proof. The first assertion can be proved componentwise. The second one follows from (2.4). □

It follows from definition as well as from theorem 2.2 that the variance matrix is symmetric.

Theorem 2.3 *Variance matrix is positive semidefinite.*

Proof. Choose an arbitrary vector $\boldsymbol{c} = (c_1, \ldots, c_n)'$. Since the variance of any random variable is nonnegative also the variance of $\boldsymbol{c}'\boldsymbol{X}$ is nonnegative. Using formula (2.5) we obtain

$$0 \leq \mathsf{var}\,\boldsymbol{c}'\boldsymbol{X} = \boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}. \qquad \square$$

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ be random vectors with finite second moments. Then *the covariance matrix* of the vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is the matrix

$$\mathsf{cov}(\boldsymbol{Y}, \boldsymbol{Z}) = (\mathsf{cov}(Y_i, Z_j)).$$

It is easy to check that

$$\mathsf{cov}(\boldsymbol{Y}, \boldsymbol{Z}) = \mathsf{E}(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})(\boldsymbol{Z} - \mathsf{E}\boldsymbol{Z})'$$

and that

$$\mathsf{cov}(\boldsymbol{Y}, \boldsymbol{Z}) = \mathsf{E}\boldsymbol{Y}\boldsymbol{Z}' - (\mathsf{E}\boldsymbol{Y})(\mathsf{E}\boldsymbol{Z})'.$$

From the definition it follows that

$$\mathsf{cov}(\boldsymbol{Z}, \boldsymbol{Y}) = [\mathsf{cov}(\boldsymbol{Y}, \boldsymbol{Z})]'.$$

Notice that the variance matrix is a special case of the covariance matrix, since

$$\mathsf{cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathsf{var}\,\boldsymbol{X}.$$

The covariance matrices are used in the following theorem.

Theorem 2.4 *Let random vectors $\boldsymbol{X} = (X_1, \ldots, X_n)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ have finite second moments. Then*

$$\mathsf{var}(\boldsymbol{X} + \boldsymbol{Y}) = \mathsf{var}\,\boldsymbol{X} + \mathsf{cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathsf{cov}(\boldsymbol{Y}, \boldsymbol{X}) + \mathsf{var}\,\boldsymbol{Y}.$$

Proof. The assertion follows from formula (2.4). □

Theorem 2.5 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ be random vectors with finite second moments. If $\boldsymbol{a}$ is an $r \times 1$ vector, $\boldsymbol{c}$ an $s \times 1$ vector, $\boldsymbol{B}$ an $r \times m$ matrix and $\boldsymbol{D}$ an $s \times n$ matrix, then*

$$\mathsf{cov}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{Y}, \boldsymbol{c} + \boldsymbol{D}\boldsymbol{Z}) = \boldsymbol{B}\,\mathsf{cov}(\boldsymbol{Y}, \boldsymbol{Z})\,\boldsymbol{D}'.$$

Proof. The result follows from the formula for covariance. □

2.3 Independence

Let us return to the model for random events which was introduced in section 2.1. We say that two events $A_1 \in \mathcal{A},\ A_2 \in \mathcal{A}$ are *independent*, if

$$\mathsf{P}(A_1 \cap A_2) = \mathsf{P}(A_1)P(A_2).$$

This is a mathematical definition of the notion independence which is used in everyday language. If $\mathsf{P}(A_2) > 0$ then we define the conditional probability of the event A_1 given A_2 as

$$\mathsf{P}(A_1|A_2) = \frac{\mathsf{P}(A_1 \cap A_2)}{\mathsf{P}(A_2)}. \tag{2.6}$$

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Let $F(x_1, \ldots, x_n)$ be its simultaneous distribution function

Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be a random vector, $F(x_1, \ldots, x_n)$ its simultaneous distribution function and $F_i(x_i)$ marginal distribution of X_i, $i = 1, \ldots, n$. We say that $X_1, \ldots, X_n$ are *independent random variables*, if we have for all real $x_1, \ldots, x_n$

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2)\ldots F_n(x_n). \tag{2.7}$$

Theorem 2.6 *Let a random vector $\boldsymbol{X}$ have simultaneous density $f(x_1, \ldots, x_n)$. Denote $f_i(x_i)$ the marginal density of the variable X_i, $i = 1, \ldots, n$. Then variables $X_1, \ldots, X_n$ are independent if and only if*

$$f(x_1, \ldots, x_n) = f_1(x_1)\ldots f_n(x_n) \qquad \textit{almost everywhere.}$$

Proof. The assertion follows from definition (2.7) and from the formula (2.2). □

The independence of random vectors is defined similarly as the independence of random variables. We say that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are *independent random vectors*, if their simultaneous distribution function is equal to the product of the distribution function of vector $\boldsymbol{Y}$ and the distribution function of vector $\boldsymbol{Z}$. If the simultaneous distribution of the vector $(\boldsymbol{Y}', \boldsymbol{Z}')'$ is continuous then the vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent if and only if their simultaneous density is equal to the product of marginal densities of the vectors $\boldsymbol{Y}$ a $\boldsymbol{Z}$.

Theorem 2.7 *Let $X_1, \dots, X_n$ be independent random variables with finite first moments. Then*

$$\mathsf{E}(X_1 \dots X_n) = (\mathsf{E}X_1) \dots (\mathsf{E}X_n).$$

Proof. Let F be the simultaneous distribution function of the vector $\boldsymbol{X} = (X_1, \dots, X_n)'$ and F_i the distribution function of the random variable X_i, $i = 1, \dots, n$. It follows from (2.3), (2.7) and from Fubini theorem that

$$\begin{aligned}
\mathsf{E}(X_1 \dots X_n) &= \int \cdots \int x_1 \dots x_n \,\mathrm{d}F(x_1, \dots, x_n) \\
&= \int \cdots \int x_1 \dots x_n \,\mathrm{d}F_1(x_1) \dots \,\mathrm{d}F_n(x_n) \\
&= \left[\int x_1 \,\mathrm{d}F_1(x_1)\right] \dots \left[\int x_n \,\mathrm{d}F_n(x_n)\right] = (\mathsf{E}X_1) \dots (\mathsf{E}X_n). \qquad \square
\end{aligned}$$

A statistician must frequently decide if two random variables are independent or not. Their simultaneous distribution function is usually not known. Some procedures are based on the following assertion.

Theorem 2.8 *Let X and Y be independent random variables with finite second moments. Then we have* $\mathsf{cov}(X, Y) = 0$.

Proof. If X and Y are independent, then the variables $X - \mathsf{E}X$ and $Y - \mathsf{E}Y$ are also independent. From the definition of covariance and from theorem 2.7 we obtain

$$\mathsf{cov}(X, Y) = \mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y) = \mathsf{E}(X - \mathsf{E}X)\mathsf{E}(Y - \mathsf{E}Y) = 0. \qquad \square$$

The covariance itself as a measure of independence is not used but after some normalization we get *correlation coefficient*. We shall consider it in section 2.6.

If $\mathsf{cov}(X, Y) = 0$, then we say that X and Y are *uncorrelated variables*. Unfortunately, examples show that uncorrelated variables may not be necessarily independent.

2.4 Conditional density

The main idea will be demonstrated on a random vector $\boldsymbol{X} = (Y, Z)'$ with two components. Let $\boldsymbol{X}$ have the density $p(y, z)$. Denote $q(z) = \int p(y, z)\,\mathrm{d}y$ the marginal density of the variable Z. First we define the *conditional distribution function* of the variable Y given $Z = z$, which we denote $F(y \,|\, z) = \mathsf{P}(Y \le y \,|\, Z = z)$. We cannot use formula (2.6), because $\mathsf{P}(Z = z) = 0$ since Z has a continuous distribution. Choose points z_1, z_2 in such a way that $z_1 < z < z_2$ and assume that $\mathsf{P}(z_1 < Z < z_2) > 0$ for such points z_1, z_2. Using(2.6) we get

$$\begin{aligned}
\mathsf{P}(Y \le y \,|\, z_1 < Z < z_2) &= \frac{\mathsf{P}(Y \le y,\, z_1 < Z < z_2)}{\mathsf{P}(z_1 < Z < z_2)} \\
&= \frac{\int_{-\infty}^{y} \int_{z_1}^{z_2} p(u, v)\,\mathrm{d}v\,\mathrm{d}u}{\int_{z_1}^{z_2} q(v)\,\mathrm{d}v}.
\end{aligned}$$

If p is a smooth function then according to the mean value theorem for every u there exists a point $z_u \in (z_1, z_2)$ such that

$$\int_{z_1}^{z_2} p(u,v)\,\mathrm{d}v = (z_2 - z_1)p(u, z_u).$$

If q is a smooth function, then analogously there exists a point $z^* \in (z_1, z_2)$ such that

$$\int_{z_1}^{z_2} q(v)\,\mathrm{d}v = (z_2 - z_1)q(z^*).$$

Thus

$$\mathsf{P}(Y \le y \,|\, z_1 < Z < z_2) = \frac{\int_{-\infty}^{y} p(u, z_u)\,\mathrm{d}u}{q(z^*)}.$$

If $z_1 \to z^-$, $z_2 \to z^+$, then also $z_u \to z$, $z^* \to z$ and because of assumed smoothness of functions p and q we will have $p(u, z_u) \to p(u, z)$, $q(z^*) \to q(z)$. If we may change limit and integral, we obtain

$$\lim_{\substack{z_1 \to z^- \\ z_2 \to z^+}} \mathsf{P}(Y \le y \,|\, z_1 < Z < z_2) = \int_{-\infty}^{y} \frac{p(u,z)\,\mathrm{d}u}{q(z)}.$$

It is natural to define the conditional distribution function $F(y \,|\, z)$ as the limit of the probability $\mathsf{P}(Y \le y \,|\, z_1 < Z < z_2)$ when $z_1 \to z^-$, $z_2 \to z^+$. Thus we obtained

$$F(y \,|\, z) = \int_{-\infty}^{y} \frac{p(u,z)}{q(z)}\,\mathrm{d}u.$$

Now we can introduce the *conditional density* as the derivative of the conditional distribution function analogously as in formula (1.4), which describes connection between the common distribution function and the common density. The result is that the conditional density r of the variable Y given $Z = z$ is

$$r(y \,|\, z) = \frac{p(y,z)}{q(z)}. \tag{2.8}$$

In our derivation it was necessary to use many mathematical assumptions. However, it can be proved (see Anděl 2007), that conditional density is always given by formula (2.8), if $q(z) \neq 0$. In the case $q(z) = 0$ we define $r(y \,|\, z) = 0$.

2.5 Approximation of random variables

Consider a random variable Y and a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Assume that Y and $\boldsymbol{X}$ have finite second moments. It happens that it is necessary to predict the variable Y using the known vector $\boldsymbol{X}$. We can have several reasons for it. Sometimes the measurement of Y is very difficult, hardly accessible or expensive. In other cases we obtain information about Y after many years whereas the vector $\boldsymbol{X}$ is known immediately.

We restrict ourselves to find an approximation of the variable Y using a linear function $\hat{Y} = \alpha + \beta_1 X_1 + \cdots + \beta_n X_n$. Our problem is to find the coefficients $\alpha, \beta_1, \ldots, \beta_n$ so that the difference between Y and $\hat{Y}$ is as small as possible. The criterion is $\mathsf{E}(Y - \hat{Y})^2$.

Theorem 2.9 *Let $\boldsymbol{V} = var\ \boldsymbol{X}$ be a regular matrix. Then*

$$\mathsf{E}(Y - \hat{Y})^2 \geq \mathsf{var}\, Y - \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)$$

and left hand side is equal to right hand side if and only if

$$\boldsymbol{\beta} = \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y), \qquad \alpha = \mathsf{E}Y - \boldsymbol{\beta}'\mathsf{E}\boldsymbol{X}. \tag{2.9}$$

Proof. If a random variable Z has a finite second moment, it fulfills $\mathsf{E}Z^2 = \mathsf{var}\, Z + (\mathsf{E}Z)^2$. Thus $\mathsf{E}Z^2 \geq \mathsf{var}\, Z$ and we have equality if and only if $\mathsf{E}Z = 0$. Define $Z = Y - \hat{Y} = Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}$. Then

$$\mathsf{E}(Y - \hat{Y})^2 \geq \mathsf{var}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}).$$

The equality holds if and only if $\mathsf{E}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}) = 0$, i. e. in the case

$$\alpha = \mathsf{E}Y - \boldsymbol{\beta}'\mathsf{E}\boldsymbol{X}.$$

We know that the variance does not depend on the shift, and so

$$\mathsf{var}(Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}) = \mathsf{var}(Y - \boldsymbol{\beta}'\boldsymbol{X}).$$

Using theorems 2.4 and 2.5 we get

$$\begin{aligned}\mathsf{var}(Y - \boldsymbol{\beta}'\boldsymbol{X}) &= \mathsf{var}\, Y - \boldsymbol{\beta}'\mathsf{cov}(\boldsymbol{X}, Y) - \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta} \\ &= [\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)]'\boldsymbol{V}[\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)] \\ &\quad + \mathsf{var}\, Y - \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y).\end{aligned}$$

Since $\boldsymbol{V}$ is a variance matrix, it is positively semidefinite according to theorem 2.3. Moreover, now we assume that $\boldsymbol{V}$ is regular. Thus $\boldsymbol{V}$ is positively definite and

$$[\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)]'\boldsymbol{V}[\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)] \geq 0.$$

The equality holds if and only if

$$\boldsymbol{\beta} - \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y) = 0.$$

The assertion is proved. □

The expression

$$\sigma^2_{Y.\boldsymbol{X}} = \mathsf{var}\, Y - \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Y)$$

is called *residual variance*. In view of (2.9) it also holds

$$\sigma^2_{Y.\boldsymbol{X}} = \mathsf{var}\, Y - \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}. \tag{2.10}$$

If we apply the formula for determinant of the matrix divided into the blocks we obtain another formula

$$\sigma^2_{Y.\boldsymbol{X}} = \frac{|\mathsf{var}(Y, X_1, \ldots, X_n)|}{|\boldsymbol{V}|}.$$

2.6 Correlation coefficient

Let X and Y be random variables with finite second moments. We denote $\sigma_X^2 = \mathsf{var}\, X$, $\sigma_Y^2 = \mathsf{var}\, Y$, $\sigma_{XY} = \mathsf{cov}(X, Y)$. If $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$, then we define *correlation coefficient*

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{2.11}$$

Instead of ρ one writes ρ_{XY} to indicate the variables for which the coefficient is calculated. It follows from theorem 2.8 that $\rho = 0$ when the variables X and Y are independent. It can be easily verified that

$$\rho_{XX} = 1. \tag{2.12}$$

Theorem 2.10 *Let a, b, c, d be such numbers that $bd \neq 0$. If $bd > 0$, then $\rho_{a+bX,c+dY} = \rho_{XY}$; if $bd < 0$, then $\rho_{a+bX,c+dY} = -\rho_{XY}$.*

Proof. The assertion follows from theorem (2.11). □

Theorem 2.11 *We have $-1 \leq \rho \leq 1$. If $b > 0$ then the equality $\rho = 1$ holds if and only if $Y = a + bX$ with probability 1. Similarly, if $b < 0$ then $\rho = -1$ holds if and only if $Y = a + bX$ with probability 1.*

Proof. From Schwarz inequality we get

$$|\mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y)|^2 \leq \mathsf{E}(X - \mathsf{E}X)^2 \mathsf{E}(Y - \mathsf{E}Y)^2.$$

It implies that $-1 \leq \rho \leq 1$. The equality is reached either in the case that $X - \mathsf{E}X = 0$ with probability 1, or in the case that $Y - \mathsf{E}Y = b(X - \mathsf{E}X)$ with probability 1. The first case cannot be realized since the variable X would have vanishing variance and the correlation coefficient would not be defined. From the same reason in the second case the value $b = 0$ must be excluded. If $b > 0$, then theorem 2.11 gives $\rho = 1$; if $b < 0$, then $\rho = -1$. □

The correlation matrix is introduced similarly as the variance and covariance matrices. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ with finite second moments and positive variances. Then the *correlation matrix* of the vector $\boldsymbol{X}$ is the matrix $\boldsymbol{P} = (\rho_{ij})$ of type $n \times n$, where $\rho_{ij} = \rho_{X_i, X_j}$. It follows from (2.12) that the matrix $\boldsymbol{P}$ has ones on the diagonal and is symmetric.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ be random vectors with finite second moments and positive variances. Then the *correlation matrix* of these vectors is $\mathsf{cor}(\boldsymbol{Y}, \boldsymbol{Z}) = (\rho_{Y_i Z_j})$ and has the type $m \times n$.

2.7 Coefficient of multiple correlation

Classical correlation coefficient ρ measures dependence between two random variables. Very often it is necessary to describe dependence between a random variable Y and a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Let $\boldsymbol{V} = \mathsf{var}\, \boldsymbol{X}$ be a regular matrix. According theorem 2.9 the variable $\hat{Y} = \alpha + \boldsymbol{\beta}' \boldsymbol{X}$ is the best linear approximation. Here α a $\boldsymbol{\beta}$ are

introduced in formula (2.9). The *coefficient of multiple correlation* $\rho_{Y,\boldsymbol{X}}$ is defined as the usual correlation coefficient between the variables Y and $\hat{Y}$. The definition is

$$\rho_{Y,\boldsymbol{X}} = \rho_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}}.$$

In the case $\boldsymbol{\beta} = \mathbf{0}$ we define $\rho_{Y,\boldsymbol{X}} = 0$.

Theorem 2.12 *Coefficient of multiple correlation is given by the formula*

$$\rho^2_{Y,\boldsymbol{X}} = \frac{\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}{\sigma^2_Y}. \tag{2.13}$$

Proof. If $\boldsymbol{\beta} = \mathbf{0}$, the assertion is valid. Let $\boldsymbol{\beta} \neq \mathbf{0}$. According to the definition we have

$$\rho^2_{Y,\boldsymbol{X}} = \frac{[\mathsf{cov}(Y, \alpha + \boldsymbol{\beta}'\boldsymbol{X})]^2}{\sigma^2_Y \,\mathsf{var}(\alpha + \boldsymbol{\beta}'\boldsymbol{X})}.$$

It follows from (2.9) that $\mathsf{cov}(\boldsymbol{X}, Y) = \boldsymbol{V}\boldsymbol{\beta}$, and thus

$$\mathsf{cov}(Y, \boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{V}. \tag{2.14}$$

This gives

$$\mathsf{cov}(Y, \alpha + \boldsymbol{\beta}'\boldsymbol{X}) = \mathsf{cov}(Y, \boldsymbol{\beta}'\boldsymbol{X}) = \mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta} \tag{2.15}$$

and similarly

$$\mathsf{var}(\alpha + \boldsymbol{\beta}'\boldsymbol{X}) = \boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}. \tag{2.16}$$

The assertion is proved. □

Theorem 2.13 *Coefficient of multiple correlation satisfies* $0 \leq \rho_{Y,\boldsymbol{X}} \leq 1$.

Proof. In the case $\boldsymbol{\beta} = \mathbf{0}$ the assertion is valid. Otherwise $\rho_{Y,\boldsymbol{X}}$ is equal to the correlation coefficient between Y a $\hat{Y}$, which is smaller or equal 1 according to theorem 2.11. For $\boldsymbol{\beta} \neq \mathbf{0}$ from (2.15) and (2.16) we get

$$\rho_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}} = \frac{\mathsf{cov}(Y, \alpha + \boldsymbol{\beta}'\boldsymbol{X})}{\sqrt{\sigma^2_Y \,\mathsf{var}(\alpha + \boldsymbol{\beta}'\boldsymbol{X})}} = \frac{\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}{\sqrt{\sigma^2_Y \,\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}}.$$

This expression is positive since $\boldsymbol{V}$ is the positive definite matrix. □

Theorem 2.14 *Coefficient of multiple correlation satisfies* $\rho_{Y,\boldsymbol{X}}$ *and residual variance* $\sigma^2_{Y.\boldsymbol{X}}$ *satisfy the relation*

$$\rho^2_{Y,\boldsymbol{X}} = 1 - \frac{\sigma^2_{Y.\boldsymbol{X}}}{\sigma^2_Y}. \tag{2.17}$$

Proof. Assertion follows from formulas (2.10) a (2.13). □

Formula (2.17) can be also written in the form

$$\sigma^2_{Y.\boldsymbol{X}} = \sigma^2_Y(1 - \rho^2_{Y,\boldsymbol{X}}).$$

Here we can see that for increasing value of the coefficient of multiple correlation the residual variance $\sigma^2_{Y.\boldsymbol{X}}$ decreases.

Theorem 2.15 *Coefficient of multiple correlation $\rho_{Y,\boldsymbol{X}}$ is the largest one from all correlation coefficients between Y and an arbitrary linear function of the vector $\boldsymbol{X}$, which is not constant.*

Proof. The assertion is obvious if $\boldsymbol{\beta} = \mathbf{0}$; in this case $\rho_{Y,\boldsymbol{X}} = 0$ and also $\rho_{Y,a+\boldsymbol{b}'\boldsymbol{X}} = 0$ for an arbitrary vector $\boldsymbol{b} \neq \mathbf{0}$. In the case $\boldsymbol{b} = \mathbf{0}$ the correlation coefficient is not defined. So assume that $\boldsymbol{\beta} \neq \mathbf{0}$. We prove that then $\rho^2_{Y,a+\boldsymbol{b}'\boldsymbol{X}} \leq \rho^2_{Y,\alpha+\boldsymbol{\beta}'\boldsymbol{X}}$ for an arbitrary vector $\boldsymbol{b} \neq \mathbf{0}$. Since numbers a and α have no influence on the correlation coefficients it suffices to prove the inequality $\rho^2_{Y,\boldsymbol{b}'\boldsymbol{X}} \leq \rho^2_{Y,\boldsymbol{\beta}'\boldsymbol{X}}$. We have

$$\rho^2_{Y,\boldsymbol{b}'\boldsymbol{X}} = \frac{[\mathsf{cov}(Y, \boldsymbol{b}'\boldsymbol{X})]^2}{\sigma^2_Y \ \mathsf{var}\, \boldsymbol{b}'\boldsymbol{X}} = \frac{[\mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{b}]^2}{\sigma^2_Y \ \boldsymbol{b}'\boldsymbol{V}\boldsymbol{b}}.$$

From (2.14) we have

$$[\mathsf{cov}(Y, \boldsymbol{X})\boldsymbol{b}]^2 = (\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{b})^2.$$

Matrix $\boldsymbol{V}$ is positive definite. Thus there exists a matrix $\boldsymbol{V}^{1/2}$, which is symmetric, positive definite and which fulfills $\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2} = \boldsymbol{V}$. Using Schwarz inequality we get

$$\begin{aligned}(\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{b})^2 &= (\boldsymbol{\beta}'\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}\boldsymbol{b})^2 = [(\boldsymbol{V}^{1/2}\boldsymbol{\beta})'(\boldsymbol{V}^{1/2}\boldsymbol{b})]^2 \\ &=\leq (\boldsymbol{V}^{1/2}\boldsymbol{\beta})'(\boldsymbol{V}^{1/2}\boldsymbol{\beta}) \cdot (\boldsymbol{V}^{1/2}\boldsymbol{b})'(\boldsymbol{V}^{1/2}\boldsymbol{b}) = (\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta})(\boldsymbol{b}'\boldsymbol{V}\boldsymbol{b}).\end{aligned}$$

Thus

$$\rho^2_{Y,\boldsymbol{b}'\boldsymbol{X}} \leq \frac{(\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta})(\boldsymbol{b}'\boldsymbol{V}\boldsymbol{b})}{\sigma^2_Y \ \boldsymbol{b}'\boldsymbol{V}\boldsymbol{b}} = \frac{\boldsymbol{\beta}'\boldsymbol{V}\boldsymbol{\beta}}{\sigma^2_Y} = \rho^2_{Y,\boldsymbol{X}}.$$

The last equality is ensured by formula (2.13). □

Theorem 2.15 guarantees that the coefficient $\rho_{Y,\boldsymbol{X}}$ is never smaller than the absolute value of any correlation coefficient ρ_{Y,X_i}, $i = 1, \ldots, n$. This result is sometimes used for checking the calculations.

Theorem 2.16 *We have*

$$\rho^2_{Y,\boldsymbol{X}} = \mathsf{cor}(Y, \boldsymbol{X})\, \boldsymbol{P}^{-1} \mathsf{cor}(\boldsymbol{X}, Y), \tag{2.18}$$

where $\boldsymbol{P} = \mathsf{cor}\, \boldsymbol{X}$.

Proof. The assertion follows from formula (2.13). For details see Anděl (2007). □

Very often we have the case when the vector $\boldsymbol{X}$ has two components. To simplify notation we write $X_0 = Y$. Then ρ_{ij}, $i, j = 0, 1, 2$ is the correlation coefficient between X_i and X_j and instead of $\rho_{Y,\boldsymbol{X}}$ we write $\rho_{0.1,2}$. Inserting into (2.18) one gets

$$\rho^2_{0.1,2} = \frac{\rho^2_{01} + \rho^2_{02} - 2\rho_{01}\rho_{02}\rho_{12}}{1 - \rho^2_{12}}. \tag{2.19}$$

The coefficient $\rho_{0.1,2}$ measures the total dependence of the variable Y on the complete vector $(X_1, X_2)'$. This dependence can be very large even when the dependence of Y on every component X_1, X_2 is quite small. We demonstrate it on an example.

Example 2.17 Let $\rho_{01} = 0$ and $\rho_{02} \neq 0$. Then (2.19) gives

$$\rho^2_{0.1,2} = \frac{\rho^2_{02}}{1 - \rho^2_{12}}.$$

Theorem 2.13 ensures that the fraction $\rho^2_{02}/(1-\rho^2_{12})$ is never larger than 1. But the number $\rho^2_{0.1,2}$ can be arbitrary close to one if ρ^2_{12} is sufficiently large. The explanation can be the following one. Let Y and X_1 be independent variables. Define $X_2 = Y - X_1$. Denote $\mathsf{var}\, Y = \sigma_0^2$, $\mathsf{var}\, X_1 = \sigma_1^2$. An easy calculation gives $\rho_{01} = 0$, $\rho_{02} = (\sigma_0^2 + \sigma_1^2)^{-1/2}\sigma_0$, $\rho_{12} = -(\sigma_0^2 + \sigma_1^2)^{-1/2}\sigma_1$, $\rho^2_{0.1,2} = 1$. Choosing σ_0 sufficiently small and σ_1 sufficiently large, the correlation coefficient ρ_{02} will have arbitrary small value. The variable Y is not correlated with X_1, its correlation with X_2 is very small, but both the variables simultaneously define it uniquely since $Y = X_1 + X_2$. $\diamondsuit$

2.8 Coefficient of partial correlation

Coefficient of partial correlation measures dependence of two random variables. The dependence may not be causal, it can arise under the influence of other confounding factors. Consider two random variables Y and Z and a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Assume that $\boldsymbol{X}$ may have influence on both Y and Z. We want to ask how large would be dependence between Y and Z without influence of $\boldsymbol{X}$. One possibility how to investigate it is to create such situation that the vector $\boldsymbol{X}$ remains constant. In most cases it is not possible and so it is necessary to find a mathematical solution.

We know from theorem 2.9 that the best linear approximation of the variable Y is $\hat{Y} = \alpha + \boldsymbol{\beta}'\boldsymbol{X}$, where α and $\boldsymbol{\beta}$ are introduced in (2.9). From this reason we can interprete $Y - \hat{Y}$ as such a part of variable Y which is cleared from influence of the vector $\boldsymbol{X}$. Similarly, let $\hat{Z} = \gamma + \boldsymbol{\delta}'\boldsymbol{X}$ be the best linear approximation of variable Z based on $\boldsymbol{X}$. It follows from (2.9) that

$$\boldsymbol{\delta} = \boldsymbol{V}^{-1}\mathsf{cov}(\boldsymbol{X}, Z), \quad \gamma = \mathsf{E}Z - \boldsymbol{\delta}'\mathsf{E}\boldsymbol{X}. \tag{2.20}$$

The part of variable Z which is not explained by the vector $\boldsymbol{X}$ can be interpreted as $Z - \hat{Z}$. From this reason the dependence between Y a Z after elimination of influence of vector $\boldsymbol{X}$ is measured by the correlation coefficient between $Y - \hat{Y}$ and $Z - \hat{Z}$. It is called *partial correlation coefficient* between Y and Z given $\boldsymbol{X}$. It is denoted by $\rho_{Y,Z.\boldsymbol{X}}$.

Theorem 2.18 *Let random variables $Y, Z, X_1, \ldots, X_n$ have finite second moments and the regular variance matrix. Denote $\boldsymbol{P} = \mathsf{cor}\,\boldsymbol{X}$. Then it holds*

$$\rho_{Y,Z.\boldsymbol{X}} = \frac{\rho_{YZ} - \mathsf{cor}(Y, \boldsymbol{X})\,\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Z)}{\sqrt{[1 - \mathsf{cor}(Y, \boldsymbol{X})\,\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Y)][1 - \mathsf{cor}(Z, \boldsymbol{X})\,\boldsymbol{P}^{-1}\mathsf{cor}(\boldsymbol{X}, Z)]}}.$$

Proof. Theorem follows from the formula for the correlation coefficient between $Y - \alpha - \boldsymbol{\beta}'\boldsymbol{X}$ and $Z - \gamma - \boldsymbol{\delta}'\boldsymbol{X}$. Using (2.9) and (2.20) we obtain the result. $\square$

For the coefficient of partial correlation we have no such inequalities like in case of the correlation of multiple correlation. From this reason $\rho_{Y,Z.\boldsymbol{X}}$ may be sometimes smaller and sometimes greater than ρ_{YZ}.

Let $n = 1$, so that the vector $\boldsymbol{X}$ has only one component. Then we get from theorem 2.18 that

$$\rho_{Y,Z.X} = \frac{\rho_{YZ} - \rho_{XY}\rho_{XZ}}{\sqrt{(1-\rho_{XY}^2)(1-\rho_{XZ}^2)}}.$$

Example 2.19 In England investigated how the harvest of hay (variable Y) depends on temperature (variable Z, which is equal to the sum of temperaturs larger than 42 °F $\doteq$ 5,6 °C in individual days of the considered spring season). It was found that $\rho_{YZ} = -0.40$. The negative correlation does not correspond our experience that the gras grows faster if the weather is warmer. Then the calculation included precipitation (variable X) and the other correlation coefficients were calculated. The result was $\rho_{XY} = 0.80$, $\rho_{XZ} = -0.56$. If it rains, more gras grows, but it is colder. The result is $\rho_{Y,Z.X} = 0.10$. The details can be found in Hooker (1907) and in Yule, Kendall (1950). □

2.9 Multinomial distribution

Assume that in an experiment one of the events $A_1, \dots, A_k$ can be realized. Let the events be disjoint and one of them must be realized. We denote their probabilities $p_i = \mathsf{P}(A_i)$, $i = 1, \dots, k$. Our assumptions imply that $p_1 + \cdots + p_k = 1$. Moreover, let all probabilities p_i be positive.

Consider the case that the experiments are n-times independently repeated. Let X_i be number of realizations of the event A_i in such series of experiments. Then we have

$$\mathsf{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!}\, p_1^{x_1} \dots p_k^{x_k} \tag{2.21}$$

for nonnegative integers $x_1, \dots, x_k$ the sum of which equals to n. In any other case this probability is zero. The distribution defined by formula (2.21) is called *multinomial*.

In the special case $k = 2$ we can the event A_1 call the success and the event A_2 failure. Then (2.21) will be the *binomial distribution*. However, this situation can arise also for general k, if we introduce events $A = A_1$ and $B = A_2 \cup \cdots \cup A_k$. Thus the marginal distribution of the random variable X_1 is $\mathsf{Bi}(n, p_1)$ and similarly $X_i \sim \mathsf{Bi}(n, p_i)$ also for other values i.

Theorem 2.20 *In multinomial distribution we have*

$$\mathsf{E}X_i = np_i, \quad \mathsf{var}\, X_i = np_i(1-p_i) \quad \textit{for } i = 1, \dots, k \tag{2.22}$$

and

$$\mathsf{cov}(X_i, X_j) = -np_ip_j \quad \textit{for } i \neq j. \tag{2.23}$$

Proof. Formula (2.22) follows from $X_i \sim \mathsf{Bi}(n, p_i)$. We are going to prove (2.23). We introduce random variables ξ_{hi} in the following way. Define $\xi_{hi} = 1$ if in the h-th experiment the event A_i appeared and $\xi_{hi} = 0$ in other cases. The experiments are independent and so

$$\mathsf{cov}(\xi_{hi}, \xi_{mj}) = 0 \quad \text{for } h \neq m. \tag{2.24}$$

We can see that

$$\xi_{hi}\xi_{hj} = 0 \quad \text{for } i \neq j, \tag{2.25}$$

since in the h-th experiment the events A_i and A_j could not appear simultaneously. An easy calculation gives $\mathsf{E}\xi_{hi} = p_i$. Since

$$X_i = \sum_{h=1}^{n} \xi_{hi}, \tag{2.26}$$

using (2.26), (2.24) and (2.25) for $i \neq j$ we get

$$\begin{aligned}\mathsf{cov}(X_i, X_j) &= \mathsf{cov}\left(\sum_{h=1}^{n}\xi_{hi}, \sum_{m=1}^{n}\xi_{mj}\right) = \sum_{h=1}^{n}\sum_{m=1}^{n}\mathsf{cov}(\xi_{hi}, \xi_{mj}) \\ &= \sum_{h=1}^{n}\mathsf{cov}(\xi_{hi}, \xi_{hj}) = \sum_{h=1}^{n}(\mathsf{E}\xi_{hi}\xi_{hj} - \mathsf{E}\xi_{hi}\mathsf{E}\xi_{hj}) \\ &= -\sum_{h=1}^{n} p_i p_j = -np_ip_j. \qquad \square\end{aligned}$$

A matrix $\boldsymbol{A}$ is called *idempotent* if it is a square matrix and if $\boldsymbol{A}^2 = \boldsymbol{A}$ holds.

Theorem 2.21 *Denote*

$$\boldsymbol{u} = (\sqrt{p_1}, \ldots, \sqrt{p_k})', \qquad \boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}', \qquad \boldsymbol{D} = \mathsf{Diag}\{\sqrt{np_1}, \ldots, \sqrt{np_k}\}.$$

Then the matrix $\boldsymbol{Q}$ is idempotent and its rank is $k-1$. Variance matrix $\boldsymbol{V}$ of the random vector $\boldsymbol{X} = (X_1, \ldots, X_k)'$ with multinomial distribution is $\boldsymbol{V} = \boldsymbol{D}\boldsymbol{Q}\boldsymbol{D}$.

Proof. It can be easily checked that the matrix $\boldsymbol{Q}$ is idempotent. The rank of the matrix $\boldsymbol{Q}$ is equal to its trace and it is $k-1$. Elements of the matrix $\boldsymbol{V}$ were calculated in Theorem 2.20, and thus the relation $\boldsymbol{V} = \boldsymbol{D}\boldsymbol{Q}\boldsymbol{D}$ can be easily verified. $\square$

Define $\boldsymbol{Y} = \boldsymbol{D}^{-1}\boldsymbol{X}$. We can see that

$$\boldsymbol{Y} = \begin{pmatrix} \dfrac{X_1}{\sqrt{np_1}} \\ \vdots \\ \dfrac{X_k}{\sqrt{np_k}} \end{pmatrix}, \qquad \mathsf{E}\boldsymbol{Y} = \begin{pmatrix} \sqrt{np_1} \\ \vdots \\ \sqrt{np_k} \end{pmatrix}$$

and an easy calculation gives

$$\mathsf{var}\,\boldsymbol{Y} = \boldsymbol{D}^{-1}\boldsymbol{V}\boldsymbol{D}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{D}\boldsymbol{D}^{-1} = \boldsymbol{Q}.$$

Define $\chi^2 = (\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})'(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})$. This expression can be written in the form

$$\chi^2 = \sum_{i=1}^{n}(Y_i - \mathsf{E}Y_i)^2 = \sum_{i=1}^{n}\left(\frac{X_i}{\sqrt{np_i}} - \sqrt{np_i}\right)^2 = \sum_{i=1}^{n}\frac{(X_i - np_i)^2}{np_i}.$$

Chapter 3

Transformations

3.1 Transformations of random variables

Let X be a random variable and t a measurable function. If $Y = t(X)$, then we say that the variable Y is the *transformation of the variable* X. Our task is to derive statistical characteristics of the variable Y from the known characteristics of the variable X. We met two important cases already in section 1.3 (see theorems 1.10 and 1.11). Now, we are going to investigate general cases.

Theorem 3.1 *Let X have a continuous distribution function F. Let $F'(x) = f(x)$ exist everywhere with exception maximally finite number of points. Let t be a monotonous function that has everywhere a nonvanishing derivative. Let τ be the inverse function to t. Then the random variable $Y = t(X)$ has the density*

$$g(y) = f[\tau(y)]|\tau'(y)|.$$

Proof. Assume first that t is increasing. Then the distribution function G of the random variable Y is equal to

$$G(y) = \mathsf{P}(Y \leq y) = \mathsf{P}[t(X) \leq y] = \mathsf{P}[X \leq \tau(y)] = F[\tau(y)].$$

The assumptions of theorem ensure that G is continuous and differentiable everywhere except maximally in finite many points. From this reason the density $g(y)$ of the variable Y is equal to

$$g(y) = G'(y) = f[\tau(y)]\tau'(y).$$

Since t is increasing, τ is also increasing and $\tau'(y) = |\tau'(y)|$. If t is decreasing, the proof is similar. □

The proof of theorem 3.1 was very simple. In many cases we can derive distribution of $Y = t(X)$, even if the mapping t is not monotonous.

Example 3.2 Let $X \sim \mathsf{N}(0, 1)$. As introduced in section 1.2, the density of X is denoted φ and its distribution function Φ. Define $Y = X^2$. Let G denote the distribution function of variable Y. It is clear that $G(y) = 0$ if $y \leq 0$. In the case $y > 0$ we get

$$G(y) = \mathsf{P}(Y < y) = \mathsf{P}(X^2 < y) = \mathsf{P}(-\sqrt{y} < X < \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}).$$

Since $\Phi(-x) = 1 - \Phi(x)$, we have together

$$G(y) = 2\Phi(\sqrt{y}) - 1.$$

It implies that the density g of the variable Y vanishes on the interval $(-\infty, 0]$. If $y > 0$ then

$$g(y) = G'(y) = \frac{1}{\sqrt{y}}\varphi(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}}\mathrm{e}^{-y/2}. \qquad \diamondsuit$$

The distribution of the random variable Y with the density g is called *chi-square distribution with one degree of freedom* and it is denoted χ^2_1. The expectation of this distribution is $\mu = \mathsf{E}Y = \mathsf{E}X^2 = 1$, the variance is $\sigma^2 = \mathsf{var}\ Y = \mathsf{E}Y^2 - (\mathsf{E}Y)^2 = \mathsf{E}X^4 - (\mathsf{E}X^2)^2 = 3 - 1^2 = 2$. $\diamondsuit$

3.2 Transformations of random vectors

Theorem 3.3 *Assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ has the density $f(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)'$. Let t be a regular one-to-one mapping from $\mathbb{R}_n$ onto $\mathbb{R}_n$. Denote τ the inverse mapping to t. Then the random vector $\boldsymbol{Y} = t(\boldsymbol{X})$ has the density*

$$g(\boldsymbol{y}) = f[\tau(\boldsymbol{y})]|D_\tau(\boldsymbol{y})|,$$

where

$$D_\tau(\boldsymbol{y}) = \begin{vmatrix} \dfrac{\partial\tau_1}{\partial y_1} & \cdots & \dfrac{\partial\tau_1}{\partial y_n} \\ \cdots\cdots & & \cdots\cdots \\ \dfrac{\partial\tau_n}{\partial y_1} & \cdots & \dfrac{\partial\tau_n}{\partial y_n} \end{vmatrix}$$

is the jacobi determinant of the mapping τ.

Proof. Let $B \in \mathcal{B}_n$ be an arbitrary borel set. Substitution theorem in multiple integrals gives

$$\begin{aligned} \mathsf{P}(\boldsymbol{Y} \in B) &= \mathsf{P}[t(\boldsymbol{X}) \in B] = \mathsf{P}[\boldsymbol{X} \in \tau(B)] \\ &= \int_{\tau(B)} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_B f[\tau(\boldsymbol{y})]|D_\tau(\boldsymbol{y})|\,\mathrm{d}\boldsymbol{y} = \int_B g(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}. \end{aligned} \tag{3.1}$$

It suffices to choose

$$B = (-\infty, x_1] \times \cdots \times (-\infty, x_n],$$

so that $\mathsf{P}(\boldsymbol{Y} \in B)$ is the value of the distribution function of the random vector $\boldsymbol{Y}$ in point $\boldsymbol{x}$. Formula (3.1) makes sure that g is the density corresponding to this distribution function. $\square$

Let $X_1, \ldots, X_n$ be independent random variables with distribution $\mathsf{N}(0, 1)$. Define $\boldsymbol{X} = (X_1, \ldots, X_n)'$. Then $\mathsf{E}\boldsymbol{X} = \boldsymbol{0}$, $\mathsf{var}\,\boldsymbol{X} = \boldsymbol{I}$. We say that $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$. The density of the vector $\boldsymbol{X}$ is

$$f(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}\right\}.$$

Let $\boldsymbol{\mu} \in \mathbb{R}_n$ and let $\boldsymbol{V}$ be a positive definite matrix. Denote $\boldsymbol{B} = \boldsymbol{V}^{1/2}$ and introduce the vector $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{BX}$. Obviously $\mathsf{E}\boldsymbol{Y} = \boldsymbol{\mu}$, $\mathsf{var}\,\boldsymbol{Y} = \boldsymbol{BIB}' = \boldsymbol{V}$. The inverse transformation is $\boldsymbol{X} = \boldsymbol{B}^{-1}\boldsymbol{Y} - \boldsymbol{B}^{-1}\boldsymbol{\mu}$ and its jacobian is $\det \boldsymbol{B}^{-1}$. The simultaneous density of the vector $\boldsymbol{Y}$ is

$$g(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}[\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})]'[\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})]\right\} |\det \boldsymbol{B}^{-1}|.$$

Since $|\det \boldsymbol{B}^{-1}| = |\boldsymbol{V}|^{-1/2}$, we get

$$g(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'(\boldsymbol{B}^{-1})'\boldsymbol{B}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}.$$

In view of $(\boldsymbol{B}^{-1})'\boldsymbol{B}^{-1} = \boldsymbol{V}^{-1}$, we have

$$g(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}. \tag{3.2}$$

We say that $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{V})$. The density of the vector $\boldsymbol{Y}$ is given by formula (3.2). Because the matrix $\boldsymbol{V}$ is regular, the distribution $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{V})$ is called *regular normal distribution*.

Let $\boldsymbol{a} \in \mathbb{R}_n$ and let $\boldsymbol{C}$ be a $n \times n$ regular matix. Define $\boldsymbol{Z} = \boldsymbol{a} + \boldsymbol{CY}$. It is clear that $\mathsf{E}\boldsymbol{Z} = \boldsymbol{a} + \boldsymbol{C\mu}$ and $\mathsf{var}\,\boldsymbol{Z} = \boldsymbol{CVC}'$. Theorem about transformation gives that the density of the vector $\boldsymbol{Z}$ is

$$h(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{CVC}'|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{a}-\boldsymbol{C\mu})'(\boldsymbol{CVC}')^{-1}(\boldsymbol{z}-\boldsymbol{a}-\boldsymbol{C\mu})\right\}.$$

We see that $\boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{a} + \boldsymbol{C\mu}, \boldsymbol{CVC}')$.

We inform without proof that this result is true also in the case when $\boldsymbol{a}$ is a vector with m components and $\boldsymbol{C}$ is a $m \times n$ matrix. If the rank of the matrix $\boldsymbol{CVC}'$ is smaller than m, the vector $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{CX}$ has the *singular normal distribution*.

We divide $\boldsymbol{X}$ into two blocks so that $\boldsymbol{X} = (\boldsymbol{Y}', \boldsymbol{Z}')'$, where $\boldsymbol{Y}$ contains first k components of the vector $\boldsymbol{X}$ and $\boldsymbol{Z}$ remaining $n-k$ components. The matrix $\boldsymbol{V}$ and vector $\boldsymbol{\mu}$ are divided analogously. We have

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\nu} \\ \boldsymbol{\tau} \end{pmatrix}, \qquad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix},$$

where $\boldsymbol{\nu}$ has k components, $\boldsymbol{\tau}$ has $n-k$ components, $\boldsymbol{V}_{11}$ is $k \times k$ matrix and $\boldsymbol{V}_{22}$ is a $(n-k) \times (n-k)$ matrix.

It can be proved that $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{\nu}, \boldsymbol{V}_{11})$ and that $\boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{\tau}, \boldsymbol{V}_{22})$. All the marginal distributions in multidimensional normal distribution are also normal. Marginal distribution of the variable X_i is $\mathsf{N}(\mu_i, \sigma_i^2)$, where $\sigma_i^2 = \sigma_{ii}$.

If $\boldsymbol{V}_{12} = \boldsymbol{0}$, then the vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are uncorrelated. Since $\boldsymbol{V}_{21} = \boldsymbol{V}_{12}' = \boldsymbol{0}$, the density of the vector $\boldsymbol{X}$ can be written in the form of the product of vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$. This proves that under assumption of normal distribution of the vector $\boldsymbol{X}$ uncorrelated vectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent.

Frequently we deal with the case $n = 2$, $k = 1$. Then $\boldsymbol{X} = (X_1, X_2)'$. Variance matrix $\boldsymbol{V}$ has the form

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where ρ is the correlation coefficient of the variables X_1 and X_2. If $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, and $\rho \in (-1, 1)$, then inserting into (3.2) we get the density of the *two-dimensional normal distribution*

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}.$$

Since the marginal density of the variable X_2 is

$$q(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2}\exp\left[-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right],$$

according to the section 2.4 we get the conditioned density of the variable X_1 given $X_2 = x_2$ as

$$r(x_1|x_2) = f(x_1, x_2)/q(x_2).$$

We can see that $r(x_1|x_2)$ is the density of the distribution

$$\mathsf{N}\left[\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2), \sigma_1^2(1-\rho^2)\right]. \qquad \diamondsuit$$

Theorems 3.1 and 3.3 have rather restrictive assumptions. We have made them to simplify the proofs and they can be removed.

Theorem 3.4 *Assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ has the density $f(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)'$. Let t be a mapping from $\mathbb{R}_n$ into $\mathbb{R}_n$, which is regular and one-to-one on such disjoint open sets $G_1, G_2, \ldots$, that for $G = \cup G_i$ we have*

$$\int_G f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1. \tag{3.3}$$

Let τ_j be the inverse mapping to $t : G_j \to t(G_j)$. Then the random vector $\boldsymbol{Y} = t(\boldsymbol{X})$ has the density

$$g(\boldsymbol{y}) = \sum_{j=1}^{\infty} g_j(\boldsymbol{y}),$$

where

$$g_j(\boldsymbol{y}) = \begin{cases} f[\tau_j(\boldsymbol{y})]|D_{\tau_j}(\boldsymbol{y})| & \text{for } \boldsymbol{y} \in t(G_j), \\ 0 & \text{for } \boldsymbol{y} \notin t(G_j). \end{cases}$$

3.3 Functions of random variables

3.3.1 Method of calculation

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ which has density $f(\boldsymbol{x})$. We must determine the density of a function $T_1 = t_1(X_1, \ldots, X_n)$. Usually, the calculation is made in two steps.

(i) We look for a mapping $\boldsymbol{T} = t(\boldsymbol{X})$ such that $\boldsymbol{T} = (T_1, \ldots, T_n)'$ and such that the assumptions of theorem 3.3 or theorem 3.4 are fulfilled. The first component of the vector $\boldsymbol{T}$ is the same as the function $t_1(X_1, \ldots, X_n)$, the distribution of which we are looking for. From this we determine the simultaneous density $g(t_1, \ldots, t_n)$ of the vector $\boldsymbol{T}$.

(ii) Marginal density g_1 of the variable T_1 can be calculated using formula

$$g_1(t_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(t_1, \ldots, t_n)\, \mathrm{d}t_2 \ldots \mathrm{d}t_n$$

(see section 2.1).

In practical calculations the choice of the mapping t in the first step is very important. If the procedure is not suitable, it may be difficult to calculate g or integral which defines g_1.

3.3.2 Sum of variables

Theorem 3.5 (convolution theorem) *Let X_1 and X_2 be independent random variables. Assume that X_1 has the density f_1 and X_2 has the density f_2. Then the random variable $Y = X_1 + X_2$ has the density*

$$h(y) = \int f_1(z) f_2(y - z)\, \mathrm{d}z. \tag{3.4}$$

Proof. Introduce the vector $\boldsymbol{X} = (X_1, X_2)'$. It follows from theorem 2.6 that the variables X_1, X_2 have the simultaneous density $f(x_1, x_2) = f_1(x_1) f_2(x_2)$. Define

$$\boldsymbol{T} = \begin{pmatrix} X_1 + X_2 \\ X_1 \end{pmatrix} = \begin{pmatrix} Y \\ Z \end{pmatrix}.$$

The inverse mapping is $X_1 = Z$, $X_2 = Y - Z$. Absolute value of jacobian is 1 and thus the simultaneous density $g(y, z)$ of the vector $\boldsymbol{T}$ is $g(y, z) = f_1(z) f_2(y - z)$. Marginal density of Y is $h(y) = \int g(y, z)\, \mathrm{d}z$. The formula 3.4) is proved. □

The formula

$$h(y) = \int f_1(y - z) f_2(z)\, \mathrm{d}z$$

can be proved analogously. This result also follows from the fact that $X_1 + X_2 = X_2 + X_1$. We say that the density h is the *convolution of densities* f_1 and f_2.

Let $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$ be independent random variables with normal distribution. We prove that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. If $\sigma_1^2 = 0$ or $\sigma_2^2 = 0$, the assertion is clear. If $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, then the assertion follows from theorem 3.5.

Let $X_1, \ldots, X_n$ be independent random variables with normal distribution $\mathsf{N}(0, 1)$. We define $Y = X_1^2 + \cdots + X_n^2$. The density of the variable Y will be denoted $f_n(y)$. It was already proved that

$$f_1(y) = \frac{1}{\sqrt{2\pi y}} \mathrm{e}^{-y/2} \qquad \text{for } y > 0.$$

We prove that

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n}{2}-1} \mathrm{e}^{-y/2} \qquad \text{for } y > 0. \tag{3.5}$$

The formula is valid if $n = 1$. For $n > 1$ it will be proved by complete induction. Assume that it is valid for all $n \leq k$. Convolution theorem gives

$$f_{k+1}(y) = \int f_k(z) f_1(y - z) \, \mathrm{d}z.$$

After some calculations we obtain formula 3.5), where $n = k + 1$.

Function $f_n(y)$ is the density of *chi-squared distribution* with n *degrees of freedom*, which is denoted as χ_n^2. We already know that $\mathsf{E}X_i^2 = 1$, $\mathsf{var}\, X_i^2 = 2$ for all i. Thus $\mu = \mathsf{E}Y = n$, $\sigma^2 = \mathsf{var}\, Y = 2n$. The expectation of χ_n^2 distribution is equal to the number of degrees of freedom, variance is twice so much.

Notice that $f_n(y)$ is density for all real $n > 0$, not only for integers.

Theorem 3.6 *Let X_1 and X_2 be independent random variables such that $X_1 \sim \chi_m^2$, $X_2 \sim \chi_n^2$. Then $X_1 + X_2 \sim \chi_{m+n}^2$.*

Proof. Using convolution theorem the density of $Y = X_1 + X_2$ is

$$h(y) = \int f_m(z) f_n(y - z) \, \mathrm{d}z.$$

After some calculations we get $h(y) = f_{m+n}(y)$. □

Analyzing χ^2 distribution we introduce a few theorems which are not connected directly with convolution.

Theorem 3.7 *Let the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ have the n-dimensional normal distribution $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{V})$ with the density (3.2). Then the random variable*

$$Y = (\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$$

has the χ_n^2 distribution.

Proof. Introduce the vector $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$. Then it can be proved that $\boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$ and the density of $\boldsymbol{Z}$ is

$$g(\boldsymbol{z}) = \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-\boldsymbol{z}'\boldsymbol{z}/2} = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} z_i^2 \right\} \right].$$

Theorem 2.6 gives that the components $Z_1, \ldots, Z_n$ are independent random variables and each of them has $\mathsf{N}(0,1)$ distribution. Since

$$\boldsymbol{Z}'\boldsymbol{Z} = Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2,$$

we also have

$$Y = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z}'\boldsymbol{Z} \sim \chi_n^2. \qquad \square$$

Theorem 3.8 *Let $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{V})$, where $\boldsymbol{V}$ is an idempotent matrix of rank $r \geq 1$. Then $\boldsymbol{X}'\boldsymbol{X} \sim \chi_r^2$.*

Proof. Every symmetric matrix $\boldsymbol{V}$ can be written in the form $\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'$, where $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{I}$ is the unit matrix and $\boldsymbol{D}$ is a diagonal matrix having eigenvalues of the matrix $\boldsymbol{V}$ on the diagonal.

First we show that the eigenvalues of an idempotent matrix are only zeroes and ones. Indeed, λ is an eigenvalue of the matrix $\boldsymbol{V}$ if and only if there exists a vector $\boldsymbol{x} \neq \boldsymbol{0}$ such that $\boldsymbol{V}\boldsymbol{x} = \lambda\boldsymbol{x}$. We multiply this equality by $\boldsymbol{V}$ from the left and we apply the fact that $\boldsymbol{V}$ is idempotent. This gives $\boldsymbol{V}^2\boldsymbol{x} = \lambda\boldsymbol{V}\boldsymbol{x}$, so that $\boldsymbol{V}\boldsymbol{x} = \lambda^2\boldsymbol{x}$. Thus $\lambda\boldsymbol{x} = \lambda^2\boldsymbol{x}$. Since $\boldsymbol{x} \neq \boldsymbol{0}$,we must have $\lambda = 0$ or $\lambda = 1$.

We can see that $h(\boldsymbol{V}) = r$ and that $\boldsymbol{U}$ is regular. Thus $\boldsymbol{D}$ has rank r. Since $\boldsymbol{D}$ is diagonal and there are only 0 and 1 on the diagonal, without loss of generality we can assume that $\boldsymbol{D} = \mathsf{Diag}\{1, \ldots, 1, 0, \ldots, 0\}$; number of 1 is r.

Define $\boldsymbol{Y} = \boldsymbol{U}'\boldsymbol{X}$. We know that $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{U}'\boldsymbol{V}\boldsymbol{U})$. But $\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U} = \boldsymbol{U}'\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{D}$. Matrix $\boldsymbol{D}$ is diagonal, and so all the components of the vector $\boldsymbol{Y}$ are uncorrelated. The vector $\boldsymbol{Y}$ has normal distribution and so its components are independent. Vector $\boldsymbol{Y}$ has vanishing expectation; r of its components have unit variance, other components have vanishing variances. Thus $n-r$ components of the vector $\boldsymbol{Y}$ are zeros almost surely. The form of $\boldsymbol{D}$ ensures that these zeros are the last variables $Y_{r+1}, \ldots, Y_n$. Finally, we have

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{X} = \boldsymbol{Y}'\boldsymbol{Y} = Y_1^2 + \cdots + Y_r^2 \sim \chi_r^2. \qquad \square$$

Theorem 3.9 *Let $\boldsymbol{X} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$ and let $\boldsymbol{A}$ be a symmetric idempotent matrix of rank $r \geq 1$. Then $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} \sim \chi_r^2$.*

Proof. We write $\boldsymbol{A}$ in the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'$, where $\boldsymbol{U}$ fulfills the condition $\boldsymbol{U}\boldsymbol{U}' = \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{D} = \mathsf{Diag}\{1, \ldots, 1, 0, \ldots, 0\}$ with r ones on the diagonal. We define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)' = \boldsymbol{U}'\boldsymbol{X}$, so that $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$. Then we have

$$\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{X} = \boldsymbol{Y}'\boldsymbol{D}\boldsymbol{Y} = Y_1^2 + \cdots + Y_r^2.$$

Since $Y_1, \ldots, Y_r$ are independent $\mathsf{N}(0,1)$ variables, the theorem is proved. $\square$

3.3.3 Quotient of random variables

Theorem 3.10 *Let X and Z be independent random variables such that $X \sim \mathsf{N}(0,1)$ and $Z \sim \chi^2_k$. Then the random variable*

$$T = \frac{X}{\sqrt{Z/k}}$$

has Student t distribution *with k degrees of freedom, which is denoted t_k and has the density*

$$h_k(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left(1+\frac{t^2}{k}\right)^{-(k+1)/2}.$$

Proof. Since X a Z are independent variables, their simultaneous density $g(x,z)$ is equal to product of their marginal densities

$$g(x,z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\frac{1}{2^{k/2}\Gamma(k/2)}z^{\frac{k}{2}-1}\mathrm{e}^{-z/2} \quad \text{for } z>0.$$

We use the transform

$$T = \frac{X}{\sqrt{Z/k}}, \qquad U = Z.$$

The inverse transformation is $X = T\sqrt{U/k}$, $Z = U$, and absolute value of jacobian is $\sqrt{u/k}$. Then the simultaneous density of the variables T and U is

$$p(t,u) = \frac{1}{\sqrt{2\pi k}\,2^{k/2}\Gamma(k/2)}u^{(k-1)/2}\exp\left\{-\frac{u}{2}\left(1+\frac{t^2}{k}\right)\right\}.$$

From here we obtain the marginal density T using the formula $h_k(t) = \int p(t,u)\,\mathrm{d}u$. □

In the case $k = 1$ we obtain $h_1(t) = 1/[\pi(1+t^2)]$. It is the density of *Cauchy distribution* $\mathsf{C}(0,1)$. If we use Stirling formula for Γ function, then we get $\lim_{k\to\infty} h_k(t) = \varphi(t)$. Densities of the Student distribution converge to the density of $\mathsf{N}(0,1)$. If $k = 1$, then $\mathsf{E}T$ does not exist, for $k > 1$ we have $\mathsf{E}T = 0$. Similarly, variance exists for $k > 2$; then $\mathsf{var}\,T = k/(k-2)$.

Let X and Z be independent random variables such that $X \sim \mathsf{N}(\delta,1)$ and $Z \sim \chi^2_k$. Then the random variable $T = \frac{X}{\sqrt{Z/k}}$ has *Student non-central* t *distribution* with k degrees of freedom with non-centrality parameter δ. This distribution is denoted as $\mathsf{t}_{k,\delta}$.

Theorem 3.11 *Let X and Y be independent random variables such that $X \sim \chi^2_m$, $Y \sim \chi^2_n$. Then the random variable*

$$Z = \frac{X/m}{Y/n}$$

has Fisher-Snedecor F distribution *with m and n degrees of freedom, which is denoted as $\mathsf{F}_{m,n}$ and the density of which is*

$$f_{m,n}(z) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{m/2} z^{\frac{m}{2}-1}\left(1+\frac{m}{n}z\right)^{-(m+n)/2} \quad \textit{for } z>0.$$

Proof. The simultaneous density of the variables X and Y is

$$p(x,y) = \frac{1}{2^{(m+n)/2}\Gamma(m/2)\Gamma(n/2)} x^{\frac{m}{2}-1} y^{\frac{n}{2}-1} \mathrm{e}^{-(x+y)/2}$$

for $x > 0$, $y > 0$. We make the transform

$$Z = \frac{X/m}{Y/n}, \qquad U = Y.$$

The inverse transform is $X = mZU/n$, $Y = U$ and absolute value of its jacobian is mu/n. Inserting for x a y into $p(x,y)$ and multiplying mu/n we get simultaneous density $g(z,y)$ of the variables Z and Y. Integrating with respect to y we get marginal density of variable Z, which is introduced in theorem. □

Moments of $\mathsf{F}_{m,n}$ distribution are

$$\mathsf{E}Z = \frac{n}{n-2} \qquad \text{for } n > 2,$$

$$\mathsf{var}\, Z = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \text{for } n > 4.$$

For $n \leq 2$ the expectation is not finite and for $n \leq 4$ the second moment is not finite.

3.4 Transformation stabilizing variance

Assume that a random variable X has a distribution which depends on a parameter θ. The parameter is chosen in such a way that $\mathsf{E}X = \theta$. In many cases the variance of variable X also depends on θ and we can write $\mathsf{var}\, X = \sigma^2(\theta)$. Usually, $\sigma(\theta)$ is a smooth function of θ. We can ask the question if it is possible to find such a non-trivial function g that the random variable $Y = g(X)$ has the variance independent of θ. We excluded constant functions g which would lead to variables with vanishing variance. Generally, this problem has no solution. However, some approximations are useful. Taylor formula gives

$$g(X) \doteq g(\theta) + (X-\theta)g'(\theta),$$

so that

$$\mathsf{E}g(X) \doteq g(\theta), \qquad \mathsf{var}\, g(X) \doteq [g'(\theta)]^2\sigma^2(\theta). \tag{3.6}$$

The expression $[g'(\theta)]^2\sigma^2(\theta)$ will not depend on θ, if we ensure

$$g'(\theta)\sigma(\theta) = c,$$

where c is a constant. From this condition we have the solution

$$g(\theta) = c\int \frac{\mathrm{d}\theta}{\sigma(\theta)}. \tag{3.7}$$

The constant c is chosen so that the function g calculated from (3.7) has a nice form. We get $\mathsf{var}\, g(X) \doteq c^2$.

It was shown that the function g calculated from (3.7) stabilizes variance and $\mathsf{var}\, g(X)$ depends on θ only a little, but at the same time the distribution of the random variable $Y = g(X)$ is nearly normal.

We apply this method to some special cases.

Example 3.12 Let the random variable X have *Poisson distribution* with parameter $\lambda > 0$. It means that

$$\mathsf{P}(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \qquad k = 0, 1, 2, \ldots$$

We write shortly $X \sim \mathsf{Po}(\lambda)$. Since it holds

$$\mathsf{E}X = \lambda, \qquad \mathsf{var}\, X = \lambda,$$

instead of parameter θ we have the parameter λ and $\sigma^2(\lambda) = \lambda$. From (3.7) we get

$$g(\lambda) = c \int \frac{\mathrm{d}\lambda}{\sqrt{\lambda}} = 2c\sqrt{\lambda}.$$

Usually, we choose $c = \frac{1}{2}$, and we deal with the function $g(\lambda) = \sqrt{\lambda}$. It is the well-known *square-root transformation*. In view of (3.6) we have

$$\mathsf{E}\sqrt{X} \doteq \sqrt{\lambda}, \qquad \mathsf{var}\,\sqrt{X} \doteq \frac{1}{4}. \qquad \diamondsuit$$

Example 3.13 Let ξ have the *binomial distribution* $\mathsf{Bi}(n, p)$, so that ξ can be only $0, 1, \ldots, n$ and the corresponding probabilities are

$$\mathsf{P}(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

It is assumed that $p \in (0, 1)$ and that n is a natural number. The variable ξ can represent number of successes in n independent experiments when in every experiment probability of success is p. A simple calculation gives

$$\mathsf{E}\xi = np, \qquad \mathsf{var}\,\xi = np(1-p).$$

The parameter p is estimated using the variable $X = \xi/n$. Thus we have

$$\mathsf{E}X = p, \qquad \mathsf{var}\, X = \frac{p(1-p)}{n}.$$

Instead of θ we have in our case p a $\sigma^2(p) = p(1-p)/n$. According (3.7) we get

$$g(p) = c\sqrt{n} \int \frac{\mathrm{d}p}{\sqrt{p(1-p)}} = 2c\sqrt{n}\arcsin\sqrt{p}.$$

Usually, one chooses $c = 1/(2\sqrt{n})$, so that finally

$$g(p) = \arcsin\sqrt{p}.$$

From (3.6) we calculate

$$\mathsf{E}\arcsin\sqrt{X} \doteq \arcsin\sqrt{p}, \qquad \mathsf{var}\,\arcsin\sqrt{X} \doteq \frac{1}{4n}. \qquad \diamondsuit$$

Example 3.14 *Sample correlation coefficient* r calculated from a sample of size n from *two-dimensional normal distribution* having theoretical correlation coefficient $\rho \in (-1, 1)$, fulfills

$$\mathsf{E}r \doteq \rho, \qquad \mathsf{var}\, r \doteq \frac{(1-\rho^2)^2}{n}.$$

Since $\sigma(\rho) \doteq n^{-1/2}(1-\rho^2)$, we have

$$g(\rho) \doteq c\sqrt{n} \int \frac{\mathrm{d}\rho}{1-\rho^2} = \frac{1}{2} c\sqrt{n}\, \ln \frac{1+\rho}{1-\rho}.$$

If $c = 1/\sqrt{n}$ then

$$g(\rho) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho},$$

or $g(\rho) = \mathrm{arctgh}\, \rho$. If we define

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r},$$

then we get *Fisher z-transformation*. Using (3.6) we have

$$\mathsf{E}Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \qquad \mathsf{var}\, Z \doteq \frac{1}{n}.$$

More detailed calculation gives

$$\mathsf{E}Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \qquad \mathsf{var}\, Z \doteq \frac{1}{n-3}.$$

In practical computations the approximations

$$\mathsf{E}Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \qquad \mathsf{var}\, Z \doteq \frac{1}{n-3}.$$

are used. See Winterbottom (1979). $\diamondsuit$

Example 3.15 Let $X \sim \chi^2_n$. If we write θ instead of n then $\mathsf{E}X = \theta$, $\mathsf{var}\, X = 2\theta$ and

$$g(\theta) = c \int \frac{\mathrm{d}\theta}{\sqrt{2\theta}} = c\sqrt{2\theta}.$$

Usually, $c = 1$ is chosen. Then $g(\theta) = \sqrt{2\theta}$, $Y = \sqrt{2X}$ and $\mathsf{E}Y \doteq \sqrt{2n}$, $\mathsf{var}\, Y \doteq 1$.

R. A. Fisher recommended to use transformation $Y = \sqrt{2X} - \sqrt{2n-1}$, because its distribution tends to the normal distribution $\mathsf{N}(0,1)$ rather quickly. Today, Wilson-Hilferty transformation is used, i.e.

$$U = 3\sqrt{\frac{n}{2}} \left(\sqrt[3]{\frac{X}{n}} + \frac{2}{9n} - 1 \right),$$

and its distribution converges to $\mathsf{N}(0,1)$ even quicker. (See Kendall, Stuart 1969 I, Rao 1978, Wilson, Hilferty 1931.) $\diamondsuit$

Normalization transformations for other types of distributions are given in Konishi (1981).

Chapter 4

Random sample

4.1 Simple random sample

A series of independent identically distributed random variables $X_1, \dots, X_n$ is called *a simple random sample*. Number n is *size of the sample*. Introduce variables

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The variable $\bar{X}$ is *sample mean*. Variable S^2 is defined only for $n \geq 2$.

Theorem 4.1 *Let $X_1, \dots, X_n$ be the random sample from the distribution with expectation μ and a finite variance σ^2. Then it holds*

$$\mathsf{E}\bar{X} = \mu, \qquad \mathsf{var}\,\bar{X} = \frac{\sigma^2}{n}, \qquad \mathsf{E}S^2 = \sigma^2.$$

Proof. First, we calculate

$$\mathsf{E}\bar{X} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}X_i = \frac{1}{n}n\mu = \mu.$$

Using theorems 2.4 and 2.8 we get

$$\mathsf{var}(X_1 + \cdots + X_n) = \mathsf{var}\,X_1 + \cdots + \mathsf{var}\,X_n = n\sigma^2.$$

Thus $\mathsf{var}\,\bar{X} = n^{-2}n\sigma^2 = \sigma^2/n$. Further

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2,$$

so that

$$\mathsf{E}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\mathsf{E}(X_i - \mu)^2 - n\mathsf{E}(\bar{X} - \mu)^2 = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

This gives the last assertion. □

Because $\mathsf{E}\bar{X} = \mu$ we say that $\bar{X}$ is unbiased estimator of the parameter μ. Similarly, S^2 is an unbiased estimator of the parameter σ^2, since $\mathsf{E}S^2 = \sigma^2$. In the general case we have the following definition. Let $(X_1, \ldots, X_n)'$ be a random vector, the distribution of which depends on a parameter θ. Let $g(x_1, \ldots, x_n)$ be a borel measurable function which does not depend on θ. We say that $T = g(X_1, \ldots, X_n)$ is an *estimator* of the parameter θ. If $\mathsf{E}T = \theta$ holds for every θ then the estimator T is unbiased.

In the multidimensional case the simple random sample is a series $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of independent identically distributed random vectors. This sample is characterized by means of

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i, \qquad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.$$

Theorem 4.2 *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample from distribution with expectation $\boldsymbol{\mu}$ and a variance matrix $\boldsymbol{V}$. Then we have*

$$\mathsf{E}\bar{\boldsymbol{X}} = \boldsymbol{\mu}, \qquad \mathsf{var}\,\bar{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{V}, \qquad \mathsf{E}\boldsymbol{S} = \boldsymbol{V}.$$

Proof is similar as in Theorem 4.1. □

4.2 Ordered random sample

Let $X_1, \ldots, X_n$ be a random sample from a distribution having the distribution function F. Random variables $X_1, \ldots, X_n$ will be ordered and the smallest will be denoted $X_{(1)}$, the second smallest $X_{(2)}$, and finally the largest $X_{(n)}$. We have

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

The variables $X_{(1)}, \ldots, X_{(n)}$ are called *ordered random sample*.

Theorem 4.3 *Let $r \in \{1, 2, \ldots, n\}$. Then the distribution function G_r of the random variable $X_{(r)}$ is*

$$G_r(x) = \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i}. \tag{4.1}$$

Proof. Probability that there will be i variables among $X_1, \ldots, X_n$ such that their values are smaller or equal x, is

$$\binom{n}{i} F^i(x)[1 - F(x)]^{n-i},$$

because the number of such variables has the binomial distribution. The variable $X_{(r)}$ will be smaller than x if among $X_1, \ldots, X_n$ either r, or $r + 1$, etc. or n variables smaller than x will be found. These cases are disjoint and so the resulting probability is equal to sum of their probabilities. □

Formulas for distribution functions of the smallest and the largest member of ordered random sample follow from formula (4.1) as special cases. The results are

$$G_1(x) = 1 - [1 - F(x)]^n, \qquad G_n(x) = F^n(x).$$

Both the formulas can be also derived directly without using theorem 4.3.

If n is large, formula (4.1) contains many members. Calculations containing $G_r(x)$ may be rather complex. If the distribution function F has a density f, then also the density g_r exists and its formula is quite simple.

Theorem 4.4 *Let F be distribution function of a continuous distribution with the density f. Then the density g_r of the variable $X_{(r)}$ also exists and for $n \geq 2$ we have*

$$g_r(x) = n\binom{n-1}{r-1} f(x) F^{r-1}(x)[1 - F(x)]^{n-r}. \tag{4.2}$$

Proof. We start with formula $g_r(x) = G_r'(x)$ and insert (4.1). Then all members disappear except for one which is equal to (4.2). □

Especially,

$$g_1(x) = nf(x)[1 - F(x)]^{n-1}, \qquad g_n(x) = nf(x)F^{n-1}(x).$$

It is possible to derive simultaneous distribution function and simultaneous density of several members of ordered random sample. Here we restrict to two members.

Theorem 4.5 *Let F be the distribution function of the continuous distribution with the density f. If $n \geq 2$ and if $1 \leq r < s \leq n$, then the simultaneous density $g_{r,s}$ of the variables $X_{(r)}$, $X_{(s)}$ exists and equals to*

$$\begin{aligned} g_{r,s}(x,y) = &\frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} \\ &\times f(x)f(y)F^{r-1}(x)[F(y) - F(x)]^{s-r-1}[1 - F(y)]^{n-s}, \end{aligned}$$

if $x < y$. In the case $x \geq y$ we have $g_{r,s}(x,y) = 0$.

Proof. The proof is similar to that of theorem 4.4. □

Theorem is often used for $r = 1$, $s = n$. Then the simultaneous density of variables $X = X_{(1)}$ and $Y = X_{(n)}$ is

$$g_{1n}(x,y) = n(n-1)f(x)f(y)[F(y) - F(x)]^{n-2}.$$

An important characteristic is the *range* $R = Y - X$. We derive a formula for the density h and for the distribution function H of the variable R. We start with the density $g_{1n}(x,y)$. Consider the transform $R = Y - X$, $T = X$. The inverse transform is $X = T$, $Y = R + T$ and the corresponding jacobian is

$$\begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial t} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} = -1.$$

Thus the simultaneous density of the variables R and T is

$$s(r,t) = n(n-1)f(t)f(r+t)[F(r+t) - F(t)]^{n-2},$$

marginal density of R is

$$h(r) = \int_{-\infty}^{\infty} n(n-1)f(t)f(r+t)[F(r+t) - F(t)]^{n-2}\,\mathrm{d}t, \qquad r > 0,$$

and the distribution function of the variable R is

$$H(y) = \int_0^y h(r)\,\mathrm{d}r = n\int_{-\infty}^{\infty} f(t)\left\{\int_0^y (n-1)f(r+t)[F(r+t) - F(t)]^{n-2}\,\mathrm{d}r\right\}\,\mathrm{d}t$$

for $y > 0$. Because

$$\frac{\partial}{\partial r}[F(r+t) - F(t)]^{n-1} = (n-1)f(r+t)[F(r+t) - F(t)]^{n-2},$$

we obtain

$$H(y) = n\int_{-\infty}^{\infty} f(t)\left\{[F(r+t) - F(t)]^{n-1}\right\}_0^y\,\mathrm{d}t = n\int_{-\infty}^{\infty} f(t)[F(y+t) - F(t)]^{n-1}\,\mathrm{d}t.$$

It can be derived that the simultaneous distribution function $G_{1,n}$ of the variables $X_{(1)}$ and $X_{(n)}$ is

$$G_{1,n}(x,y) = F^n(y) - [F(y) - F(x)]^n \qquad \text{for } x < y.$$

In connection with random sample we introduce *rank* of the random variable. If the random variable X_i is in the random sample jth (i.e., if $X_i = X_{(j)}$), then the rank R_i of this variable equals to j. Value R_i equals the number of variables in the random sample which are smaller or equal X_i. This procedure is applied when the distribution is continuous where all values $X_1(\omega), \ldots, X_n(\omega)$ are different with probability one (no ties). If several values are identical then every variable from this group usually obtains arithmetical mean from the corresponding ranks.

4.3 Random sample from the normal distribution

Consider problems investigated in section 4.1. If we specify distribution which generates the sample, we can specify some assertions and some other derive.

Theorem 4.6 *Let $X_1, \ldots, X_n$ be the sample from the distribution $\mathsf{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$. Then the following asertions hold:*

(a) $\bar{X} \sim \mathsf{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.

(b) If $n \geq 2$, then $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

(c) If $n \geq 2$, then $\bar{X}$ and S^2 are independent.

Proof. (a) We know that $X_1 + \cdots + X_n \sim \mathsf{N}(n\mu, n\sigma^2)$. Then $\bar{X} = n^{-1}(X_1 + \cdots + X_n) \sim \mathsf{N}(\mu, \sigma^2/n)$.

(b) Since

$$S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2 = \frac{1}{n-1}\sum [(X_i - \mu) - (\bar{X} - \mu)]^2,$$

we can consider variables $X_i - \mu$ instead of original X_i. The equality ensures that S^2 is not changed when X_i is substituted by $X_i - \mu$. Vector $\boldsymbol{b}_1 = (n^{-1/2}, \ldots, n^{-1/2})'$ has the unit length. Thus there exist vectors $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ such that the matrix $\boldsymbol{B}' = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$ is orthonormal. Make the transform $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X}$. Obviously, the first component of the vector $\boldsymbol{Y}$ is $Y_1 = \sqrt{n}\bar{X}$. Since $\boldsymbol{B}$ is orthonormal, we see that $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}$. We have $\boldsymbol{X} = \boldsymbol{B}'\boldsymbol{Y}$ and

$$\begin{aligned}(n-1)S^2 &= \sum X_i^2 - n\bar{X}^2 = \boldsymbol{X}'\boldsymbol{X} - n\bar{X}^2 = \boldsymbol{Y}'\boldsymbol{B}\boldsymbol{B}'\boldsymbol{Y} - n\bar{X}^2 \\ &= \boldsymbol{Y}'\boldsymbol{Y} - n\bar{X}^2 = Y_1^2 + Y_2^2 + \cdots + Y_n^2 - Y_1^2 = Y_2^2 + \cdots + Y_n^2.\end{aligned}$$

The simultaneous density of variables $X_1, \ldots, X_n$ is equal to the product of their densities. We compare this product with the formula (3.2). Under our simplifying assumption $\mathsf{E}X_i = 0$ we can see that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ has distribution $\mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then $\boldsymbol{Y}$ has distribution $\mathsf{N}(\boldsymbol{0}, \boldsymbol{B}\sigma^2\boldsymbol{I}\boldsymbol{B}')$, which is distribution $\mathsf{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Variables $Y_1, \ldots, Y_n$ are independent and each of them has distribution $\mathsf{N}(0, \sigma^2)$. Thus $Y_i/\sigma \sim \mathsf{N}(0,1)$ and we have

$$(n-1)S^2/\sigma^2 = (Y_2/\sigma)^2 + \cdots + (Y_n/\sigma)^2 \sim \chi^2_{n-1}.$$

(c) It follows from (b) that the variables $Y_1, \ldots, Y_n$ are independent. Thus the variables $Y_1 = \sqrt{n}\bar{X}$ and $Y_2^2 + \cdots + Y_n^2 = (n-1)S^2$ are independent. If we add to all variables X_i the original expectation μ, then $\bar{X}$ changes μ and S^2 remains the same. Their independence remains. □

Chapter 5

Estimating theory

5.1 Statistics and unbiased estimators

Let the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ have the density $f(\boldsymbol{x}, \boldsymbol{\theta})$ with respect to a σ-finite measure μ, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is an unknown parameter. Using vector $\boldsymbol{X}$ the best estimator of the parameter $\boldsymbol{\theta}$ should be found. It is only known that $\boldsymbol{\theta}$ is an element of a parametric space $\Omega \subset \mathbb{R}_m$. If the problem is to find *point estimator*, it means to find a measurable mapping $g : (\mathbb{R}_n, \mathcal{B}_n) \to (\mathbb{R}_m, \mathcal{B}_m)$ such that the random vector $\boldsymbol{T} = g(\boldsymbol{X})$ is the best approximation of the value $\boldsymbol{\theta}$. In the case of *interval estimator* we wish to find interval or another appropriate set which covers $\boldsymbol{\theta}$ with sufficiently large probability. In this chapter we deal with point estimators.

Usually, we create a new random vector $\boldsymbol{S} = [S_1(\boldsymbol{X}), \ldots, S_k(\boldsymbol{X})]'$. This new vector $\boldsymbol{S}$ is called *statistics*. We emphasize that statistics is not a function of the parameter $\boldsymbol{\theta}$. We choose functions $S_i(\boldsymbol{x})$ such that the vector $\boldsymbol{S}$ has lower dimension than $\boldsymbol{X}$ and no information about $\boldsymbol{\theta}$ is lost. If we intend to take statistics $\boldsymbol{S}$ as estimator of $\boldsymbol{\theta}$, we must have $k = m$.

We say that the estimator $\boldsymbol{T}$ of the parameter $\boldsymbol{\theta}$ is *unbiased* if $\mathsf{E}\boldsymbol{T} = \boldsymbol{\theta}$ holds for every $\boldsymbol{\theta} \in \Omega$. If we have $\mathsf{E}\boldsymbol{T} = \boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})$, where the function $\boldsymbol{b}$ is not identically vanishing on the set Ω the estimator $\boldsymbol{T}$ is called *biased*. Vector $\boldsymbol{b}(\boldsymbol{\theta})$ is called *bias* of $\boldsymbol{T}$ at $\boldsymbol{\theta}$.

Instead of $\mathsf{E}\boldsymbol{T}$ we should write $\mathsf{E}_{\boldsymbol{\theta}}\boldsymbol{T}$ to indicate that the expectation is calculated when the parameter has value $\boldsymbol{\theta}$. However, we shall mostly use $\mathsf{E}\boldsymbol{T}$.

A requirement on estimator is its unbiasedness. However, it can happen that there exists another estimator which is biased but better from another point of view.

5.2 Examples

Example 5.1 Let $X_1, \ldots, X_n$ be a sample from $\mathsf{N}(\mu, \sigma^2)$, where $n \geq 2$ and $\sigma > 0$. Assume that none of parameters μ a σ^2 is known and we should estimate σ^2. Denote

$$Y = \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The variable $S^2 = \frac{1}{n-1} Y$ has been used as estimator of σ^2 and we justified it by the property that S^2 is unbiased estimator for σ^2 (theorem 4.1). Another estimator for σ^2

could be the *sample variance* M_2 defined as

$$M_2 = \frac{1}{n}Y.$$

We show that

$$\mathsf{E}(M_2 - \sigma^2)^2 < \mathsf{E}(S^2 - \sigma^2)^2,$$

i.e., that M_2 has smaller mean square error than S^2. Theorem 4.6 b) on p. 48 gives that $Y/\sigma^2 \sim \chi^2_{n-1}$. It implies

$$\mathsf{E}Y = \sigma^2(n-1), \quad \mathsf{var}\, Y = 2\sigma^4(n-1), \quad \mathsf{E}Y^2 = \mathsf{var}\, Y + (\mathsf{E}Y)^2 = \sigma^4(n^2-1).$$

Inserting for M_2 and S^2 we get

$$\begin{aligned}\mathsf{E}(M_2 - \sigma^2)^2 &= \mathsf{E}M_2^2 - 2\sigma^2\mathsf{E}M_2 + \sigma^4 = \frac{1}{n^2}\mathsf{E}Y^2 - \frac{2\sigma^2}{n}\mathsf{E}Y + \sigma^4 \\ &= \frac{2n-1}{n^2}\sigma^4, \\ \mathsf{E}(S^2 - \sigma^2)^2 &= \mathsf{var}\, S^2 = \frac{1}{(n-1)^2}\mathsf{var}\, Y = \frac{2}{n-1}\sigma^4.\end{aligned}$$

Since for every positive integer n we have

$$\frac{2n-1}{n^2} < \frac{2}{n-1},$$

the mean square error of the variable M_2 is smaller than that of S^2.

It is interesting that the coefficient $\frac{1}{n}$ by Y is not optimal in sense of mean square error. Look for a number k such that the expression $\mathsf{E}(kY - \sigma^2)^2$ is minimal. Since

$$\begin{aligned}\mathsf{E}(kY - \sigma^2)^2 &= k^2\mathsf{E}Y^2 - 2k\sigma^2\mathsf{E}Y + \sigma^4 = \sigma^4[k^2(n^2-1) - 2k(n-1) + 1] \\ &= \sigma^4\left[(n^2-1)\left(k - \frac{1}{n+1}\right)^2 + \frac{2}{n+1}\right],\end{aligned}$$

it is clear that $k = \frac{1}{n+1}$ gives the minimum and the minimum is $\frac{2}{n+1}\sigma^4$.

This example is special case of more general situation (see Williams 2001, p. 191). Let T be such an unbiased estimator of the parameter θ that $0 \neq \mathsf{E}T^2 < \infty$ for every $\theta \in \Omega$. Then $\mathsf{E}(aT - \theta)^2$ is minimal in the case

$$a = \frac{\theta^2}{\mathsf{E}T^2}.$$

In our case we had $\theta = \sigma^2$, $T = S^2$. It happens that a does not depend on θ. This was also in our special case. $\diamondsuit$

Example 5.2 Let random variable X have *geometric distribution* $\mathsf{Ge}(p)$, where $p \in (0,1)$ is an unknown parameter. Define $q = 1 - p$. Random variable X is number of failures in sample from *Bernoulli distribution* before the first success. We have $\mathsf{P}(X = k) = q^k p$,

$k = 0, 1, \ldots$. Let us look for an unbiased estimator for p. If $T(X)$ is such estimator, it must satisfy

$$\mathsf{E}T(X) = \sum_{k=0}^{\infty} T(k)q^k p = p, \qquad p \in (0,1).$$

From here we get

$$\sum_{k=0}^{\infty} T(k)q^k = 1, \qquad q \in (0,1),$$

and the estimator satisfies $T(0) = 1$ and $T(k) = 0$ for $k \geq 1$. This is a poor estimator because the number of failures before the first success is taken only as information if the success was already in the first experiment or not. $\diamondsuit$

Example 5.3 Let a random variable X have *Poisson distribution* $\mathsf{Po}(\lambda)$. Let $\lambda > 0$ be an unknown parameter. We look for an unbiased estimator for $\mathrm{e}^{-2\lambda}$. If $T(X)$ is such an estimator, we must have $\mathsf{E}T(X) = \mathrm{e}^{-2\lambda}$. It means that

$$\sum_{x=0}^{\infty} T(x)\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \mathrm{e}^{-2\lambda},$$

and so

$$\sum_{x=0}^{\infty} T(x)\frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}.$$

Finally, we obtain

$$\sum_{x=0}^{\infty} T(x)\frac{\lambda^x}{x!} = \sum_{x=0}^{\infty}(-1)^x\frac{\lambda^x}{x!}.$$

This equality must hold for all $\lambda > 0$ and so coefficients by λ^x must be the same. It gives $T(x) = (-1)^x$. This estimator is not good. It can be negative although it is estimator of a non-negative function. Further, small change of variable X causes large change of estimator.

There are also other modifications of this example. Williams (2001) on p. 188 introduces that Lehmann is the author of the following result. Let

$$\mathsf{P}(X = k) = \frac{\lambda^k}{k!}\frac{1}{\mathrm{e}^{\lambda} - 1}, \qquad k = 1, 2, 3, \ldots$$

(it is *truncated Poisson distribution*). We look for an unbiased estimator T of the parametric function $1 - \mathrm{e}^{-\lambda}$ based on X. Since T must be an unbiased estimator, we must have

$$\frac{1}{\mathrm{e}^{\lambda} - 1}\sum_{k=1}^{\infty} T(k)\frac{\lambda^k}{k!} = 1 - \mathrm{e}^{-\lambda},$$

i.e.

$$\sum_{k=1}^{\infty} T(k)\frac{\lambda^k}{k!} = 2\sum_{n=1}^{\infty}\frac{\lambda^{2n}}{(2n)!}.$$

It implies that

$$T = \begin{cases} 0, & \text{when } X \text{ is an odd number,} \\ 2, & \text{when } X \text{ is an even number.} \end{cases}$$

Since $1 - \mathrm{e}^{-\lambda} \in (0,1)$, the estimator T is not acceptable. $\diamondsuit$

Example 5.4 Let $X \sim \mathsf{Bi}(n,p)$, i.e.,

$$\mathsf{P}(X=x)=\binom{n}{x}p^x(1-p)^{n-x}, \qquad n\geq 1, \quad 0<p<1, \quad x=0,1,\ldots,n.$$

We show that there exists no unbiased estimator for the parametric function $\frac{1}{p}$. Assume that there exists a function T, that $\mathsf{E}T(X)=\frac{1}{p}$ for every $p\in(0,1)$. Then

$$\sum_{x=0}^{n}T(x)\binom{n}{x}p^x(1-p)^{n-x}=\frac{1}{p}, \qquad 0<p<1.$$

The left hand side is a polynomial in p having degree maximally n, and it cannot be the same as the function $\frac{1}{p}$ on the interval $(0,1)$. $\diamondsuit$

Remark 5.5 Let $X_1, X_2, \ldots$ be independent random variables with *Bernoulli distribution* with probability of success $p\in(0,1)$. Then $X=X_1+\cdots+X_n\sim\mathsf{Bi}(n,p)$. Example 5.4 does not imply, that for the parametric function $\frac{1}{p}$ is not possible to find an unbiased estimator.

Let ξ be the number of failures in the sample from the Bernoulli distribution before the first success. Then ξ has geometric distribution $\mathsf{Ge}(p)$, i.e., $\mathsf{P}(\xi=k)=p(1-p)^k$, $k=0,1,2,\ldots$. It is well known that $\mathsf{E}\xi=\frac{1-p}{p}=\frac{1}{p}-1$. Thus $\xi+1$ is an unbiased estimator for $\frac{1}{p}$. $\diamondsuit$

Let us remark that in examples 5.2 — 5.4 the sample sizes were only 1. In the remark 5.5 the sample had not a fixed size.

5.3 Consistent estimators

5.3.1 Definition

Let θ be an univariate parameter. Assume that $X_1, X_2, \ldots$ is a sample from a distribution which depends on θ. Let us have an estimator $T_n=g_n(X_1,\ldots,X_n)$ defined for all nonnegative integers. We say that the estimator T_n is *consistent*, if $T_n\to\theta$ in probability for $n\to\infty$.

Theorem 5.6 *Let $\mathsf{E}T_n^2<\infty$ for every nonnegative integer n. If $\mathsf{E}T_n\to\theta$ and $\mathsf{var}\,T_n\to 0$, then T_n is a consistent estimator of the parameter θ.*

Proof. For every $\varepsilon>0$ we have

$$\begin{aligned}
\mathsf{P}(|T_n-\theta|>\varepsilon) &= \int_{|T_n-\theta|>\varepsilon}\mathrm{d}\mathsf{P}\leq\int_{|T_n-\theta|>\varepsilon}\varepsilon^{-2}(T_n-\theta)^2\,\mathrm{d}\mathsf{P}\\
&\leq\varepsilon^{-2}\int[(T_n-\mathsf{E}T_n)+(\mathsf{E}T_n-\theta)]^2\,\mathrm{d}\mathsf{P}\\
&=\varepsilon^{-2}\left[\int(T_n-\mathsf{E}T_n)^2\,\mathrm{d}\mathsf{P}\right.\\
&\qquad\left.+2(\mathsf{E}T_n-\theta)\int(T_n-\mathsf{E}T_n)\,\mathrm{d}\mathsf{P}+(\mathsf{E}T_n-\theta)^2\right]\\
&=\varepsilon^{-2}[\mathsf{var}\,T_n+(\mathsf{E}T_n-\theta)^2]\to 0. \qquad \square
\end{aligned}$$

5.3.2 Example

Example 5.7 Let $X_1, \ldots, X_n$ be a sample from the *rectangular distribution* $\mathsf{R}(0, \theta)$, where $\theta > 0$ is an unknown parameter. Let $X_{(n)} = \max(X_1, \ldots, X_n)$. We show that $X_{(n)}$ is a biased consistent estimator of the parameter θ. Since

$$\mathsf{P}(X_{(n)} < x) = \frac{x^n}{\theta^n}, \qquad 0 < x < \theta,$$

the density of the variable $X_{(n)}$ is

$$p(x) = \frac{nx^{n-1}}{\theta^n}, \qquad 0 < x < \theta.$$

This implies

$$\mathsf{E}X_{(n)} = \frac{n}{\theta^n} \int_0^\theta x^n \, \mathrm{d}x = \frac{n\theta}{n+1}, \qquad \mathsf{E}X_{(n)}^2 = \frac{n}{\theta^n} \int_0^\theta x^{n+1} \, \mathrm{d}x = \frac{n\theta^2}{n+2},$$

and thus

$$\mathsf{var}\, X_{(n)} = \mathsf{E}X_{(n)}^2 - [\mathsf{E}X_{(n)}]^2 = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Using theorem 5.6 we have that $X_{(n)}$ is a consistent estimator for θ. However, in this case it is not difficult to modify $X_{(n)}$ to obtain from $X_{(n)}$ an unbiased consistent estimator. If we define $T = \frac{n+1}{n} X_{(n)}$, then

$$\mathsf{E}T = \theta, \qquad \mathsf{var}\, T = \frac{\theta^2}{n(n+2)}. \quad \diamondsuit$$

Chapter 6

Empirical estimators

6.1 Empirical distribution function

We have the numbers $x_1, \ldots, x_n$. We order them in such a way that $x_{(1)} \leq \cdots \leq x_{(n)}$. *Empirical distribution function* (ecdf — empirical cumulative distribution function) is defined by the formula

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi_{(-\infty,x]}(x_i),$$

where

$$\chi_A(x) = \begin{cases} 1 & \text{for } x \in A, \\ 0 & \text{for } x \notin A \end{cases}$$

is *characteristic function of the set* A. It means that $F_n(x) = 0$ for $x < x_{(1)}$. If all the values $x_1, \ldots, x_n$ are different, then at each of them the function $F_n(x)$ has a jump $1/n$. If the value x_i is in collection $x_1, \ldots, x_n$ together k-times, then $F_n(x)$ has at point x_i jump k/n. For all values $x \geq x_{(n)}$ we have $F_n(x) = 1$.

6.2 Sample quantiles

Consider numbers $x_1, \ldots, x_n$. The sample quantiles (sometimes called *empirical quantiles*) could be defined as values of quantile function belonging to the empirical distribution function $F_n(x)$. This definition is introduced in the book Venables, Ripley (2002) on p. 108. People who use program `R` have a little different definition. We explain it using publication Verzani (2002).

Sample median is the point in the data that splits it into half. For example, data

$$10, \quad 17, \quad 18, \quad 25, \quad 28$$

have sample median 18, since two values are larger and two are less. In case of data

$$10, \quad 17, \quad 18, \quad 25, \quad 28, \quad 28 \tag{6.1}$$

median could be any number between 18 and 25, for concreteness it is taken as the average 21,5.

The p quantile (also known as the $100p\%$ - percentile) is the point in the data where $100p\%$ is less and $100(1-p)\%$ is larger. If there are n data points, then the p quantile

occurs at the position $1 + (n-1)p$ with weighted averaging if this is between integers. For example the .25 quantile of the numbers $10, 17, 18, 25, 28, 28$ occurs at the position $1 + (6-1)(.25) = 2.25$. That is 1/4 of the way between the second and third number which in this example is 17.25.

Program `R` has 9 types of *sample quantiles*. The type 7 is default. Details can be found in Hyndman, Fan (1996) and in help to function `quantile`.

Quantile corresponding to $p = 0.5$ is *median* $Q(0.5)$, for $p = 0.25$ we get *lower quartile* (or *the first quartile*) $Q(0.25)$ and for $p = 0.75$ we get *upper quartile* (or *third quartile*) $Q(0.75)$. The difference $Q(0.75) - Q(0.25)$ is *interquartile range*.

6.3 Sample correlation coefficient

Let $(X_1, Y_1)', \dots, (X_n, Y_n)'$ be a sample from a two-dimensional distribution. Denote

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

If

$$\mathsf{E}X_i = \mu_X, \quad \mathsf{E}Y_i = \mu_Y, \quad \mathsf{var}\, X_i = \sigma_X^2, \quad \mathsf{var}\, Y_i = \sigma_Y^2, \quad \mathsf{cov}(X_i, Y_i) = \sigma_{XY},$$

then according theorem 4.1 we have

$$\mathsf{E}\bar{X} = \mu_X, \qquad \mathsf{E}\bar{Y} = \mu_Y, \qquad \mathsf{E}S_X^2 = \sigma_X^2, \qquad \mathsf{E}S_Y^2 = \sigma_Y^2$$

and similarly can be proved that

$$\mathsf{E}S_{XY} = \sigma_{XY}.$$

Correlation coefficient ρ was defined in section 2.6 defined as

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}.$$

It is natural to define the sample correlation coefficient r using an analogous formula in which unknown variances and unknown covariance are substituted by their unbiased estimators. If $S_X^2 > 0$, $S_Y^2 > 0$, define

$$r = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

If $S_X^2 = 0$ or $S_Y^2 = 0$, the sample correlation coefficient is not defined. An elementary arrangement gives

$$r = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}}. \tag{6.2}$$

Schwarz inequality implies $-1 \le r \le 1$.

Theorem 6.1 *Let $(X_1, Y_1)', \dots, (X_n, Y_n)'$ be a sample from a two-dimensional normal distribution with positive variances and the correlation coefficient $\rho \in (-1, 1)$. Then r has the density*

$$f_n(r) = \frac{2^{n-3}}{(n-3)!\,\pi}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}\sum_{k=0}^{\infty}\Gamma^2\left(\frac{n+k-1}{2}\right)\frac{(2\rho r)^k}{k!} \tag{6.3}$$

for $-1 < r < 1$.

Proof. See Cramér (1946). □

Theorem 6.2 *Let assumptions of theorem 6.1 hold and let $\rho = 0$. Then the random variable*

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \tag{6.4}$$

has the distribution t_{n-2}.

Proof. The Γ function satisfies

$$\Gamma(2p) = \frac{2^{2p-1}}{\sqrt{\pi}}\Gamma(p)\Gamma\left(p+\frac{1}{2}\right), \qquad p > 0.$$

We apply this formula to $(n-3)! = \Gamma(n-2) = \Gamma\left(2\frac{n-2}{2}\right)$. For $\rho = 0$ we can write the density (6.3) in the form

$$f_n(r) = \frac{\Gamma\left(\dfrac{n-1}{2}\right)}{\Gamma\left(\dfrac{n-2}{2}\right)\sqrt{\pi}}\,(1-r^2)^{(n-4)/2}, \qquad -1 < r < 1. \tag{6.5}$$

We use the transform (6.4) and obtain the density of t_{n-2} distribution, which was derived in theorem 3.10 □

Theorem 6.2 shows how to test the hypothesis $H_0 : \rho = 0$ against alternative $H_1 : \rho \neq 0$. First, the sample correlation coefficient r and the variable T are calculated using (6.4) and in the case $|T| \geq t_{n-2}(\alpha)$ hypothesis H_0 is rejected on the level α. The assumption that the sample $(X_1, Y_1)', \dots, (X_n, Y_n)'$ is from the normal distribution is very important. Test that correlation coefficient ρ is zero is identical with the test $\beta_1 = 0$ in the regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$.

Test that coefficient ρ is zero is very frequent. From this reason critical values for r were tabulated using theorem 6.2. Since in this procedure critical values of t are used, critical value $r_n(\alpha)$ is defined as the number for which under H_0

$$\mathsf{P}[|r| \geq r_n(\alpha)] = \alpha$$

holds. This corresponds to the two-sided test. One-sided tests one obtains using elementary modification. However, a computer gives directly p-value of the test.

Let us remark that for a sample from the normal distribution in the case $\rho = 0$ for $n \geq 3$ one obtains

$$\mathsf{E}r = 0, \qquad \mathsf{var}\, r = \frac{1}{n-1}.$$

In the case $\rho \in (-1, 1)$ can be proved that

$$\mathsf{E}r = \rho + O(n^{-1}) \qquad \mathsf{var}\, r = \frac{(1-\rho^2)^2}{n} + o(n^{-1})$$

(see Cramér 1946). If the remaining members are ignored, the *transformation stabilizing variance* can be derived. We dealt with this problem in section 3.4 (see example 3.14 on p. 43). It was derived that the variable Z calculated by means of *Fisher z-transformation*

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

has the distribution with the moments

$$\mathsf{E}Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \qquad \mathsf{var}\, Z \doteq \frac{1}{n-3},$$

and this distribution with $n \to \infty$ fast converges to the normal distribution. This can be demonstrated on the following fact. We know that the normal distribution has skewness $\alpha_3 = 0$ and kurtosis $\alpha_4 = 3$. Calculations show that Z has skewness and kurtosis given by formulas

$$\alpha_3 \doteq \frac{\rho^6}{(n-1)^3}, \qquad \alpha_4 \doteq 3 + \frac{2}{n-1} + \frac{4 + 2\rho^2 - 3\rho^4}{(n-1)^2}.$$

If we want to test the hypothesis $H_0 : \rho = \rho_0$, where $\rho_0 \in (-1, 1)$ is a given number against alternative $H_1 : \rho \neq \rho_0$, we calculate first

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}, \qquad \zeta_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}.$$

If H_0 holds then the random variable

$$U = \sqrt{n-3}(Z - \zeta_0)$$

has approximately normal distribution $\mathsf{N}(0, 1)$. Thus H_0 will be rejected in the case when

$$|U| \geq u\left(\frac{\alpha}{2}\right).$$

Using z-transformation an approximate confidence interval for ρ can be constructed. If ρ is the actual value of the correlation coefficient, then

$$\mathsf{P}\left[\sqrt{n-3}\left|Z - \frac{1}{2} \ln \frac{1+\rho}{1-\rho}\right| < u\left(\frac{\alpha}{2}\right)\right] \doteq 1 - \alpha.$$

One gets from this formula that the confidence interval is

$$\left(\frac{D-1}{D+1}, \frac{H-1}{H+1}\right),$$

where

$$D = \exp\left\{2Z - \frac{2u\left(\frac{\alpha}{2}\right)}{\sqrt{n-3}}\right\}, \qquad H = \exp\left\{2Z + \frac{2u\left(\frac{\alpha}{2}\right)}{\sqrt{n-3}}\right\}.$$

Table 6.1: Concentration of milk acid

ξ_i	40	64	34	15	57	45
η_i	33	46	23	12	56	40

Example 6.3 Concentration of the milk acid in the blood of mothers (values ξ_i) and their newborn children (values η_i) was measured. The results are introduced in tab. 6.1.

Calculation gives $r = 0.935$. Critical value is equal to $r_6(0{,}05) = 0.811\,4$. Since $|r| \geq r_6(0.05)$, we reject the hypothesis that the concentration of milk acid in blood of mothers and their newborns are uncorrelated values.

Fisher z-transformation gives $Z = 1.695$. If we want to test $H_0 : \rho = 0$, we have $\rho_0 = 0$, $\zeta_0 = 0$, $U = 2.937$. Since $|U| \geq u(0.025) = 1.96$, we reject H_0 also using this test.

For the construction of confidence interval for ρ with confidence coefficient 0.95 we get $D = 3.088$, $H = 285.439$. From this we calculate that the confidence interval is (0.511, 0.993). $\diamondsuit$

6.4 Sample coefficient of multiple correlation

Let the vectors

$$\begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \boldsymbol{X}_n \end{pmatrix} \tag{6.6}$$

be a sample from $(p+1)$-dimensional distribution. Remember that the random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are p-dimensional. The sample correlation coefficient between i-th a j-th components $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ will be denoted r_{ij}. Let r_{0i} be a sample correlation coefficient between Y-th variables and i-th components of vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

Introduce *sample correlation matrices*

$$\boldsymbol{R}_{\boldsymbol{X}\boldsymbol{X}} = (r_{ij})_{i,j=1}^{p} \qquad \boldsymbol{R}_{Y\boldsymbol{X}} = (r_{0i})_{i=1}^{p}, \qquad \boldsymbol{R}_{\boldsymbol{X}Y} = \boldsymbol{R}'_{Y\boldsymbol{X}}.$$

The diagonal of the matrix $\boldsymbol{R}_{\boldsymbol{X}\boldsymbol{X}}$ contains units and the matrix is symmetric. We shall assume that it is also regular. This holds under very general conditions, see Anděl (1978).

If we substitute in (2.18) unknown theoretical correlation matrices by their sample versions, we can define *sample coefficient of multiple correlation* $r_{Y,\boldsymbol{X}} = r_{0.1,2,\ldots,k}$ as non-negative number satisfying

$$r^2_{Y,\boldsymbol{X}} = \boldsymbol{R}_{Y\boldsymbol{X}} \boldsymbol{R}^{-1}_{\boldsymbol{X}\boldsymbol{X}} \boldsymbol{R}_{\boldsymbol{X}Y}.$$

If $p = 2$, we obtain

$$r^2_{0.1,2} = \frac{r^2_{01} + r^2_{02} - 2r_{01}r_{02}r_{12}}{1 - r^2_{12}}\,.$$

Theorem 6.4 *If random vectors (6.6) are the sample from the regular normal distribution, if $n > p + 1$, and if $\rho_{Y,\boldsymbol{X}} = 0$, then the random variable*

$$Z = \frac{n-p-1}{p} \frac{r^2_{Y\boldsymbol{X}}}{1 - r^2_{Y\boldsymbol{X}}}$$

has $\mathsf{F}_{p,n-p-1}$ distribution.

Proof. See Anděl (1978). □

Theorem 6.4 is an instruction how to test the hypothesis H_0, that the variable Y does not depend on vector $\boldsymbol{X}$. If we are sure that the assumption concerning normality is fulfilled, we calculate Z. In the case $Z \geq F_{p,n-p-1}(\alpha)$ we reject H_0. The value $r^2_{Y,\boldsymbol{X}}$ is called *sample coefficient of determination*. Further the *adjusted coefficient of determination*

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}\,(1 - r^2_{Y\boldsymbol{X}})$$

is introduced. Although authors of statistical software prefer adjusted coefficient of determination, the use of R^2_{adj} may be dangerous. For example, in the case $r^2_{Y\boldsymbol{X}} < p/(n-1)$ we get $R^2_{\text{adj}} < 0$. But R^2_{adj} is an estimator of theoretical coefficient of determination $\rho^2_{Y\boldsymbol{X}}$, which is nonnegative.

Example 6.5 In year 1957 statisticians investigated expenses Y_i for food and drinks in households in dependence on number of people in the household x_i and on net earnings z_i. The data concerning 7 randomly chosen households are given in Tab. 6.2.

Table 6.2: Expenses for food and drinks

Y_i	4	3	4	1	6	4	5
x_i	4	2	4	1	5	3	4
z_i	10	8	12	3	15	12	13

The coefficient of multiple correlation which describes dependence of expenses on number of people and earnings is 0.983. This is statistically significant value, its p-value is 0.001179. ♢

6.5 Sample coefficient of partial correlation

Assume that

$$\begin{pmatrix} Y_1 \\ \boldsymbol{X}_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \boldsymbol{X}_n \\ Z_n \end{pmatrix} \tag{6.7}$$

is a random sample from a $(p+2)$-dimensional distribution. Vectors $\boldsymbol{X}_i$ have again p components. Let $r_{p+1,i}$ be sample correlation coefficient between Z-variables and i-th components of vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Denote

$$\boldsymbol{R}_{Z\boldsymbol{X}} = (r_{p+1,i})_{i=1}^{p}, \qquad \boldsymbol{R}_{\boldsymbol{X}Z} = \boldsymbol{R}'_{Z\boldsymbol{X}}.$$

In view of theorem 2.18 on p. 29 define *sample coefficient of partial correlation* $r_{Y,Z.\boldsymbol{X}}$ by the formula

$$r_{Y,Z.\boldsymbol{X}} = \frac{r_{YZ} - \boldsymbol{R}_{Y\boldsymbol{X}}\boldsymbol{R}^{-1}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}Z}}{\sqrt{\left(1 - \boldsymbol{R}_{Y\boldsymbol{X}}\boldsymbol{R}^{-1}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}Y}\right)\left(1 - \boldsymbol{R}_{Z\boldsymbol{X}}\boldsymbol{R}^{-1}_{\boldsymbol{X}\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}Z}\right)}},$$

if the denominator is not zero.

If $p = 1$, then we have

$$r_{Y,Z.X} = \frac{r_{YZ} - r_{YX}r_{ZX}}{\sqrt{(1 - r_{YX}^2)\,(1 - r_{ZX}^2)}}.$$

Theorem 6.6 *Let random vectors (6.7) be a sample from regular normal distribution. If* $\rho_{Y,Z.\boldsymbol{X}} = 0$ *and if* $n > p + 2$ *then the random variable*

$$T = \frac{r_{Y,Z.\boldsymbol{X}}}{\sqrt{1 - r_{Y,Z.\boldsymbol{X}}^2}} \sqrt{n - p - 2} \tag{6.8}$$

has $\mathtt{t}_{n-p-2}$ *distribution.*

Proof. See Anděl (1978). □

Theorem 6.6 is used for testing hypothesis $H_0 : \rho_{Y,Z.\boldsymbol{X}} = 0$. From formula (6.8) we calculate T and in the case $|T| \geq t_{n-p-2}(\alpha)$ we reject H_0. The variable $r_{Y,Z.\boldsymbol{X}}$ can be also directly compared with the critical value for the common correlation coefficient, only n must be substituted by the number $n - p$. It means that H_0 is rejected in the case

$$|r_{Y,Z.\boldsymbol{X}}| \geq r_{n-p}(\alpha).$$

Example 6.7 We use again data from example 6.5 on p. 62. First, we calculate the matrix of the partial correlation coefficients.

```
          [,1]        [,2]        [,3]
[1,] 1.0000000 0.517452512 0.831167731
[2,] 0.5174525 1.000000000 0.008143009
[3,] 0.8311677 0.008143009 1.000000000
```

Sample coefficients of partial correlation are $r_{Y,X.Z} = 0.517$ and $r_{Y,Z.X} = 0.831$. The corresponding two-sided critical value is 0.8114. We can see that there exists statistically significant dependence between expenses for food and drinks and for earning even if the influence of the number of the family is eliminated. However, on the basis of given data we cannot reject the hypothesis that expenses for food and drinks do not depend on number of people living in the household if the influence of earning is eliminated. But the sample size was very small, we used it only for illustrations. ♢

Chapter 7

Interval estimators

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ such that its distribution depends on the parameter $\boldsymbol{\theta}$. We say that the random variable S is *pivotal* (for $\boldsymbol{\theta}$), if S is function only of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ and the distribution of S does not depend on $\boldsymbol{\theta}$.

If the distribution of the vector $\boldsymbol{X}$ depends on the parameter $(\boldsymbol{\theta}, \boldsymbol{\delta})$, then S is pivotal for $\boldsymbol{\theta}$, if S is a function only of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ and in the same time the distribution of S does not depend on $(\boldsymbol{\theta}, \boldsymbol{\delta})$. In this case $\boldsymbol{\delta}$ is called the *nuisance parameter*.

Pivotal statistics can be used for constructing confidence intervals and confidence sets. Assume first that $\boldsymbol{\theta} = \theta$ is a univariate parameter and $S = S(\boldsymbol{X}, \theta)$ pivotal statistics. If $\alpha \in (0, 1)$ is given then we find such numbers q_L, q_U, that it holds

$$\mathsf{P}\{q_L \leq S(\boldsymbol{X}, \theta) \leq q_U\} = 1 - \alpha.$$

The inequality $q_L \leq S(\boldsymbol{X}, \theta) \leq q_U$ will be transformed into form

$$L(\boldsymbol{X}, q_L, q_U) \leq \theta \leq U(\boldsymbol{X}, q_L, q_U),$$

which presents the confidence interval for θ with *confidence coefficient* $1 - \alpha$. If the parameter $\boldsymbol{\theta}$ is multidimensional, then the inequality $q_L \leq S(\boldsymbol{X}, \boldsymbol{\theta}) \leq q_U$ is transformed to the form $\boldsymbol{\theta} \in Z(\boldsymbol{X}, q_L, q_U)$, which gives the confidence set Z.

Example 7.1 Let $X_1, \ldots, X_n$ $(n \geq 2)$ be a sample from $\mathsf{N}(\mu, \sigma^2)$, where $\sigma > 0$. If the parameter σ is known, then

$$S = (\bar{X} - \mu)/(\sigma/\sqrt{n}\,)$$

is pivotal statistics for μ. Variable S has distribution $\mathsf{N}(0, 1)$. Denote $u = u_{\alpha/2}$ the critical value of the distribution $\mathsf{N}(0, 1)$. Then

$$\mathsf{P}\left\{-u \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq u\right\} = 1 - \alpha.$$

This implies

$$\mathsf{P}\left\{\bar{X} - \frac{u\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{u\sigma}{\sqrt{n}}\right\} = 1 - \alpha,$$

and thus

$$\left[\bar{X} - \frac{u\sigma}{\sqrt{n}}, \bar{X} + \frac{u\sigma}{\sqrt{n}}\right]$$

is confidence interval for μ with confidence coefficient $1-\alpha$. *The length of this interval* is $\Delta = 2u\sigma/\sqrt{n}$. From equation $2u\sigma/\sqrt{n} = \delta$ we get $n = 4u^2\sigma^2/\delta^2$.

If neither parameter μ nor parameter σ are known, then $T = (\bar{X}-\mu)/(S/\sqrt{n}\,)$ is the pivotal statistics for μ, and now σ is nuisance parameter. We know, that $T \sim \mathsf{t}_{n-1}$, so that distribution of T depends neither on μ, nor on σ. In this case we have

$$\mathsf{P}\left\{-t_{n-1}(\alpha) \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1}(\alpha)\right\} = 1-\alpha.$$

Similarly we get that

$$\left[\bar{X} - \frac{t_{n-1}(\alpha)S}{\sqrt{n}}, \bar{X} + \frac{t_{n-1}(\alpha)S}{\sqrt{n}}\right]$$

is confidence interval for μ with confidence coefficient $1-\alpha$. $\diamondsuit$

Chapter 8

Testing hypotheses

Let the distribution of the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)'$ depend on an unknown parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. It is known in advance that the parameter belongs to the *parametric space* Ω, which is a subset of $\mathbb{R}_k$. This symbol Ω has nothing to do with the space Ω, which was introduced as the space of elementary events.

A detailed analysis can lead to hypothesis that $\boldsymbol{\theta}$ belongs to a subset ω of the space Ω. We are not sure with this assertion and so we call the assertion $\boldsymbol{\theta} \in \omega$ *null hypothesis*. Briefly it is denoted by $H_0 : \boldsymbol{\theta} \in \omega$. The other possibility is *alternative hypothesis* $H_1 : \boldsymbol{\theta} \notin \omega$. Equivalently one can write $H_1 : \boldsymbol{\theta} \in \Omega \setminus \omega$. If the set ω has just one element, we say that the hypothesis H_0 is *simple*. If the set $\Omega \setminus \omega$ has one element, we say that the hypothesis H_1 is *simple*.

The test of hypothesis H_0 on the basis of vector $\boldsymbol{X}$ is usually made in the following way. A suitable set $W \in \mathcal{B}_n$ is found such that is called *critical set*. If the event $\{\boldsymbol{X} \in W\}$ is realized, then we reject H_0. Otherwise we do not reject H_0. In this decision one of the following cases is realized.

(i) H_0 is valid and the test do not reject it. The decision is correct.
(ii) H_0 is not valid and the test rejects it. The decision is correct.
(iii) H_0 is valid and the test rejects it. We say that it is *error of the first kind*.
(iv) H_0 is not valid and the test does not reject it. We say that it is *error of the second kind*.

Assume that we test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against alternative $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. Let the test have critical set W. Then $\beta(\boldsymbol{\theta}_1) = \mathsf{P}(\boldsymbol{X} \in W | \boldsymbol{\theta}_1)$ is called *power of the test*.

We show connection between testing hypotheses and confidence intervals. Let θ be univariate parameter and $S = S(\boldsymbol{X}, \theta)$ pivotal statistics. Let

$$W = \{S(\boldsymbol{X}, \theta_0) \le c_1, S(\boldsymbol{X}, \theta_0) \ge c_2\}, \qquad c_1 < c_2,$$

be critical set for testing hypothesis $H_0 : \theta = \theta_0$, such that

$$\mathsf{P}\{S(\boldsymbol{X}, \theta_0) \le c_1 | \theta_0\} = \alpha_1, \qquad \mathsf{P}\{S(\boldsymbol{X}, \theta_0) \ge c_2 | \theta_0\} = \alpha_2, \qquad \alpha_1 + \alpha_2 = \alpha.$$

Since S is pivotal, we have

$$\mathsf{P}\{S(\boldsymbol{X}, \theta_0) \le c_1\} = \alpha_1, \qquad \mathsf{P}\{S(\boldsymbol{X}, \theta_0) \ge c_2\} = \alpha_2.$$

Assume that inequality $c_1 < S(\boldsymbol{X}, \theta) < c_2$ is equivalent to inequality $w_1 < \theta < w_2$, where $w_1 = w_1(\boldsymbol{X})$, $w_2 = w_2(\boldsymbol{X})$. Then

$$1 - \alpha = \mathsf{P}\{c_1 < S(\boldsymbol{X}, \theta) < c_2\} = \mathsf{P}\{w_1 < \theta < w_2\}.$$

Interval (w_1, w_2) is confidence interval for θ with confidence coefficient $1 - \alpha$ and it is complement to the critical set for testing H_0.

Statistical thinking is based on the following principle, which is applied in practical life. If an event has in the experiment only very small probability,we behave as if this event cannot come. The size of this probability denoted by α depends on consequences which the event would have. In practical situations one chooses $\alpha = 0{,}05$ and we shall do this in our text. Hypothesis H_0 is chosen so that the error of the first kind is more important than error of the second kind. If we do not want to accept the error of the first kind, we must define critical set W so that the probability of error of the first kind would not be larger than α. The best choice of the critical test W is such that under this condition the probability of the second kind is minimal. Number $\sup_{\boldsymbol{\theta} \in \omega} \mathsf{P}(\boldsymbol{X} \in W)$ is called *level of the test*. In some cases the best critical set can be found using Neyman-Pearson lemma.

Lemma 8.1 (Neyman-Pearson) *Let $\omega = \{\theta_0\}$, $\Omega \setminus \omega = \{\theta_1\}$, so that both hypotheses H_0 and H_1 are simple. Let $\boldsymbol{X}$ have density p_0 when H_0 is valid and density p_1 under H_1. Let for a given probability $\alpha \in (0, 1)$ there exists a number c such that for the set*

$$W_0 = \{\boldsymbol{x} \,:\, p_1(\boldsymbol{x}) \geq c\, p_0(\boldsymbol{x})\} \tag{8.1}$$

we have

$$\int_{W_0} p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \alpha.$$

Then for an arbitrary set $W \in \mathcal{B}_n$ which fulfills condition

$$\int_{W} p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \alpha$$

it holds

$$\int_{W_0} p_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \geq \int_{W} p_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

Proof. We have

$$\begin{aligned}
\int_{W_0} p_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int_{W} p_1(\boldsymbol{x})\, \mathrm{d}x &= \int_{W_0 \setminus W} p_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int_{W \setminus W_0} p_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&\geq \int_{W_0 \setminus W} c\, p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int_{W \setminus W_0} c\, p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\
&= \int_{W_0} c\, p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \int_{W} c\, p_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = c\alpha - c\alpha = 0. \qquad \square
\end{aligned}$$

It follows from the lemma that from all critical sets ensuring probability of the first kind equal to α it is W_0 which has the smallest probability of the second kind.

Thus W_0 is the *best critical set*. The set (8.1) can be explained as follows. If for the given $\boldsymbol{x}$ the number $p_1(\boldsymbol{x})$ is substantially larger than number $p_0(\boldsymbol{x})$, then we conclude that the parameter is θ_1.

On the other side, if $p_0(\boldsymbol{x})$ is considerably larger than $p_1(\boldsymbol{x})$, we decide that the parameter is θ_0. It is also problem of estimating parameters and the lemma 8.1 shows that the reasonable procedure can be based on comparison of densities with given value $\boldsymbol{x}$. This leads to the *maximum likelihood method*.

The critical set is often defined so that a test statistics $S = S(\boldsymbol{X})$ is larger (or smaller) than a calculated value. In such case it is often used *p-value*. It is the probability of obtaining a value as numerically large as or larger than the observed statistics.

Chapter 9

One-sample problem

9.1 Descriptive statistics

9.1.1 Graphs

Boxplot or more descriptively box-and-whiskers plot is a graphical summary of a distribution. The box in the middle indicates hinges, nearly quartiles, and median. The lines (whiskers) show the largest and the smallest observation that falls within a distance of 1.5 times the box size from the nearest hinge. If any observations fall farther away, the additional points are considered extreme values and are shown separately.

If the data are not too many, we can show them on the left (or right) hand side of the graph. In this way we obtain *rugplot*.

In the case of many data a picture of *histogram* is useful. Also other graphs are used, for example *stripchart*.

9.2 One-sample Kolmogorov-Smirnov test

In this section we deal with *one-sample Kolmogorov-Smirnov test*. Let $X_1, \ldots, X_n$ be a sample from a distribution with continuous distribution function.

Introduce random variables

$$\xi_i(x) = \begin{cases} 1, & \text{if} \quad X_i \leq x, \\ 0, & \text{if} \quad X_i > x \end{cases}$$

for $i = 1, \ldots, n$. Define

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \xi_i(x). \tag{9.1}$$

Function $F_n(x)$ je *empirical distribution function*. Given a sample, it is the same as the empirical distribution function which was introduced in section 6.1 on p. 57. We prove that with growing n the function $F_n(x)$ approaches to the true distribution function $F(x)$.

Theorem 9.1 *Pro every real x we have*

$$F_n(x) \to F(x) \qquad \textit{almost surely for} \quad n \to \infty.$$

Proof. For every fixed x the variables $\xi_i(x)$ are independent and identically distributed. They satisfy

$$\mathsf{P}[\xi_i(x) = 1] = F(x), \qquad \mathsf{E}\xi_i(x) = F(x).$$

Since $F_n(x)$ is defined by the formula (9.1), the assertion follows from Kolmogorov theorem. □

It is possible to prove a stronger assertion.

Theorem 9.2 (Glivenko-Cantelli) *Denote* $D_m = \sup_x |F_m(x) - F(x)|$. *Then it holds*

$$\mathsf{P}\left(\lim_{m\to\infty} D_m = 0\right) = 1.$$

Proof. See Gnedenko (1954). □

We want to test hypothesis H_0 that this distribution function is F. Let F_n be empirical distribution function corresponding to the sample $X_1, \ldots, X_n$. Define

$$D_n = \sup_x |F_n(x) - F(x)|. \tag{9.2}$$

With respect to theorems 9.1 and 9.2 large values of the variable D_n will give evidence against the hypothesis H_0. If n is small, special tables of critical values are used. If n is larger, the limit formula is applied, namely that

$$\lim_{n\to\infty} \mathsf{P}(\sqrt{n}\, D_n < \lambda) = K(\lambda),$$

where

$$K(\lambda) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2\lambda^2). \tag{9.3}$$

For large n critical values $D_n(\alpha)$ for random variable D_n are approximated by means of

$$D_n(\alpha) \doteq \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}} \tag{9.4}$$

(see Likeš, Laga 1978). It means that H_0 is rejected when $D_n \geq D_n(\alpha)$.

Notice that calculation of variable D_n can be restricted to those values of x, in which F_n has a jump. In these points it is necessary to take into account not only the difference $F_n(x) - F(x)$, but also the limit from the right.

It is necessary to emphasize that the hypothesis H_0 must determine the distribution function F completely including all parameters.

Kolmogorov-Smirnov test can be used for testing the hypothesis that the random sample $X_1, \ldots, X_n$ is from rectangular distribution $\mathsf{R}(0, 1)$. This is suitable for testing generators of random numbers. The test cannot be used for testing hypothesis that the sample is from the normal distribution. This hypothesis does not specify parameters μ a σ^2. In the case that we estimate the parameters from the sample and the function F take distribution function of the normal distribution with estimated parameters, the distribution of the test statistics D_n would be considerably changed. From this reason the critical values were determined using simulations. For the normal distribution see Lilliefors (1967) and Iman (1982), for the exponential distribution see Lilliefors (1969) and Iman (1982). The generalization to samples from discrete distributions is introduced

in papers Conover (1972) and Mantel (1974). Test of hypothesis that the sample is from the Poisson distribution is described in Campbell, Oprian (1979). Kolmogorov-Smirnov test was generalized to censored samples (see Barr, Davison 1973).

Example 9.3 We simulated the sample with size 20 from the rectangular distribution $\mathsf{R}(0,1)$.

```
0.29 0.79 0.41 0.88 0.94 0.05 0.53 0.89 0.55 0.46 0.96 0.45 0.68
0.57 0.10 0.90 0.25 0.04 0.33 0.95
```

The one-sample Kolmogorov-Smirnov test gives the results

```
D = 0.183, p-value = 0.4604
alternative hypothesis: two-sided
```

Since the p-value is larger than 0.05, the hypothesis about sample from the rectangular distribution is not rejected. See Fig. 9.1 $\diamondsuit$.

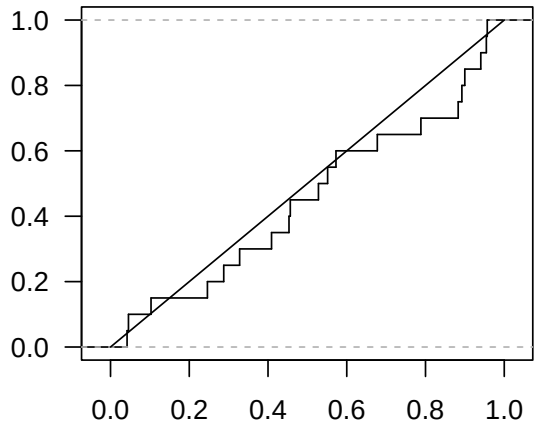

Figure 9.1: Kolmogorov-Smirnov one-sample test

9.3 One-sample t test

Let $X_1, \ldots, X_n$ be a random sample from z $\mathsf{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ and $n \geq 2$. Assume that no parameter μ and σ^2 is known. We shall investigate a test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

Theorem 9.4 *If the true expectation of the normal distribution is μ, then the random variable*

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n}$$

has t_{n-1} distribution.

Proof. It follows from Theorem 4.6 on p. 48 that

$$X = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim \mathsf{N}(0,1), \qquad Z = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and that X a Z are independent. We use theorem 40 which says that

$$X \Big/ \sqrt{Z/(n-1)} \sim \mathsf{t}_{n-1}.$$

We easily get $X \Big/ \sqrt{Z/(n-1)} = T$. □

Critical value $t_k(\alpha)$ of Student t distribution with k degrees of freedom is defined so that probability of its exceeding by the random variable with distribution t_k is $\alpha/2$. It means that in the case $T \sim \mathsf{t}_k$ is

$$\mathsf{P}\{|T| \geq t_k(\alpha)\} = \alpha.$$

This is the difference from definition of most other distributions.

Let us return to the problem formulated at the beginning of this section. If H_0 holds, then using theorem 9.4 and definition of critical values of t distribution we have

$$\mathsf{P}\left[\frac{|\bar{X} - \mu_0|}{S}\sqrt{n} \geq t_{n-1}(\alpha)\right] = \alpha. \tag{9.5}$$

Practical procedure for testing H_0 against H_1 is the following:

(i) Calculate value $T_0 = \dfrac{\bar{X} - \mu_0}{S}\sqrt{n}$.

(ii) If $|T_0| \geq t_{n-1}(\alpha)$, reject H_0. Otherwise do not reject H_0.

Formula (9.5) ensures that the level of the test is α.

If we consider the complementary event then from formula (9.5) we get

$$1 - \alpha = \mathsf{P}\left[\bar{X} - \frac{S}{\sqrt{n}}t_{n-1}(\alpha) < \mu < \bar{X} + \frac{S}{\sqrt{n}}t_{n-1}(\alpha)\right].$$

Thus the interval $\bar{X} \mp n^{-1/2}St_{n-1}(\alpha)$ is the confidence interval for μ for unknown σ^2 with confidence coefficient $1 - \alpha$.

One-sided tests and one-sided confidence intervals can be derived similarly as for the normal distribution with known variance. The results differ so that instead of σ^2 we have S^2 and instead of value $u(\alpha)$ we have $t(2\alpha)$.

Example 9.5 An automaton fills boxes with a washing powder. Each box should contain 1 kg of the powder. Five boxes were randomly taken and their contain exactly weighted. The following departures (in dkg) were found:

$$-3 \qquad 2 \quad -2 \qquad 0 \quad -1.$$

It should be detected if there is systematic departure from the considered value.

We consider our data as random sample from z $\mathsf{N}(\mu, \sigma^2)$, If there is no systematic departure, the hypothesis $H_0 : \mu = 0$ holds. If there is a departure, the hypothesis $H_1 : \mu \neq 0$ will hold. We have

$$n = 5, \qquad \bar{X} = -0.8, \qquad S^2 = 3.7, \qquad T_0 = -0.9300.$$

Since $|T_0| < t_4(0.05) = 2.776$, the hypothesis H_0 cannot be rejected. Two-sided confidence interval for the parameter μ with the confidence coefficient 0,95 is (-3.19, 1.59). Notice that it contains value $\mu_0 = 0$. ♢

9.4 Variance

Consider the random sample $X_1, \ldots, X_n$ from $\mathsf{N}(\mu, \sigma^2)$, where $n > 1$ and no parameter μ and $\sigma^2 > 0$ is known. We shall deal with the test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$. Let $\chi_f^2(\alpha)$ be *critical value of distribution* χ_f^2. It is the number which the random variable with distribution χ_f^2 exceeds with probability α. Test is based on the variable $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$, which is unbiased estimator of the parameter σ^2 and its distribution does not depend on μ. We choose critical set $S^2 \geq k$. If H_0 is true and we want

$$\alpha = \mathsf{P}(S^2 \geq k) = \mathsf{P}\left\{\frac{(n-1)S^2}{\sigma_0^2} \geq \frac{(n-1)k}{\sigma_0^2}\right\},$$

we must have

$$(n-1)k/\sigma_0^2 = \chi_{n-1}^2(\alpha),$$

so that

$$k = \sigma_0^2 \chi_{n-1}^2(\alpha)/(n-1).$$

If the true variance is σ^2, we obtain

$$\mathsf{P}\left\{\frac{(n-1)S^2}{\chi_{n-1}^2(\alpha)} < \sigma^2\right\} = 1 - \alpha.$$

Thus

$$\left((n-1)S^2/\chi_{n-1}^2(\alpha), \infty\right)$$

is the right-hand side confidence interval for σ^2.

If we want to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$, we reject H_0 in the case that either $S^2 \leq k_1$, or $S^2 \geq k_2$. The statisticians usually choose k_1 and k_2 so that

$$\mathsf{P}(S^2 \leq k_1) = \frac{\alpha}{2}, \quad \mathsf{P}(S^2 \geq k_2) = \frac{\alpha}{2}.$$

This will be satisfied when

$$k_1 = \frac{\sigma_0^2 \chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)}{n-1}, \qquad k_2 = \frac{\sigma_0^2 \chi_{n-1}^2\left(\frac{\alpha}{2}\right)}{n-1}.$$

If the true variance is σ^2, then

$$\mathsf{P}\left\{\frac{(n-1)S^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)}\right\} = 1 - \alpha.$$

It implies that

$$\left(\frac{(n-1)S^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)}, \frac{(n-1)S^2}{\chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)}\right)$$

is the two-sided confidence interval for σ^2 with confidence coefficient $1 - \alpha$.

The two-sided confidence interval satisfying condition $\mathsf{P}(S^2 \leq k_1) = \mathsf{P}(S^2 \geq k_2) = \frac{\alpha}{2}$ is not optimal in the sense that its length is not shortest possible. If the density of the corresponding random variable is unimodal and smooth, the optimal confidence interval may be calculated using theorem 9.6 (see Farnsworth 2004). There are two

complementary problems. First, we would find the interval of the given length which has maximal probability. In the second problem we want to find the shortest interval which has the given probability. It is clear that the solution of the second problem leads to the solution of the first one.

Theorem 9.6 *Let the random variable X have absolutely continuous distribution function F and the differentiable density f. Let $w > 0$ be the given number. Consider all couples $(x, x+w)$ such that $f'(x) > 0$, $f'(x+w) < 0$. If $f(x) = f(x+w)$ holds, then the interval $[x, x+w]$ has locally maximal probability among all intervals of length w.*

Proof. Define

$$A(x) = \mathsf{P}(X \le x + w) - \mathsf{P}(x \le x) = F(x+w) - F(x).$$

When w is fixed, $A(x)$ depends only on x. We have $A'(x) = f(x+w) - f(x)$, so that the condition $A'(x) = 0$ gives $f(x+w) = f(x)$. Since $A''(x) = f'(x+w) - f'(x) < 0$, local maximum of $A(x)$ is at point x for which $f(x+w) = f(x)$. $\square$.

In many cases we have $f(x) = 0$ for $x < 0$, $f(x)$ increasing on $(0, x_0)$ and decreasing on (x_0, ∞). Then $A(x)$ has exactly one local maximum which is also global maximum.

9.5 Sign test

Let $X_1, \dots, X_n$ be a sample from a continuous distribution with *median* $\tilde{x}$. It means that

$$\mathsf{P}(X_i < \tilde{x}) = \mathsf{P}(X_i > \tilde{x}) = \frac{1}{2}, \qquad i = 1, \dots, n.$$

We test the hypothesis $H_0 : \tilde{x} = x_0$, where x_0 is a given number. We start with the two-sided test where the alternative hypothesis is $H_1 : \tilde{x} \ne x_0$. We form the differences $X_1 - x_0, \dots, X_n - x_0$. If some differences are zeros then they are dropped. Number of positive differences will be denoted as Y. If H_0 holds, Y has *binomial distribution* $\mathsf{Bi}(n, \frac{1}{2})$. Hypothesis H_0 will be rejected, if Y is nearly zero, or nearly n. In such cases tables of critical values k_1 a k_2 can be used. They satisfy

$$\mathsf{P}(Y \le k_1) \le \frac{\alpha}{2}, \qquad \mathsf{P}(Y \ge k_2) \le \frac{\alpha}{2}. \tag{9.6}$$

Here k_1 is the largest and k_2 the smallest number for which (9.6) is valid. Because the distribution $\mathsf{Bi}(n, \frac{1}{2})$is symmetric, we have $k_2 = n - k_1$.

Hypothesis H_0 will be rejected in the cases when $Y \le k_1$ or $Y \ge k_2$. The level of test is maximally α. Usually, it is considerable smaller than α.

Introduce random variables $\xi_1, \dots, \xi_n$ so that $\xi_i = 0$ in case $X_i - x_0 \le 0$ and $\xi_i = 1$ in the case $X_i - x_0 > 0$. Thus $Y = \xi_1 + \dots + \xi_n$. Since $\mathsf{E}\xi_i = \frac{1}{2}$, $\mathsf{var}\,\xi_i = \frac{1}{4}$, central limit theorem ensures that the variable $\left(Y - \frac{n}{2}\right) / \sqrt{n}$ has asymptotically distribution $\mathsf{N}(0, \frac{1}{4})$. Thus the variable

$$U = \frac{2Y - n}{\sqrt{n}} \tag{9.7}$$

has asymptotically distribution $\mathsf{N}(0, 1)$. The hypothesis H_0 will be rejected when $|U| \ge u(\frac{\alpha}{2})$. The level of this test tends to α. The procedure is used if $n \ge 20$. Practically, H_0 is rejected when $U^2 \ge \chi^2_1(\alpha)$.

The sign test is used specially in the case when the distribution of random variables X_i is considerably skew. However, the power of this test as rather small. Since this test has considerably large probability of the error of the second kind, it is recommended to have in the disposal larger number of observations n.

If from the technical reasons some differences $X_i - x_0$ are zeros then these values are left out and number n is correspondingly decreased.

The sign test can be one-sided. *Multidimensional sign test* is introduced in Bennet (1962).

Example 9.7 Ten persons should independently estimate time one minute. The following results (in seconds) were registered:

$$53, \quad 48, \quad 45, \quad 55, \quad 63, \quad 51, \quad 66, \quad 56, \quad 50, \quad 58.$$

We have $\tilde{x} = 60$, $n = 10$, number of differences with positive sign is $Y = 2$. For $n = 10$ and $\alpha = 0{,}05$ we have $k_1 = 1$, $k_2 = 9$. Since $Y \in (k_1, k_2)$, the sign test does not reject hypothesis H_0. It can be calculated that in this case the actual level of the test is not 0,05, but only 0,02. But the test which would use values $k_1 = 2$, $k_2 = 8$, would have level 0,11.

$\diamondsuit$

9.6 One-sample Wilcoxon test

Let $X_1, \ldots, X_n$ be random sample from a continuous distribution with the density f, which is symmetric around the point a. Then we have $f(a + x) = f(a - x)$. Then obviously a must be *median* $\tilde{x}$. If there exists finite expectation of this distribution then for each i it must hold $\mathsf{E}X_i = a$. Finiteness of the expectation is not assumed, however. The one-sample Wilcoxon test is designed to testing hypothesis $H_0 : \tilde{x} = x_0$ against the alternative hypothesis $H_1 : \tilde{x} \neq x_0$.

Assume that no variable X_i is equal to x_0. Define $Y_i = X_i - x_0$. The variables Y_i order into non-decreasing sequence with respect to their absolute value

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \cdots \leq |Y|_{(n)}.$$

Let R_i^+ be the rank of variable $|Y_i|$. Define

$$S^+ = \sum_{Y_i \geq 0} R_i^+, \qquad S^- = \sum_{Y_i < 0} R_i^+.$$

It holds $S^+ + S^- = n(n+1)/2$. If the number $\min(S^+, S^-)$ is smaller or equal to critical value $w_n(\alpha)$, we reject H_0.

Our assumptions imply that $Y_1, \ldots, Y_n$ are independent identically distributed random variables and their distribution is symmetric around zero.

Theorem 9.8 *Vectors* $(\operatorname{sign} Y_1, \ldots, \operatorname{sign} Y_n)'$ *and* $(|Y|_{(1)}, \ldots, |Y|_{(n)})'$ *are independent.*

Proof. Since the variables Y_i are independent, vectors $(\operatorname{sign} Y_i,\ |Y_i|)'$ are also independent. From continuity and symmetry of distribution it follows that

$$\mathsf{P}(\operatorname{sign} Y_i = 1) = \mathsf{P}(\operatorname{sign} Y_i = -1) = \frac{1}{2}.$$

For an arbitrary $y > 0$ we have

$$\begin{aligned}\mathsf{P}(\operatorname{sign} Y_i = 1, |Y_i| < y) &= \mathsf{P}(0 < Y_i < y) = \frac{1}{2}\mathsf{P}(-y < Y_i < y)\\ &= \frac{1}{2}\mathsf{P}(|Y_i| < y) = \mathsf{P}(\operatorname{sign} Y_i = 1)\,\mathsf{P}(|Y_i| < y).\end{aligned}$$

Thus the variables $\operatorname{sign} Y_i$ and $|Y_i|$ are for every i independent. The result is that the vectors $(\operatorname{sign} Y_1, \dots, \operatorname{sign} Y_n)'$ and $(|Y_1|, \dots, |Y_n|)'$ are independent. Since the vector $(|Y|_{(1)}, \dots, |Y|_{(n)})'$ is a function of the vector $(|Y_1|, \dots, |Y_n|)'$, the proof is finished. □

Theorem 9.9 *Let* $S = \sum_{i=1}^{n} R_i^+ \operatorname{sign} Y_i$. *Then*

$$S^+ = \frac{1}{2}S + \frac{n(n+1)}{4}.$$

Proof. We have $S^+ - S^- = S$, $S^+ + S^- = n(n+1)/2$. From here S^+ can be easily calculated. □

Theorem 9.10 *If* H_0 *holds, then*

$$\mathsf{E}S^+ = \frac{1}{4}n(n+1), \qquad \mathsf{var}\, S^+ = \frac{1}{24}n(n+1)(2n+1).$$

Proof. First we notice that $\mathsf{E}\operatorname{sign} Y_i = 0$ for every i. From theorem 9.8 we get that $\mathsf{E}(R_i^+ \operatorname{sign} Y_i) = (\mathsf{E}R_i^+)(\mathsf{E}\operatorname{sign} Y_i)$, and thus

$$\mathsf{E}(R_i^+ \operatorname{sign} Y_i) = 0. \tag{9.8}$$

From here

$$\mathsf{E}S = \sum_{i=1}^{n} \mathsf{E}(R_i^+ \operatorname{sign} Y_i) = 0.$$

In view of (9.8) we have

$$\begin{aligned}\mathsf{var}(R_i^+ \operatorname{sign} Y_i) &= \mathsf{E}(R_i^+ \operatorname{sign} Y_i)^2 = \mathsf{E}(R_i^+)^2\mathsf{E}(\operatorname{sign} Y_i)^2 = \mathsf{E}(R_i^+)^2\\ &= 1^2\frac{1}{n} + 2^2\frac{1}{n} + \cdots + n^2\frac{1}{n} = \frac{1}{6}(n+1)(2n+1).\end{aligned}$$

Similarly can be proved that

$$\mathsf{cov}(R_i^+ \operatorname{sign} Y_i, R_j^+ \operatorname{sign} Y_j) = 0 \qquad \text{for } i \neq j.$$

Now, we apply theorem 2.4 on p. 22. □

It can be proved (see Hájek, Šidák 1967), that S^+ has asymptotically normal distribution. The hypothesis H_0 can be tested by means of statistics

$$U = \frac{S^+ - \mathsf{E}S^+}{\sqrt{\mathsf{var}\, S^+}},$$

where $\mathsf{E}S^+$ and $\mathsf{var}\, S^+$ are introduced in theorem 9.10. If $|U| \geq u(\frac{\alpha}{2})$, hypothesis H_0 is rejected on the level, which is approaching to α as $n \to \infty$.

It is necessary to emphasize that one of the assumptions of the on-sample Wilcoxon test is the symmetry of the density f around the median. Hypothesis H_0 can be correctly rejected also in the case when the median is equal to x_0, but the density f is strongly non-symmetric.

If some of the variables X_i is equal to x_0, usually this observation is left out.

Example 9.11 We show the analysis of the data introduced in example 9.7. We remember that estimates of time of one minute (in seconds) were

$$53, \quad 48, \quad 45, \quad 55, \quad 63, \quad 51, \quad 66, \quad 56, \quad 50, \quad 58.$$

If the median of this distribution is $x_0 = 60$ seconds, we would obtain variables $Y_i = X_i - 60$. They are equal to

$$-7, \quad -12, \quad -15, \quad -5, \quad 3, \quad -9, \quad 6, \quad -4, \quad -10, \quad -2.$$

We order the variables in a non-decreasing series with respect to their absolute values. We get

$$-2, \quad 3, \quad -4, \quad -5, \quad 6, \quad -7, \quad -9, \quad -10, \quad -12, \quad -15.$$

The order of the number 3 is 2, the order of 6 has the order 5. Thus $S^+ = 2 + 5 = 7$. We get $S^- = 10 \times 11/2 - S^+ = 48$. Critical value is $w_{10}(0{,}05) = 8$. Since $\min(S^+, S^-) = 7 \leq w_{10}(0{,}05) = 8$, we reject the hypothesis that in human population half persons the length of one minute undervalues and half overvalues. The asymptotic procedure would give

$$\mathsf{E}S^+ = 27{,}5, \qquad \mathsf{var}\, S^+ = 96{,}25, \qquad U = -2{,}09.$$

Since $|U| \geq u(0.025) = 1.96$, hypothesis H_0 would be rejected also by this procedure.

♢

9.7 Hodges-Lehmann estimator

Let $X_1, \ldots, X_n$ be random variables and $X_{(1)} \leq \cdots \leq X_{(n)}$ the ordered random sample. Define

$$\tilde{X} = \begin{cases} X_{(k+1)}, & \text{if } n = 2k+1, \\ [X_{(k)} + X_{(k+1)}]/2, & \text{if } n = 2k. \end{cases}$$

Then $\tilde{X}$ is called *median* of random variables $X_1, \ldots, X_n$. We use denotation $\tilde{X} = \text{median}(X_1, \ldots, X_n)$.

Let X_1, X_2 be independent random variables with identical distribution function F, which has median $\tilde{x}$. Consider the variable $Y = (X_1 + X_2)/2$ and denote G the distribution

function of the variable Y. Then $\tilde{y}$ is *pseudomedian* of the distribution, which has the distribution function F. Notice that pseudomedian was introduced in the paper Høyland (1965).

Distribution of the random variable X is *symmetric about the point* μ, if we have for all real x

$$\mathsf{P}(X \geq \mu + x) = \mathsf{P}(X \leq \mu - x).$$

The distribution is called *symmetric*, if there exists such μ that this distribution is symmetric about μ.

Theorem 9.12 *If the distribution defined by the distribution function F is symmetric and if there exist median and pseudomedian uniquely, then median and pseudomedian are the same. If the distribution that corresponds to the distribution function F is not symmetric then generally $\tilde{x} \neq \tilde{y}$.*

Proof. Assume without loss of generality that the distribution given by the distribution function F is symmetric about zero. A necessary and sufficient condition for it is that its characteristic function is real. Because characteristic function of the sum $X_1 + X_2$ is the product of characteristic functions of individual summands, the distribution of the sum $X_1 + X_2$ is symmetric about zero. If median and pseudomedian are defined uniquely, each of them must be zero.

Let X_1 and X_2 be independent identically distributed random variables with exponential distribution with the parameter $\lambda = 2$, which has density f and the distribution function F given by formulas

$$f(x) = \frac{1}{2}\mathrm{e}^{-x/2}, \qquad F(x) = 1 - \mathrm{e}^{-x/2}, \qquad x > 0.$$

Median of this distribution $\tilde{x}$ is the root of the equation $F(x) = 1/2$. We get $\tilde{x} = 2\ln 2 = 1.386294$.

Now, we take into account that f is the density of the distribution χ^2_2. Consequently, the variable $Z = X_1 + X_2$ has distribution χ^2_4 with density $g(z) = \frac{1}{4}z\mathrm{e}^{-z/2}$ for $z > 0$. Let $Y = Z/2$. Then Y has density $h(y) = y\mathrm{e}^{-y}$ for $y > 0$ and the distribution function

$$H(y) = 1 - (1 + y)\mathrm{e}^{-y}, \qquad y > 0.$$

Median $\tilde{y}$ of the distribution with the distribution function H one gets by solving equation $H(y) = 1/2$. The result is $\tilde{y} = 1.67835$. Median of the distribution $\mathsf{Ex}(2)$ is not the same as pseudomedian of this distribution. □

Let us return to the original formulation of the test. We defined $Y_i = X_i - x_0$, and the unknown median of the distribution X_i is a. If it holds $x_0 = a$, from theorem 9.10 we would have $\mathsf{E}S^+ = n(n+1)/4$. It suggest to take as estimator of parameter a such value $x_0 = \hat{a}$,for which we should have $S^+ = n(n+1)/4$. The solution is to choose

$$\hat{a} = \operatorname{medin}\left\{\frac{X_i + X_j}{2}, \quad i \leq j\right\}.$$

Each of $n(n+1)/2$ means $(X_i + X_j)/2$ is called *Walsh mean* (Walsh 1949). The variable $\hat{a}$ is called *Hodges-Lehmann estimator* (it belongs to the class of estimators introduced

in paper Hodges, Lehmann 1963). Based on principles leading to Hodges-Lehmann estimator it is possible to construct nonparametric confidence interval for pseudomedian (see Hollander, Wolfe 1973, p. 35). We remark, that the Hodges-Lehmann estimator and nonparametric confidence interval can be constructed also in the case of two samples in connection with two-sample Wilcoxon test (see Hollander, Wolfe 1973, p. 75).

Example 9.13 We continue with the example 9.11. The calculations are made by program `r`.

```
p-value = 0.03711,
95 percent confidence interval: 50.0 59.5
sample estimates: (pseudo)median
                  54
```

Hodges-Lehmann estimator for pseudomedian is 54, the confidence interval for pseudomedian with confidence level 0.95 is $(50.0, 59.5)$. $\diamondsuit$

Chapter 10

Discrete one-sample problem

10.1 Confidence intervals and tests for p

10.1.1 Tests in binomial distribution

Assume that $X \sim \mathsf{Bi}(n, p)$. We describe a test $H_0 : p = p_0$ against $H_1 : p \neq p_0$, where p_0 is given value from interval $(0, 1)$. The sign test was a special case of this problem with $p_0 = \frac{1}{2}$.

If n is a small number, we calculate for the given α the largest integer k_1 such that

$$S_1(k_1) = \sum_{i=0}^{k_1} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2}$$

and the smallest integer k_2 such, that

$$S_2(k_2) = \sum_{i=k_2}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \frac{\alpha}{2}.$$

Hypothesis H_0 will be rejected on the level maximally α, if $X \leq k_1$ or $X \geq k_2$.

A direct calculation of sums $S_1(k_1)$ and $S_2(k_2)$ can be mathematically very difficult. A better procedure is to use the relation

$$\sum_{i=0}^{s} \binom{n}{i} p^i (1-p)^{n-i} = F^*_{2(n-s),2(s+1)} \left[\frac{(s+1)(1-p)}{p(n-s)} \right], \tag{10.1}$$

which is valid for $s = 0, 1, \ldots, n-1$; here $F^*_{m,n}(x)$ is distribution function of the Fisher-Snedecor distribution. See Ling (1992), Peizer, Pratt (1968) and Pratt (1968).

This method is equivalent to the following procedure. Denote

$$D = \frac{X}{X + (n - X + 1) F_{2(n-X+1),2X} \left(\frac{\alpha}{2} \right)},$$

$$H = \frac{(X+1) F_{2(X+1),2(n-X)} \left(\frac{\alpha}{2} \right)}{n - X + (X+1) F_{2(X+1),2(n-X)} \left(\frac{\alpha}{2} \right)},$$

where $F_{m,n}(\beta)$ denotes the critical value of the $\mathsf{F}_{m,n}$ distribution on the level β. If $0 < X < n$, then (D, H) is confidence interval for parameter p with the confidence coefficient $1-\alpha$. In the case $p_0 \notin (D, H)$, we reject the hypothesis H_0. Here $(D, 1)$ is *right-hand side confidence interval* for parameter p with confidence coefficient $1-\frac{\alpha}{2}$ (when $X = n$) and $(0, H)$ is *left-hand side confidence interval* for p with confidence coefficient $1-\frac{\alpha}{2}$ (also when $X = 0$).

Example 10.1 Simonoff (2003) on p. 66 introduces that among the 13 patients, that had Lymski borelioza, to a given diagnostic test 8 reacted. Estimated sensitivity of the test is 8/13=61.5 %. $\diamondsuit$

10.1.2 Wald confidence interval

Let $X \sim \mathsf{Bi}(n, p)$. Let u_α be critical value of the distribution $\mathsf{N}(0, 1)$ on level α. If value $X = x$ ($x \neq 0$, $x \neq n$), is observed, then the maximum-likelihood estimator of the parameter p is $\hat{p} = x/n$.

Denote $q = 1-p$, $\hat{q} = 1-\hat{p}$,

$$\xi = \frac{\hat{p}-p}{\sqrt{pq/n}}, \qquad Z = \frac{\hat{p}-p}{\sqrt{\hat{p}\hat{q}/n}}.$$

Then $\hat{p} \xrightarrow{\mathsf{P}} p$, $\hat{q} \xrightarrow{\mathsf{P}} q$, $\xi \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1)$. Since

$$Z = \frac{\frac{\hat{p}-p}{\sqrt{pq/n}}}{\sqrt{\frac{\hat{p}\hat{q}}{pq}}},$$

from Cramér-Slucky theorem it follows that $Z \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1)$. From this reason an approximate confidence interval for p with confidence coefficient $1-\alpha$ is equal

$$\hat{p} \pm u_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}. \tag{10.2}$$

It is so called *Wald confidence interval* or *standard confidence interval*.

It is not recommended to use the Wald interval when n is small. The problem is that the confidence coefficient can substantially differ from its nominal value and in many cases is rather small.

10.1.3 Wilson confidence interval

Because of above mentioned problems the statisticians looked for other methods for constructing confidence interval. Since

$$\mathsf{P}\left(\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right| \leq u_{\alpha/2}\right) \approx 1-\alpha,$$

we have

$$\mathsf{P}\big(|\hat{p}-p|^2 \leq u_{\alpha/2}^2 p(1-p)/n\big) \approx 1-\alpha.$$

The roots of the equation

$$|\hat{p} - p|^2 = u_{\alpha/2}^2 p(1-p)/n$$

determine new confidence interval. Define shortly $u = u_{\alpha/2}$. Then we get the equation

$$(n + u^2)p^2 - (2n\hat{p} + u^2)p + n\hat{p}^2 = 0$$

with roots

$$\begin{aligned} p_{12} &= \frac{2n\hat{p} + u^2 \pm \sqrt{(2n\hat{p} + u^2)^2 - 4(n + u^2)n\hat{p}^2}}{2(n + u^2)} \\ &= \hat{p}\frac{n}{n + u^2} + \frac{1}{2}\frac{u^2}{n + u^2} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\frac{n^2u^2}{(n + u^2)^2} + \frac{1}{4}\frac{u^4}{(n + u^2)^2}}. \end{aligned}$$

We derived *Wilson confidence interval* (also called *score interval* or also *q-interval*). Denote the center of the Wilson interval by symbol c. Since c is the weighted mean of the values $\hat{p}$ and $\frac{1}{2}$, we get $|\frac{1}{2} - c| < |\frac{1}{2} - \hat{p}|$.

10.2 Confidence intervals for parameter λ

10.2.1 Standard confidence interval

Let $X_1, \ldots, X_n$ be a sample from $\mathsf{Po}(\lambda)$. Maximum likelihood estimator of the parameter λ is $\hat{\lambda} = \bar{X}$. It is known that

$$\mathsf{E}\bar{X} = \lambda, \quad \mathsf{var}\,\bar{X} = \lambda/n, \quad Z = \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1).$$

Since $\bar{X} \xrightarrow{\mathsf{P}} \lambda$, we obtain

$$\frac{\bar{X} - \lambda}{\sqrt{\bar{X}/n}} \xrightarrow{\mathsf{d}} \mathsf{N}(0, 1).$$

From here it follows that the interval with endpoints

$$\bar{X} \pm u_{\alpha/2}\sqrt{\bar{X}/n} \tag{10.3}$$

is confidence interval for λ with asymptotic confidence coefficient $1 - \alpha$. It is called *standard confidence interval*.

10.2.2 Score confidence interval

If the true value of the parameter is λ, then it follows from the asymptotic normality of the variable Z that

$$\mathsf{P}\left(\frac{|\bar{X} - \lambda|}{\sqrt{\lambda/n}} \le u_{\alpha/2}\right) \to 1 - \alpha.$$

Write $u = u_{\alpha/2}$. Then the endpoints of the *score confidence interval* are roots of the equation

$$\frac{|\bar{X} - \lambda|}{\sqrt{\lambda/n}} = u.$$

We write this equation in the form $\bar{X}^2 - 2\bar{X}\lambda + \lambda^2 = u\lambda/n$ and we get the solution

$$\lambda_{12} = \frac{1}{2}\left[2\bar{X} + \frac{u^2}{n} \pm \sqrt{\left(2\bar{X} + \frac{u^2}{n}\right)^2 - 4\bar{X}^2}\right] = \bar{X} + \frac{u^2}{2n} \pm \frac{u}{\sqrt{n}}\sqrt{\bar{X} + \frac{u^2}{4n}}. \qquad (10.4)$$

10.2.3 Clopper-Pearson confidence interval

The sums of probabilities in the Poisson distribution can be calculated using the distribution function of the χ^2 distribution. Remember that the density of χ^2_n distribution is

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1}\mathrm{e}^{-x/2}, \qquad x > 0. \qquad (10.5)$$

Integrating by parts it can be shown that

$$\sum_{k=0}^{m} \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \int_{2\lambda}^{\infty} f_{2m+2}(x)\,\mathrm{d}x.$$

Assume that $X \sim \mathsf{Po}(\lambda)$. Then the *Clopper-Pearson confidence interval* for λ based on X is (λ_d, λ_u), where the limits λ_d and λ_u are given by the conditions

$$\mathsf{P}(X \geq x|\lambda = \lambda_d) = \alpha/2, \qquad \mathsf{P}(X \leq x|\lambda = \lambda_u) = \alpha/2,$$

and $\lambda_d = 0$ for $x = 0$. Let $\chi^2_\nu(\alpha)$ be upper α percentile of χ^2_ν distribution (critical value of the χ^2_ν distribution on the level α). Then the solution is

$$\lambda_d = \begin{cases} 0, & \text{for } x = 0, \\ \frac{1}{2n}\chi^2_{2x}\left(1 - \frac{\alpha}{2}\right), & \text{for } x \neq 0, \end{cases} \qquad \lambda_u = \frac{1}{2n}\chi^2_{2x+2}\left(\frac{\alpha}{2}\right).$$

In the case that $X_1, \ldots, X_n$ is the sample from $\mathsf{Po}(\lambda)$, we define $X = \sum_{i=1}^n X_i$. Thus we have $X \sim \mathsf{Po}(n\lambda)$. The limit λ_u on the basis X is given by the condition

$$\frac{\alpha}{2} = \mathsf{P}(X \leq x|\lambda = \lambda_u) = \mathsf{P}(\chi^2_{2x+2} \geq 2n\lambda_u).$$

This gives $2n\lambda_u = \chi^2_{2x+2}(\alpha/2)$, it means $\lambda_u = \frac{1}{2n}\chi^2_{2x+2}(\alpha/2)$.

10.3 Tests χ^2 when parameters are known

Let random vector $\boldsymbol{X}$ have the multinomial distribution. In view of (2.26) on p. 31 we can apply the central limit theorem. Define $\boldsymbol{Y} = \boldsymbol{D}^{-1}\boldsymbol{X}$. The vector $\boldsymbol{Y}$ has also asymptotically normal distribution because it arises by linear transformation from $\boldsymbol{X}$. Using results derived in section 2.9 on p. 30 we have $\mathsf{E}\boldsymbol{Y} = (\sqrt{np_1}, \ldots, \sqrt{np_k})'$, $\mathsf{var}\,\boldsymbol{Y} = \boldsymbol{Q}$. Applying theorem 3.8 on p. 39, we can see that $(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})'(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})$ has asymptotically χ^2_{k-1} distribution, because the rank of the the matrix $\boldsymbol{Q}$ is $k-1$. Variable $(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})'(\boldsymbol{Y} - \mathsf{E}\boldsymbol{Y})$ is denoted as χ^2 and it equals to

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}. \qquad (10.6)$$

Variables X_i are called *empirical frequencies* and np_i are *theoretical frequencies*. In some cases the value χ^2 is calculated from (10.6) because it is important to know summands on the right hand side. The complete value χ^2 can be calculated from the modified formula (10.6)

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{k}\frac{X_i^2}{p_i} - n. \tag{10.7}$$

Using variable χ^2 defined in (10.6) and introduced in (10.7) we can test the hypothesis that the true values of probabilities of a multinomial distribution are equal to numbers $p_1, \ldots, p_k$. If we get $\chi^2 \geq \chi^2_{k-1}(\alpha)$, we reject the hypothesis H_0. This *Pearson* χ^2 *test* can be applied for testing regularity of dice, (where the probability should be 1/6 for each possibility), for checking generators of random numbers and in many other cases.

It is necessary to emphasize that the test χ^2 is asymptotic and so it can be recommended only if the sample size n is sufficiently large. Modern computers enable to find the critical value using simulations.

Example 10.2 In the book Yule, Kendall (1950) are introduced results of 4096 throws with 12 dices. For each throw it was recorded how many six's were obtained. The results are introduced in table 10.1.

Table 10.1: Results of 4096 throws with 12 dices

	Number of 6								
	0	1	2	3	4	5	6	7 and more	Total
n_i	447	1145	1181	796	380	115	24	8	4096
p_i	0.112	0.269	0.296	0.197	0.089	0.028	0.007	0.001	1.000
np_i	459	1103	1213	809	364	116	27	5	4096

Probability of a 6 is 1/6 and results are independent, number of 6 is random variable with binomial distribution $\mathsf{Bi}\left(12, \frac{1}{6}\right)$. Thus

$$p_i = \binom{12}{i}\left(\frac{1}{6}\right)^i\left(\frac{5}{6}\right)^{12-i}, \qquad i = 0, 1, \ldots, 12.$$

Using (10.6) or (10.7) we get $\chi^2 = 5.484$. Since this result is less than critical value $\chi^2_7(0.05) = 14.07$, we cannot reject the hypothesis that the dices are regular. ♢

10.4 Tests χ^2 when parameters are not known

It happens frequently that the probabilities $p_1, \ldots, p_k$ depend on an unknown parameter $\boldsymbol{a} = (a_1, \ldots, a_m)'$. We can write $p_1 = p_1(\boldsymbol{a}), \ldots, p_k = p_k(\boldsymbol{a})$. Pro each $\boldsymbol{a}$ we must have

$$p_1(\boldsymbol{a}) + \cdots + p_k(\boldsymbol{a}) = 1.$$

If the functions $p_i(\boldsymbol{a})$ are sufficiently smooth then we obtain from here

$$\frac{\partial p_1(\boldsymbol{a})}{\partial a_j} + \cdots + \frac{\partial p_k(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, \ldots, m. \tag{10.8}$$

This relation is used for derivation of other formulas. Instead of (10.7) we have now

$$\chi^2(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{k}\frac{X_i^2}{p_i(\boldsymbol{a})} - n. \tag{10.9}$$

For estimating $\boldsymbol{a}$ similar procedure can be used as it was least squares method in the case of linear model. Let $\boldsymbol{a}^*$ be the value of the parameter $\boldsymbol{a}$, which minimizes (10.9). Then $\boldsymbol{a}^*$ is called *estimator of the parameter* $\boldsymbol{a}$ *using method of minimal* χ^2. We get it solving the system of equation

$$\frac{\partial\chi^2(\boldsymbol{a})}{\partial a_j} = -\frac{1}{n}\sum_{i=1}^{k}\frac{X_i^2}{p_i^2(\boldsymbol{a})}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,\ldots,m. \tag{10.10}$$

Usually, it is difficult to solve this system and so another procedure was looked for. Instead of (10.9) we can differentiate relation

$$\chi^2(\boldsymbol{a}) = \sum_{i=1}^{k}\frac{[X_i - np_i(\boldsymbol{a})]^2}{np_i(\boldsymbol{a})}. \tag{10.11}$$

This leads to

$$-\frac{1}{2}\frac{\partial\chi^2(\boldsymbol{a})}{\partial a_j} = \sum_{i=1}^{k}\left(\frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{[X_i - np_i(\boldsymbol{a})]^2}{2np_i^2(\boldsymbol{a})}\right)\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0 \tag{10.12}$$

for $j = 1,\ldots,m$. It is the same as the system (10.10). It can be shown that with growing n the influence of the second member on the right hand side of formula (10.12) has less influence. If we leave it out, we come to a new system of equations

$$\sum_{i=1}^{k}\frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,\ldots,m.$$

Using the relation (10.8), we have finally the system

$$\sum_{i=1}^{k}\frac{X_i}{p_i(\boldsymbol{a})}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,\ldots,m. \tag{10.13}$$

The solution of the system (10.13) denote $\hat{\boldsymbol{a}}$. It is called *estimator of parameter* $\boldsymbol{a}$ by *modified method of minimum* χ^2.

System (10.13) is related to problem of estimating parameter $\boldsymbol{a}$ by maximum likelihood method. Likelihood function of multinomial distribution is

$$f(\boldsymbol{a}) = \frac{n!}{X_1!\ldots X_k!}\,[p_1(\boldsymbol{a})]^{X_1}\ldots[p_k(\boldsymbol{a})]^{X_k}.$$

Thus the logarithmic likelihood function is

$$L(\boldsymbol{a}) = \ln f(\boldsymbol{a}) = \ln\frac{n!}{X_1!\ldots X_k!} + \sum_{i=1}^{k} X_i \ln p_i(\boldsymbol{a}).$$

Likelihood equations are

$$\frac{\partial L(\boldsymbol{a})}{\partial a_j} = 0 \qquad \text{for } j = 1, \ldots, m.$$

Estimator of the parameter $\boldsymbol{a}$ using maximum likelihood method is solution of the system of equations

$$\sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0 \qquad \text{for } j = 1, \ldots, m, \tag{10.14}$$

which is the same as the system of equations (10.13) for estimator $\boldsymbol{a}$ by modified method of minimum χ^2.

Theorem 10.3 *Let $m < k - 1$ and assume that for all points $\boldsymbol{a}$ from nondegenerated finite closed interval A z $\mathbb{R}_m$ hold:*

(1) $p_1(\boldsymbol{a}) + \cdots + p_k(\boldsymbol{a}) = 1$.

(2) There exists such $c > 0$, that $p_i(\boldsymbol{a}) > c^2$ for $i = 1, \ldots, k$.

(3) Every function $p_i(\boldsymbol{a})$ has continuous derivatives

$$\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} \qquad \text{and} \qquad \frac{\partial^2 p_i(\boldsymbol{a})}{\partial a_j \partial a_s} \qquad \text{for} \quad j,\, s = 1, \ldots, m.$$

(4) Matrix $\left(\dfrac{\partial p_i(\boldsymbol{a})}{\partial a_j} \right)$, *which is of type $k \times m$, has rank m.*

Let $\boldsymbol{a}^0$ be inner point of A. Denote $p_i^0 = p_i(\boldsymbol{a}^0)$. Let $\boldsymbol{X} = (X_1, \ldots, X_k)'$ have multinomial distribution with parameters $n, p_1^0, \ldots, p_k^0$. Then in the case $n \to \infty$ there exist such sequences of positive numbers $\varepsilon_n \to 0$ and $\delta_n \to 0$, that the system (10.13) has with probability at least $1 - \varepsilon_n$ one root $\hat{\boldsymbol{a}}_n$ such that $|\hat{\boldsymbol{a}}_n - \boldsymbol{a}^0| < \delta_n$. If we insert this root into (10.14) [or equivalently to (10.13)], the variable $\chi^2(\hat{\boldsymbol{a}}_n)$ for $n \to \infty$ has asymptotically χ^2_{k-m-1} distribution.

Proof. See Anděl (1978) or Cramér (1946). □

Let us remark that assumption (4) excludes to introduce superfluous parameters a_i into model. Analogous theorem can be proved also for other kind of estimators, see Rao 1978. But there exist nice estimators of the parameter $\boldsymbol{a}$ which do not satisfy assumptions of theorem 10.3. Typically χ^2 test is used as test for distribution and as the test of independence in contingency tables. We shall introduce these methods. Also here it is necessary to have theoretical frequencies $np_i(\hat{\boldsymbol{a}})$ sufficiently large.

10.5 Test of independence

Let the random vector $\boldsymbol{X} = (Y, Z)'$ have discrete distribution, variable Y takes values $1, \ldots, r$ and variable Z values $1, \ldots, c$. We shall write

$$p_{ij} = \mathsf{P}(Y = i, Z = j), \qquad p_{i.} = \sum_j p_{ij}, \qquad p_{.j} = \sum_i p_{ij}.$$

Assume that we have sample of size n from this distribution. Let n_{ij} be number of the cases when the pair (i, j) appeared in the sample. Random variables n_{ij} then have simultaneous *multinomial distribution* with parameter n and with probabilities p_{ij}. There is a difference in comparison with section 10.4, namely that probabilities p_{ij} as well as empirical frequencies n_{ij} are written as matrices instead of vectors. Matrix (n_{ij}) is called *contingency table*. Further we write

$$n_{i.} = \sum_j n_{ij}, \qquad n_{.j} = \sum_i n_{ij}.$$

Clearly it holds

$$n = \sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij}.$$

Numbers $p_{i.}$ and $p_{.j}$ are called *marginal probabilities* and values $n_{i.}$ and $n_{.j}$ are *marginal frequencies*. Matrix of probabilities (p_{ij}) and the contingency table (n_{ij}) are introduced in tab. 10.2.

Table 10.2: Matrix of probabilities and contingency table

Matrix of probabilities

	Z			
Y	1	$\cdots$	c	Σ
1	p_{11}	$\cdots$	p_{1c}	$p_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	p_{r1}	$\cdots$	p_{rc}	$p_{r.}$
Σ	$p_{.1}$	$\cdots$	$p_{.c}$	1

Contingency table

	Z			
Y	1	$\cdots$	c	Σ
1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
Σ	$n_{.1}$	$\cdots$	$n_{.c}$	n

In practical situations the contingency table arises in the following way. We follow two characteristics. They can be discrete and have only a few values (man - woman, blood group 0 - A - B - AB) or we prepare only some categories. In many cases we give numbers $1, 2, \ldots$ only as marks, but they are not exact values.

The most frequent problem solved in contingency tables is testing hypothesis that the variables Y and Z are independent. Before derivation of the *test independence*, we prove an auxiliary assertion.

Theorem 10.4 *Variables Y and Z are independent if and only if for all pairs (i, j) the relation $p_{ij} = p_{i.}p_{.j}$ holds.*

Proof. It is known that Y and Z are independent if and only if $\mathsf{P}(Y \in A, Z \in B) = \mathsf{P}(Y \in A)\mathsf{P}(Z \in B)$ for arbitrary sets $A \subset \{1, \ldots, r\}$, $B \subset \{1, \ldots, c\}$. If we choose $A = \{i\}$, $B = \{j\}$, then it follows that in the case of independence Y and Z we must have $p_{ij} = p_{i.}p_{.j}$. Choose $A = \{1, 2\}$, $B = \{1, 2, 3\}$. The general case is similar. We have

$$\begin{aligned}\mathsf{P}(Y \in A, Z \in B) &= \sum_{i=1}^{2}\sum_{j=1}^{3} p_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{3} p_{i.}p_{.j} \\ &= \left(\sum_{i=1}^{2} p_{i.}\right)\left(\sum_{j=1}^{3} p_{.j}\right) = \mathsf{P}(Y \in A)P(Z \in B). \qquad \square\end{aligned}$$

Thus the hypothesis of independence H_0 can be written in the form

$$H_0 : p_{ij} = p_{i.}p_{.j}, \qquad i = 1, \ldots, r;\ j = 1, \ldots, c.$$

It means that probabilities p_{ij} of multinomial distribution are functions of smaller number of unknown parameters, namely of marginal probabilities $p_{i.}$ a $p_{.j}$. But the marginal probabilities are not independent, since

$$\sum_i p_{i.} = \sum_j p_{.j} = 1.$$

If we wish to fulfil condition (4) from theorem 10.3, then we cannot include probabilities $p_{r.}$ and $p_{.c}$ among unknown parameters, because they can be calculated from other probabilities. Thus we have $m = r - 1 + c - 1 = r + c - 2$ unknown parameters. However, we consider only situations where all marginal probabilities are positive. Otherwise we would skip some rows or some columns.

Unknown parameters $p_{1.}, \ldots, p_{r-1,.}$ and $p_{.1}, \ldots, p_{.,c-1}$ can be calculated from the system (10.13). Instead of X_i we have here variables n_{ij}. We obtain

$$\sum_{j=1}^{c} \left(\frac{n_{ij}}{p_{i.}} - \frac{n_{rj}}{p_{r.}} \right) = 0, \qquad i = 1, \ldots, r-1 \tag{10.15}$$

and

$$\sum_{i=1}^{r} \left(\frac{n_{ij}}{p_{.j}} - \frac{n_{ic}}{p_{.c}} \right) = 0, \qquad j = 1, \ldots, c-1, \tag{10.16}$$

since we have for $h = 1, \ldots, r-1$ that

$$\frac{\partial p_{i.}p_{.j}}{\partial p_{h.}} = \begin{cases} p_{.j} & \text{for } i = h, \\ -p_{.j} & \text{for } i = r, \\ 0 & \text{in other cases.} \end{cases}$$

Similar result can be obtained for partial derivatives with respect to $p_{.l}$. Formula (10.15) holds also for $i = r$ and (10.16) holds also for $j = c$. Instead of (10.15) we can write

$$\frac{n_{i.}}{p_{i.}} - \frac{n_{r.}}{p_{r.}} = 0, \qquad i = 1, \ldots, r. \tag{10.17}$$

From here we get

$$n_{i.} = \frac{n_{r.}}{p_{r.}} p_{i.}, \qquad i = 1, \ldots, r.$$

Summing over i we have $n = n_{r.}/p_{r.}$, so that estimator for $p_{r.}$ is $\hat{p}_{r.} = n_{r.}/n$. Finally, inserting into (10.17) we have estimators

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, \qquad i = 1, \ldots, r.$$

Similarly can be solved (10.16) and the result is

$$\hat{p}_{.j} = \frac{n_{.j}}{n}, \qquad j = 1, \ldots, c.$$

Theorem 10.3 ensures that the variable

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \tag{10.18}$$

has asymptotically distribution χ^2, and its number of degrees of freedom is

$$rc - (r + c - 2) - 1 = (r - 1)(c - 1).$$

If we get $\chi^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$, we rejct the hypothesis H_0 that the variables Y and Z are independent. Of course, the theoretical frequencies $n_{i.}n_{.j}/n$ must be sufficiently large.

Example 10.5 U 6800 mu byla zjiovna barva o a barva vlas (viz Yule, Kendall 1950). Vsledky jsou uvedeny v tab. 10.3.

Table 10.3: Colour of eyes and hair

	Color of hair				
Colour of eyes	Fair	Chestnut	Black	Red-haired	Total
Blue	1 768	807	189	47	2 811
Green	946	1 387	746	53	3 132
Brown	115	438	288	16	857
Total	2 829	2 632	1 223	116	6 800

We obtained $\chi^2 = 1073.508$, $df = 6$, p-value $< 2.2e - 16$.

Because of small p-value we reject the hypothesis that color of eyes and color of hair in men population are independent variables.

$\diamondsuit$

10.6 2×2 tables

10.6.1 Test χ^2

Test of independence in contingency table 2×2 can be based on formula (10.18). In the case $r = c = 2$ we get

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(nn_{ij} - n_{i.}n_{.j})^2}{n_{i.}n_{.j}}.$$

For every pair (i, j) we have

$$(nn_{ij} - n_{i.}n_{.j})^2 = (n_{11}n_{22} - n_{12}n_{21})^2,$$

and thus

$$\chi^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n} \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{1}{n_{i.}n_{.j}} = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n} \frac{n^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

$$= n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{10.19}$$

If hypothesis of independence is valid, then the statistic χ^2 has asymptotically χ^2_1 distribution.

10.6.2 Odds ratio

Another approach to 2×2 tables can be motivated in the following example. Assume that a man has a disease. He checks up that this disease had 18 men. Some of them were cured, some not. Some are alive, some not. Data are in tab. 10.4.

Table 10.4: Data on sick men

	Alive	Dead	Total
Cured	5	6	11
Not cured	3	4	7
Total	8	10	18

The man can think as follows. If he cures, his chance to alive can be estimated 5:6. If he does not cure, then the chance to alive is 3:4. Using division we find the better possibility: the expression

$$\frac{5:6}{3:4} = \frac{5 \times 4}{6 \times 3} = \frac{20}{18}$$

is larger than 1, and so it will be better to cure. Such expression is generally

$$b = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

and we call it *odds ratio*. Since n_{ij}/n is estimator of probability p_{ij}, b is estimator of C.

$$\beta = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Theorem 10.6 *In 2×2 table the equality $\beta = 1$ holds if and only if $p_{ij} = p_{i.}p_{.j}$ for all couples $(i, j), i \neq j$.*

Proof. If $p_{ij} = p_{i.}p_{.j}, i \neq j$ for all couples (i, j), then we get immediately that $\beta = 1$.

Now, assume that $\beta = 1$ holds. If we denote $p_{11}/p_{12} = \lambda$, then from $\beta = 1$ it follows that also $p_{21}/p_{22} = \lambda$. From here

$$p_{11} = \lambda p_{12}, \qquad p_{21} = \lambda p_{22}.$$

The corresponding table of probabilities can be written as tab. 10.5.

Table 10.5: Table of probabilities

λp_{12}	p_{12}	$(\lambda + 1)p_{12}$
λp_{22}	p_{22}	$(\lambda + 1)p_{22}$
		$(\lambda + 1)p_{.2}$

Since $(\lambda+1)p_{\cdot 2}=1$, we get that $\lambda+1=1/p_{\cdot 2}$. Because $(\lambda+1)p_{12}=p_{1\cdot}$, we see that $p_{12}=p_{1\cdot}p_{\cdot 2}$. Similarly from the relation $(\lambda+1)p_{22}=p_{2\cdot}$ it follows that $p_{22}=p_{2\cdot}p_{\cdot 2}$. Finally we get

$$\begin{aligned} p_{11}&=p_{1\cdot}-p_{12}=p_{1\cdot}-p_{1\cdot}p_{\cdot 2}=p_{1\cdot}(1-p_{\cdot 2})=p_{1\cdot}p_{\cdot 1},\\ p_{21}&=p_{2\cdot}-p_{22}=p_{2\cdot}-p_{2\cdot}p_{\cdot 2}=p_{2\cdot}(1-p_{\cdot 2})=p_{2\cdot}p_{\cdot 1}. \end{aligned}$$ □

The dependence of variables will be larger if β is far from 1. However, do not forget that $0\le\beta\le\infty$. It was proved (see Edwards 1963), that every reasonable measure of dependence in 2×2 table must be function of parameter β, in the sample the function b.

Values β are not symmetric around point 1. From this reason the characteristics logarithmic interaction d and theoretic logarithmic interaction δ were proposed, which are defined as

$$d=\ln b,\qquad \delta=\ln\beta.$$

It was proved (see Goodman 1964), that variable

$$D=\frac{d-\delta}{\sqrt{\dfrac{1}{n_{11}}+\dfrac{1}{n_{12}}+\dfrac{1}{n_{21}}+\dfrac{1}{n_{22}}}}$$

has asymptotically distribution $\mathsf{N}(0,1)$. If we want to test hypothesis independence H_0 using this method, we insert $\delta=0$ and H_0 is rejected in the case that $|D|\ge u\left(\frac{\alpha}{2}\right)$. It is important that using D it is possible to test H_0 also against one-sided alternatives. Again, the test is asymptotic and should be used when the frequencies are sufficiently large.

Let us return to the parameter β. It can be expressed in the form

$$\beta=\frac{p_{11}p_{22}}{p_{12}p_{21}}=\frac{p_{11}(1-p_{1\cdot}-p_{\cdot 1}+p_{11})}{(p_{1\cdot}-p_{11})(p_{\cdot 1}-p_{11})}.$$

Probability p_{11} is function of $\beta,p_{1\cdot},p_{\cdot 1}$. The same holds also for other probabilities p_{ij}. Conditional distribution of frequencies n_{ij} for given marginal frequencies depends only on β. Then it suffices to know only the frequency n_{11} since other frequencies n_{ij} are determined. Conditional probability $\mathsf{P}(n_{11}=t)$ depends only on $n_{1\cdot},n_{\cdot 1},n,\beta$. Fisher (1935) derived, that this probability is given by *non-central hypergeometric distribution*

$$\mathsf{P}(n_{11}=t|n_{1\cdot},n_{\cdot 1},n,\beta)=\frac{\binom{n_{1\cdot}}{t}\binom{n-n_{1\cdot}}{n_{\cdot 1}-t}\beta^t}{\sum_u\binom{n_{1\cdot}}{u}\binom{n-n_{1\cdot}}{n_{\cdot 1}-u}\beta^u}. \tag{10.20}$$

Using formula (10.20) it is possible to obtain conditional confidence interval for β (see Agresti 2002, p. 99). The value $\hat\beta$, which maximizes probability (10.20), is called *conditional maximum likelihood estimator* of the parameter β. The estimator

$$b=(n_{11}n_{22})/(n_{12}n_{21})$$

mentioned above is *unconditional maximum likelihood estimator* of the parameter β, since it is based on maximum likelihood estimators of probabilities p_{ij} of this multinomial distribution.

Example 10.7 In England it was investigated if the weight of criminals and their intellect are independent variables. Data (see Yule, Kendall 1950) are presented in tab. 10.6.

We test the hypothesis $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$. For $\delta = 0$ we get $D = -3{,}03$. Since $|D| \geq u(0{,}025) = 1{,}96$, we reject the hypothesis H_0.

Table 10.6: Data on criminals

Intellect	Weight		
	to 150 lb	over 150 lb	Total
Normal	272	124	396
Lowered	82	15	97
Total	354	139	493

We compare this result with the classical Pearson test. We get $\chi^2 = 9{,}67$, p-value $= 0.002$. Since p-value is smaller than 0.05, hypothesis of independence would be rejected also by this test. Notice that $D^2 = 9{,}18$ is not very different from the value $\chi^2 = 9{,}67$. $\diamondsuit$

The application of the odds-ratio can also sometimes give paradoxical results. Consider again the example presented at the beginning of this section. Let us imagine that a woman is ill. She would have for disposal data on 23 ill women (see tab. 10.7).

Table 10.7: Data on sick women

	Alive	Dead	Total
Cured	6	3	9
Not cured	9	5	14
Total	15	8	23

Table 10.8: Data on sick people

	Alive	Dead	Total
Cured	11	9	20
Not cured	12	9	21
Total	23	18	41

This woman calculates odds ratio $b = (6 : 3)/(9 : 5) = 30/27$. Since $b > 1$, for her it would be better to cure.

If the data are collapsed, we obtain tab. 10.8.

In the collapsed table we have $b = 99/108 < 1$. This leads to the surprising conclusion that for men and women it is better to cure but for people not to cure. This is *Simpson paradox*.

10.6.3 Fisher's factorial test

If the frequencies are small and it is not appropriate to use in the 2×2 the limiting distribution χ^2_1 , in such a case is applied the following method. Probability that given n frequencies $n_{11}, n_{12}, n_{21}, n_{22}$, will be realized is

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}\, p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}.$$

Assume that the hypothesis independence $H_0 : p_{ij} = p_{i.}p_{.j}$ is true. Denote

$$Q = p_{1.}^{n_{1.}} p_{2.}^{n_{2.}} p_{.1}^{n_{.1}} p_{.2}^{n_{.2}}.$$

Then under H_0 we get

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}\, Q. \tag{10.21}$$

Probability that a table with marginal frequencies $n_{1.}, n_{2.}, n_{.1}, n_{.2}$ arizes, is equal to

$$R = \sum_{i=\max(0, n_{.1}-n_{2.})}^{\min(n_{1.}, n_{.1})} P(i, n_{1.} - i, n_{.1} - i, i + n_{2.} - n_{.1}).$$

Inserting from (10.21) we obtain

$$R = Q \frac{n!}{n_{1.}!\, n_{2.}!} \sum_{i=\max(0, n_{.1}-n_{2.})}^{\min(n_{1.}, n_{.1})} \binom{n_{1.}}{i} \binom{n_{2.}}{n_{.1} - i}.$$

But for arbitrary nonnegative integers r, s and k fulfilling condition $r + s \geq k$ comparing coefficients by t^k in the formula

$$(1 + t)^r (1 + t)^s = (1 + t)^{r+s}$$

we get *Vandermond convolutory formula*

$$\sum_{i=\max(0, k-s)}^{\min(r,k)} \binom{r}{i} \binom{s}{k - i} = \binom{r + s}{k}.$$

This gives

$$R = Q \frac{(n!)^2}{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}.$$

Conditional probability P, that in given table with marginal frequencies $n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$ a table with frequencies n_{11}, n_{12}, n_{21}, n_{22} arises, is equal

$$P = \frac{P(n_{11}, n_{12}, n_{21}, n_{22})}{R}.$$

$$P = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}. \tag{10.22}$$

Sometimes P is introduced in an equivalent form

$$P = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n}{n_{.1}}}.$$

An advantage of the conditional probability P is that it does not contain any unknown parameters and so we have no problem in estimating them.

The testing procedure depends on the problem if we want to test hypothesis H_0 against one-sided or two-sided alternative. Let d be the logarithmic interaction of he given table. If we test H_0 against the alternative $H_1 : \delta < 0$ (where δ is the theoretical logarithmic interaction), probabilities P are summed for the tables which have the same marginal frequencies as the original table and their logarithmic interactions are smaller or equal number d. If this sum is smaller or equal α, H_0 is rejected.

When testing H_0 against $H_1 : \delta > 0$ we add probabilities P of the tables with the same marginal frequencies as the original table, the logarithmic interactions having larger or equal d. If the sum of probabilities smaller or equal α, H_0 is rejected. The two-sided test can be done analogously.

This test was derived by R. A. Fisher. Since calculation of probabilities is based on formula (10.22), the test is called Fisher's factorial test. The actual level of the test is usually less than α. Sometimes it is called Fisher's exact test.

Example 10.8 Some 24 randomly chosen students were asked if they have good or bad results in maths and if the study music. Denote $M+$ good and $M-$ bad result in math and denote $H+$ (and $H-$) the case when the student studies (or not studies) music. The results are in tab. 10.9.

Table 10.9: Mathematics and music

	$H+$	$H-$	Celkem
$M+$	6	4	10
$M-$	1	13	14
Total	7	17	24

We should test the hypothesis that the results in math and study of music are independent.

We find all tables with the same marginal frequencies as the starting table (see tab. 10.10). Then we calculate their logarithmic interactions d and Fisher's probabilities P.

You can see that the sum of all probabilities P is really 1. We do two-sided test and so we sum the probabilities P, which correspond to tables with absolute value of logarithmic interaction larger or equal to the number 2.97. We get the number

$$0.009\,916 + 0.008\,495 + 0.000\,347 = 0.018\,758.$$

The sum $0.018\,758$ is not larger than $\alpha = 0.05$, and so we reject the hypothesis about independence. We emphasize that this does not prove the causality.

$\diamondsuit$

Table 10.10: Set of contingency tables with the same marginal frequencies as tab. 10.9

$\begin{matrix} 0 & 10 \\ 7 & 7 \end{matrix}$	$\begin{matrix} 1 & 9 \\ 6 & 8 \end{matrix}$	$\begin{matrix} 2 & 8 \\ 5 & 9 \end{matrix}$	$\begin{matrix} 3 & 7 \\ 4 & 10 \end{matrix}$
$d = -\infty$	$d = -1.91$	$d = -0.80$	$d = 0.07$
$P = 0.009\,916$	$P = 0.086\,766$	$P = 0.260\,297$	$P = 0.347\,063$

$\begin{matrix} 4 & 6 \\ 3 & 11 \end{matrix}$	$\begin{matrix} 5 & 5 \\ 2 & 12 \end{matrix}$	$\begin{matrix} 6 & 4 \\ 1 & 13 \end{matrix}$	$\begin{matrix} 7 & 3 \\ 0 & 14 \end{matrix}$
$d = 0.89$	$d = 1.79$	$d = 2.97$	$d = \infty$
$P = 0.220\,858$	$P = 0.066\,258$	$P = 0.008\,495$	$P = 0.000\,347$

Chapter 11

Paired problem

11.1 Paired t test

Consider the random sample $(Y_1, Z_1)', \ldots, (Y_n, Z_n)'$ from a twodimensional distribution with vector of mean values $(\mu_1, \mu_2)'$. We want to test the hypothesis $H_0 : \mu_1 - \mu_2 = \Delta$ against alternative $H_1 : \mu_1 - \mu_2 \neq \Delta$, where Δ is given number (mostly zero). Define $X_i = Y_i - Z_i$, $i = 1, \ldots, n$. Variables $X_1, \ldots, X_n$ are independent and identically distributed. Assume that $X_i \sim \mathsf{N}(\mu, \sigma^2)$. Obviously $\mu = \mu_1 - \mu_2$. Now, we have to test $H_0' : \mu = \Delta$ against $H_1' : \mu \neq \Delta$ and the problem is changed to the one sample t test, which is in this case called *paired t test*.

The hypothesis H_0', and also the hypothesis H_0, will be rejected at level α,if $|\bar{X} - \Delta|\sqrt{n}/S \geq t_{n-1}(\alpha)$, where S is calculated from z $X_1, \ldots, X_n$. The most frequent case is the situation when $\{(Y_i, Z_i)'\}$ is the sample from a two-dimensional normal normal distribution. This assumption ensures that the paired t test can be used. The paired t test is used usually in situations when we have on each from n objects measured two variables and individual objects are considered as independent but not the measurements on the same object.

Confidence interval for Δ can be derived from the fact that

$$\mathsf{P}\{|\bar{X} - \Delta|\sqrt{n}/S \leq t_{n-1}(\alpha)\} = 1 - \alpha.$$

Thus confidence interval for Δ has endpoints

$$\bar{X} \mp \frac{1}{\sqrt{n}} S t_{n-1}(\alpha).$$

Example 11.1 It should be decided if two front tyres go down equally fast. Six new cars were chosen and after some time it were measured how much two front tyres went down.

Differences in go down can be considered as independent random variables with distribution $\mathsf{N}(\mu, \sigma^2)$. If both tyres go down equally fast, the hypothesis $H_0 : \mu = 0$ is valid. We have $n = 6$, $\Delta = 0$, $\bar{X} = 0{,}0833$, $S^2 = 0{,}0377$, $T = 1{,}0518$. Since $1{,}0518 < t_5(0{,}05) = 2{,}571$, we cannot reject the hypothesis that both front tyre go down equally fast. $\diamondsuit$

Table 11.1: Tyres go down

Number of car	1	2	3	4	5	6
right tyre :	1.8	1.0	2.2	0.9	1.5	1.6
left tyre :	1.5	1.1	2.0	1.1	1.4	1.4
Difference	0.3	−0.1	0.2	−0.2	0.1	0.2

11.2 Paired sign test

Let $(Y_1, Z_1)'$,...,$(Y_n, Z_n)'$ be a random sample from a two-dimensional distribution. Define $X_i = Y_i - Z_i$, $i = 1, \ldots, n$. Assume that the variables X_i have a continuous distribution with a unique median $\tilde{x}$. We should test the hypothesis H_0, that $\tilde{x} = x_0$, where x_0 is a given number. Thus the problem is transformed to one-dimensional sigh test. The procedure is called *paired sign test*.

11.3 Paired Wilcoxon test

Consider a random sample $(Y_1, Z_1)'$,...,$(Y_n, Z_n)'$ from a two-dimensional distribution. Denote $X_i = Y_i - Z_i$, $i = 1, \ldots, n$. Assume that the variables X_i have a continuous distribution with a density which is symmetric around the point a and positive in its neighbourhood. Then a is a median of this distribution. We want to test the hypothesis H_0, that $a = x_0$, where x_0 is a given number. The problem is transformed to one-sample Wilcoxon test. This procedure is called one-sample Wilcoxon test.

11.4 Spearman correlation coefficient

We have seen that the analysis of sample correlation coefficient has the assumption that the sample is from the normal distribution. However, it happens quite often that this assumption is not valid. And sometimes in the random sample $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ the values of the mentioned random variables it is not possible to determine, only their order is in our disposal. But if the order X-values and Y-values are very similar it indicates some dependence between X_i a Y_i.

Assume that $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ is the sample from a continuous two-dimensional distribution. Let $R_1, \ldots, R_n$ be orders of variables $X_1, \ldots, X_n$ and $Q_1, \ldots, Q_n$ orders of variables $Y_1, \ldots, Y_n$. It is quite usual to order couples $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ in advance such that $R_i = i$, $i = 1, \ldots, n$.

Spearman correlation coefficient r_S is defined as sample correlation coefficient calculated from $(R_1, Q_1)', \ldots, (R_n, Q_n)'$, it means

$$r_S = \frac{\sum R_i Q_i - n\bar{R}\bar{Q}}{\sqrt{(\sum R_i^2 - n\bar{R}^2)(\sum Q_i^2 - n\bar{Q}^2)}}.$$

Theorem 11.2 *We have*

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2. \tag{11.1}$$

Proof. We have

$$r_S = \frac{\sum R_i Q_i - n\bar{R}\bar{Q}}{\sqrt{\left(\sum R_i^2 - n\bar{R}^2\right)\left(\sum Q_i^2 - n\bar{Q}^2\right)}}, \tag{11.2}$$

where

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}, \qquad \bar{Q} = \bar{R},$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} Q_i^2 = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum R_i Q_i = \frac{1}{2}\left(\sum R_i^2 + \sum Q_i^2\right) - \frac{1}{2}\sum (R_i - Q_i)^2$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum (R_i - Q_i)^2.$$

These results we insert into (11.1) and after some simplification we get formula (11.2). □

Critical values are denoted by $r_S(\alpha)$. If $|r_S| \geq r_S(\alpha)$, we reject hypothesis of independence Y_i and X_i. For $n > 30$ we use asymptotic normality of the coefficient r_S. We calculate

$$r_S^*(\alpha) = \frac{u\left(\frac{\alpha}{2}\right)}{\sqrt{n-1}}. \tag{11.3}$$

Hypothesis of independence is rejected in the case that $|r_S| \geq r_S^*(\alpha)$. One-sided tests can be derived similarly.

If $(X_1, Y_1)', \ldots, (X_n, Y_n)'$ is the sample from the regular two-dimensional normal distribution with correlation coefficient ρ, then it can be proved (see van der Waerden 1957), that

$$\mathsf{E} r_S = \frac{6}{\pi}\,\frac{n-2}{n+1}\,\arcsin\frac{\rho}{2} + \frac{6}{\pi(n+1)}\,\arcsin\rho.$$

With growing n the second member on the right-hand side tends to zero and the sample correlation coefficient r tends almost surely to ρ. We get an approximation

$$r \doteq 2\sin\left(\frac{\pi}{6}\,r_S\right).$$

If in our data from which r_S is calculated, there are many agreements, it is recommended to use *corrected Spearman correlation coeficient* (see Kendall 1962 and Sachs 1974). It is defined by the formula

$$r_{S,\text{korig}} = 1 - \frac{6\sum (R_i - Q_i)^2}{n^3 - n - T_{x'} - T_{y'}},$$

where

$$T_{x'} = \frac{1}{2}\sum (t_{x'}^3 - t_{x'}), \qquad T_{y'} = \frac{1}{2}\sum (t_{y'}^3 - t_{y'}).$$

Here symbol $t_{x'}$ denotes number of equally large X-values. (If we have among X- values a few groups with the same number of observations, then $t_{x'}$ are sizes of the groups. Symbol $t_{x'}$ is defined analogously.)

Chapter 12

Discrete paired problem

12.1 McNemar test

In some cases we want to apply another test than the test of independence or the test of homogeneity. Sometimes we have a set of n randomly chosen statistical units and presence and absence of a characteristic is investigated. Then an intervention in every unit is executed and repeatedly it is investigated if the characteristic is present or not. The aim of the investigation is to find if the intervention changed probability of the occurrence of the characteristic.

Denote 1 presence and 0 absence of the characteristic. In this case it is necessary to have data in the form of tab. 12.1. We assume that these data are sample from a multinomial distribution with parameter n and with probabilities introduced in tab. 12.2.

Table 12.1: Frequencies for the McNemar test

Before intervention	After intervention 1	0	Total
1	n_{11}	n_{12}	$n_{1.}$
0	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Table 12.2: Probabilities for McNemar test

Before intervention	After intervention 1	0	Total
1	p_{11}	p_{12}	$p_{1.}$
0	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

We want to test $H_0 : p_{1.} = p_{.1}$. By chance H_0 is equivalent to *hypothesis of symmetry* $H_0' : p_{12} = p_{21}$. In this case all probabilities in tab. 12.2 are determined by two unknown parameters p_{11} and p_{12}. If H_0' holds then $p_{21} = p_{12}$ and $p_{22} = 1 - p_{11} - 2p_{12}$.

From formula (10.13) we get

$$\frac{n_{11}}{p_{11}} - \frac{n_{22}}{p_{22}} = 0,$$
$$\frac{n_{12}}{p_{12}} + \frac{n_{21}}{p_{21}} - 2\frac{n_{22}}{p_{22}} = 0.$$

It implies

$$n_{11} = \frac{n_{22}}{p_{22}}\, p_{11}, \tag{12.1}$$

$$n_{12} + n_{21} = 2\frac{n_{22}}{p_{22}}\, p_{12}. \tag{12.2}$$

We include a trivial equality

$$n_{22} = \frac{n_{22}}{p_{22}} p_{22}$$

and then we add all relations. This gives

$$n = \frac{n_{22}}{p_{22}},$$

so that an estimator for p_{22} is $\hat{p}_{22} = n_{22}/n$. Inserting in (12.1) and (12.2) we obtain

$$\hat{p}_{11} = \frac{n_{11}}{n}, \qquad \hat{p}_{12} = \frac{n_{12} + n_{21}}{2n}, \qquad \hat{p}_{21} = \hat{p}_{12}.$$

The variable

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{ii} - n\hat{p}_{ii})^2}{n\hat{p}_{ii}} + \frac{(n_{12} - n\hat{p}_{12})^2}{n\hat{p}_{12}} + \frac{(n_{21} - n\hat{p}_{21})^2}{n\hat{p}_{21}}$$

is equal to

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \tag{12.3}$$

and has asymptotically χ^2_1 distribution. Hypothesis H_0 is rejected, when $\chi^2 \geq \chi^2_1(\alpha)$. In textbooks it is introduced that this asymptotic result is applicable for $n_{12} + n_{21} \geq 8$. The test was published in McNemar (1947).

When frequencies are small then we use the result that under H_0 the conditional distribution of frequencies n_{12} is binomial $\mathsf{Bi}\left(n_{12} + n_{21}, \frac{1}{2}\right)$. A detailed derivation can be found in the book Anděl (2007), for example. Define $N = n_{12} + n_{21}$. Hypothesis H_0 is rejected, if $n_{12} \leq k_1$ or $n_{12} \geq k_2$, where k_1 and k_2 are corresponding critical values. This test is two-sided. Using its variant based on the binomial distribution it is easy to apply the McNemar test also against its one-sided alternatives.

Example 12.1 It was investigated if application of a medicament has as a side effect change of speed of shrinkage of blood. Randomly was determined 100 patients. Each of them was determined if his shrinkage of blood is slow or fast. Then the patients obtained the mentioned medicament and after adequate time the speed of shrinkage of blood was determined again. The results are in tab. 12.3.

Table 12.3: Shrinkage of blood

Before application	After application Slow	Fast	Total
Slow	24	28	52
Fast	12	36	48
Total	36	64	100

Using (12.3) we get $\chi^2 = 6{,}4$. This value is larger than $\chi^2_1(0{,}05) = 3{,}84$. Thus we reject the hypothesis that the application of medicament has no influence on speed of shrinkage of blood.

For $N = 28 + 12 = 40$ and $\alpha = 0.05$ critical values are $k_1 = 13$, $k_2 = 27$. Since $n_{12} = 28 \geq k_2$, we reject the hypothesis that the application of medicament has no influence on speed of shrinkage of blood, too.

The calculation using the program `R` can be following one.

```
slow.aft <- c(24,12)
quick.aft <- c(28,36)
tbl <- cbind(pomala.aft,rychla.aft)
rownames(tbl) <- c("pomala.pred","rychla.pred")
names(dimnames(tbl)) <- c("reakce.pred","reakce.po")
mcnemar.test(tbl, correct=F)
        McNemar's Chi-squared test
data:  tbl McNemar's chi-squared = 6.4, df = 1, p-value = 0.01141
```

♢

12.2 Stuart test

Consider a square contingency table with dimension $c \times c$. The data should be used for testing *hypothesis of homogeneity of marginal probabilities*

$$H_0 : p_{1.} = p_{.1}, \quad \ldots \quad p_{c.} = p_{.c}.$$

We describe the *Stuart test*, which is direct generalization of the McNemar test. Introduce the following notation:

$$d_i = n_{i.} - n_{.i} \quad \text{for } i = 1, \ldots, c, \qquad \boldsymbol{d} = (d_1, \ldots, d_{c-1})'.$$

Let $\boldsymbol{V} = (V_{ij})_{i,j=1,\ldots,c-1}$ be the matrix of dimension $(c-1) \times (c-1)$, the elements of which are

$$\begin{aligned} V_{ii} &= n_{i.} + n_{.i} - 2n_{ii}, \\ V_{ij} &= -(n_{ij} + n_{ji}) \qquad \text{for} \quad i \neq j. \end{aligned}$$

Theorem 12.2 (Stuart) *If the hypothesis H_0 is valid, then the variable $Q = \boldsymbol{d}'\boldsymbol{V}^{-1}\boldsymbol{d}$ has asymptotically χ^2_{c-1} distribution.*

Proof. See Stuart (1955) or Anděl (2007). □

It is easy to check that for $c = 2$ the variable Q is the same as the variable χ^2 which is introduced in formula (12.3). If $c = 3$ the variable Q can be calculated using the formula

$$Q = \frac{N_{23}d_1^2 + N_{13}d_2^2 + N_{12}d_3^2}{2(N_{12}N_{23} + N_{12}N_{13} + N_{13}N_{23})}, \tag{12.4}$$

where

$$N_{ij} = \frac{n_{ij} + n_{ji}}{2}.$$

A similar but more complicated explicit formula was derived also for $c = 4$ (see Krauth 1985).

Chapter 13

Two-sample problem

13.1 Descriptive statistics

The description of data and their graphical representation will be illustrated on data called `energy` available in library `ISwR`. The data are in form `data.frame` with 22 rows and 2 columns. It is a description of energy expenditure in the groups of lean and obese women. The first column called `expend` is numerical vector giving expenditure of energy in MJ during 24 hours. The second column is called `stature` and it is factor with levels `lean` a `obese`. (See Dalgaard 2002.)

```
library(ISwR)
data(energy)
attach(energy)
expend.lean <- expend[stature=="lean"]
expend.obese <- expend[stature=="obese"]
expend.lean
 [1]  7.53  7.48  8.08  8.09 10.15  8.40 10.88  6.13  7.90  7.05  7.48  7.58
[13]  8.11
expend.obese
[1]  9.21 11.51 12.79 11.85  9.97 8.79  9.69  9.68  9.19
```

For each group we prepare *histogram*. The histograms are placed in such a way that they can be compared (see fig. 13.1).

```
opar <- par(mfrow=c(2,1))
hist(expend.lean, breaks=10, xlim=c(5,13), ylim=c(0,4), col="white")
hist(expend.obese, breaks=10, xlim=c(5,13), ylim=c(0,4), col="grey90")
par(opar)
```

Data can be also represented by *boxplots* (see fig. 13.2) and using figure called *stripchart* (see fig. 13.3). Figures were created using program

```
boxplot(expend ~ stature, boxwex=0.2, las=1)
stripchart(expend ~ stature, method="jitter", jitter=0.03, vertical=T,
     las=1)
```

In the case of stripcharts the method called `jitter` was used to show also data which are identical or very nearly the same.

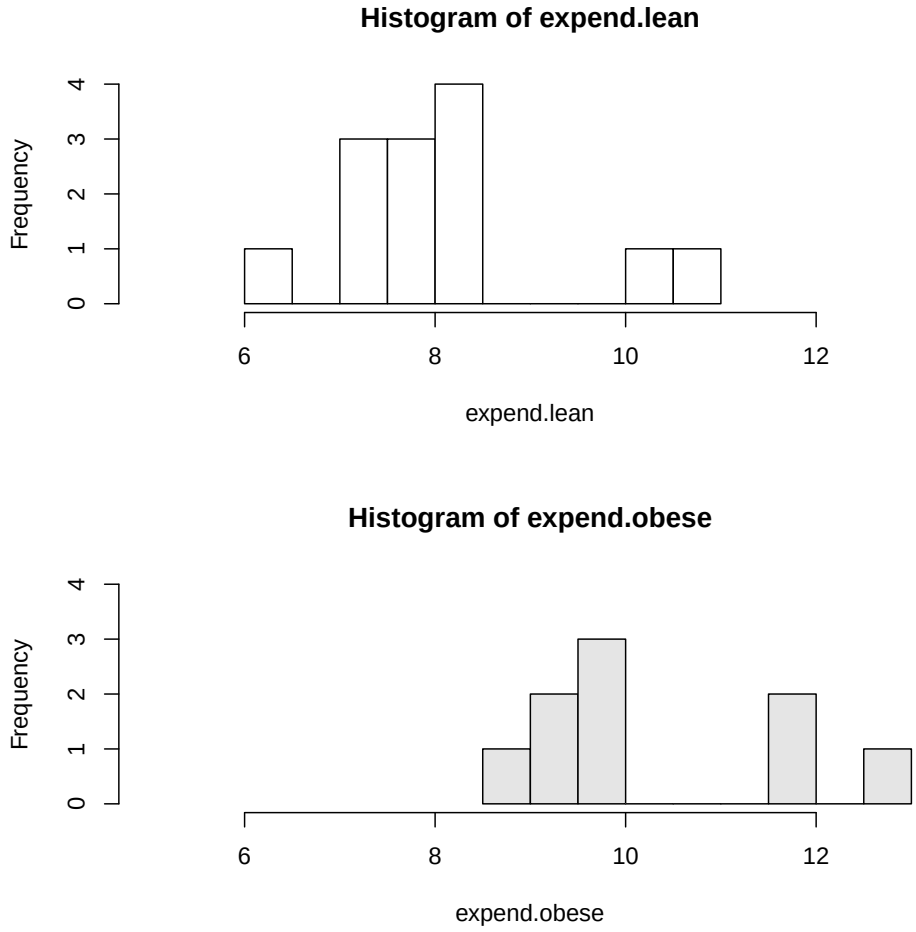

Figure 13.1: Histograms

13.2 Two-sample Kolmogorov-Smirnov test

Let $X_1, \ldots, X_m$ be a random sample from a distribution with continuous distribution function F and let $Y_1, \ldots, Y_n$ be an independent random sample with a continuous distribution function G. Consider a test of hypothesis $H_0 : F = G$ against $H_1 : F \neq G$. Let F_m be the empirical distribution function of the first sample and G_n of the second sample. Theorems 9.1 and 9.2 imply, that the functions F_m and G_n with increasing m and n converge to distribution functions F and G. Define

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|.$$

If H_0 is true then the Glivenko-Cantelli theorem gives that $D_{m,n} \to 0$ almost surely when $m \to \infty$, $n \to \infty$. An exact result is described in the following theorem.

Theorem 13.1 (Smirnov) *Let $M = mn/(m+n)$. Define*

$$K(\lambda) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2\lambda^2). \tag{13.1}$$

Then for every λ we have

$$\lim_{m,n\to\infty} \mathsf{P}(\sqrt{M}\, D_{m,n} < \lambda) = K(\lambda).$$

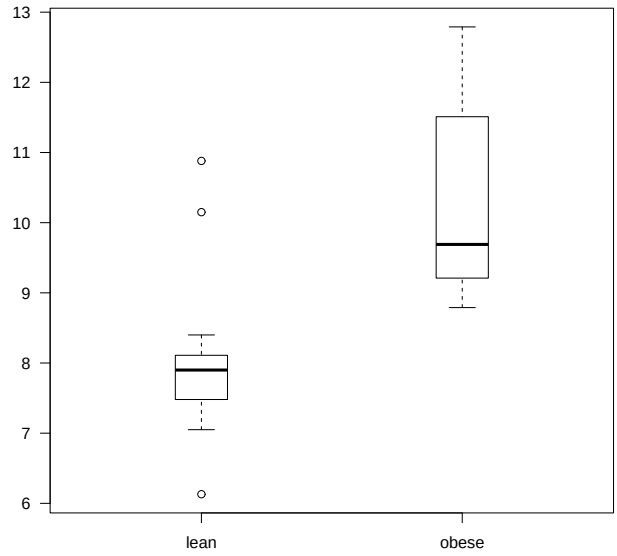

Figure 13.2: Boxplots

Proof. See Smirnov (1944). Modern proof can be found in the book Hájek, Šidák (1967). Function $K(\lambda)$ was introduced in the formula (9.3) on p. 72. □

Distribution of the variable $D_{m,n}$ for finite values m, n is introduced in the book Hájek, Šidák (1967). Function $K(\lambda)$ can be approximated by some members from the beginning of the series $1 - 2e^{-2\lambda^2}$ (see Likeš, Laga 1978). Then

$$\mathsf{P}\left(D_{m,n} < \frac{\lambda}{\sqrt{M}}\right) \doteq 1 - 2e^{-2\lambda^2}.$$

The expression on the right hand side equals to $1-\alpha$ for $\lambda = \lambda_\alpha = \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}$. The critical value is approximately

$$D^*_{m,n}(\alpha) = \frac{\lambda_\alpha}{\sqrt{M}} = \sqrt{\frac{1}{2M}\ln\frac{2}{\alpha}}.$$

Practical use of the Kolmogorov-Smirnov test is the following one. From samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ empirical distribution functions F_m and G_n and the variable $D_{m,n}$ are calculated. If the numbers m and n are small, the variable $D_{m,n}$ is compared with exact critical values $D_{m,n}(\alpha)$. If the numbers m and n are not very small, theorem 13.1 can be applied. Define $\lambda_0 = \sqrt{M}\, D_{m,n}$ and calculate $K(\lambda_0)$. If we get $K(\lambda_0) \geq 1 - \alpha$, we reject H_0 on the level which tends to α when the sizes of the samples go to infinity. The critical value for the variable $D_{m,n}$ is approximated by $D^*_{m,n}(\alpha)$. Hypothesis H_0 is rejected when $D_{m,n} \geq D^*_{m,n}(\alpha)$.

The Kolmogorov-Smirnov test was generalized to the case of comparing three and more samples in the paper Kiefer (1959). Critical values were tabulated in Wolf, Naus (1973). See also Domaǹski (1990).

Example 13.2 Two methods of fertilization were compared. Eight fields were fertilized using the new method and five by old method. The crop of wheat on hectares in tons

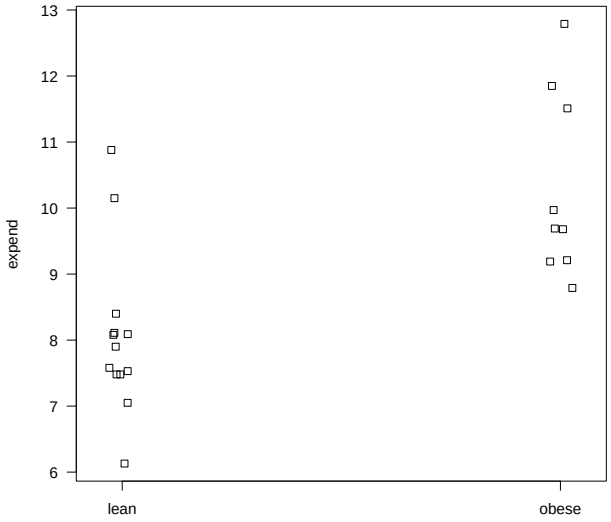

Figure 13.3: Stripcharts

Table 13.1: Yield of wheat in tons on hectare

X_i	5.7	5.5	4.3	5.9	5.2	5.6	5.8	5.1
Y_i	5.0	4.5	4.2	5.4	4.4			

are labeled X_i for new method and Y_i for the old method. The data are presented in tab. 13.1. We should find out if the method of fertilization has influence on the yield of wheat.

First, we construct graph of the both distribution functions (see fig. 13.4).

```
new <- c(5.7,5.5,4.3,5.9,5.2,5.6,5.8,5.1)
old <- c(5.0,4.5,4.2,5.4,4.4)
plot.ecdf(new, verticals=T, las=1, xlab="", ylab="", pch="", main="")
plot.ecdf(old, verticals=T, add=T, pch="", lty=2)
```

The full line shows ecdf of yields concerning the new method of fertilization and the dashed line concerns the old one.

In our example we have $D_{8,5} = 0.675$. Critical value on the level 5 per cent is 0.75. (By the way, the approximate critical value is $D^*_{8,5}(0.05) = 0.774$.) Since $D_{8,5} < 0.75$, Kolmogorov-Smirnov test does not reject the hypothesis that both samples are from populations with the same distribution functions.

Both samples have small sizes. Application of theorem 13.1 is recommended for $m + n > 35$. If we use the approximation, we get $\lambda_0 = 1.184$, $K(\lambda_0) = 0.879$. Since $K(\lambda_0) < 0{,}95$, we cannot reject the hypothesis.

Using the program `R` the calculations can be done in the following way.

```
ks.test(new,old)
        Two-sample Kolmogorov-Smirnov test
data:  new and old D = 0.675, p-value = 0.07925 alternative
hypothesis: two.sided
```

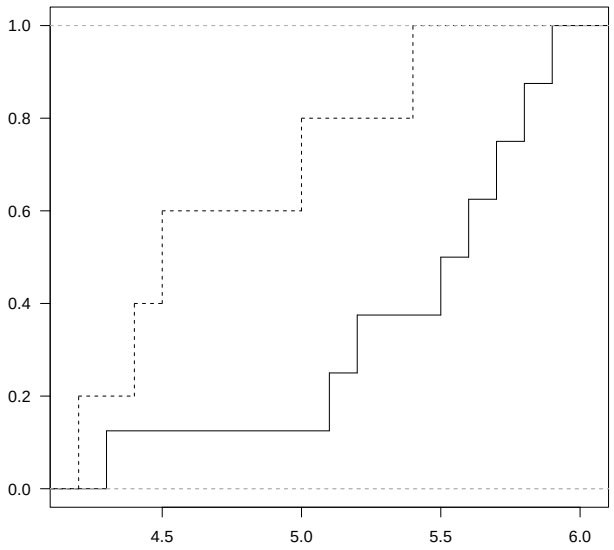

Figure 13.4: Two-sample Kolmogorov-Smirnov test

♢

13.3 Two-sample t test

Let $X_1, \ldots, X_m$ be a sample from $\mathsf{N}(\mu_1, \sigma^2)$ and let $Y_1, \ldots, Y_n$ be a sample from $\mathsf{N}(\mu_2, \sigma^2)$. Assume that $m \geq 2$, $n \geq 2$, $\sigma^2 > 0$, and that both samples are mutually independent. Denote

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$
$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Under these assumptions the following theorem holds.

Theorem 13.3 *Random variable*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}} \sqrt{\frac{mn(m+n-2)}{m+n}} \tag{13.2}$$

has distribution t_{m+n-2}.

Proof. Theorems 4.6 (b) on p. 48 and 3.6 on p. 38 give that

$$Z = [(m-1)S_X^2 + (n-1)S_Y^2]/\sigma^2 \sim \chi^2_{m+n-2}.$$

Similarly as in theorem 4.6 (c) can be proved that the variables $\bar{X} - \bar{Y}$ and $(m-1)S_X^2 + (n-1)S_Y^2$ are independent. We get

$$\bar{X} - \bar{Y} \sim \mathsf{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right).$$

In the same time we have

$$U = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{m} + \frac{\sigma^2}{n}}} \sim \mathsf{N}(0,1).$$

Theorem 3.10 ensures that the variable $U \Big/ \sqrt{Z/(m+n-2)}$ has distribution t_{m+n-2}. However, $U \Big/ \sqrt{Z/(m+n-2)} = T$. □

Two-sample t-test is the test of hypothesis $H_0 : \mu_1 - \mu_2 = \delta$, where δ is a given number (very often it is $\delta = 0$). If it is a test against alternative $H_1 : \mu_1 - \mu_2 \neq \delta$, the procedure is following one. First, we calculate T from formula (13.2) where we insert $\mu_1 - \mu_2 = \delta$. If $|T| \geq t_{m+n-2}(\alpha)$, we reject the hypothesis H_0 on the level α. The one-sided tests are analogous.

Among assumptions of two-sample t test we have that both samples come from normal populations with the same variance. The violation of these assumptions does not influence the result of the test too much. Yet in the case of substantial non-normality we prefer some non-parametric tests (very often it is two-sample Wilcoxon test, see section 13.6). Assumption that the variances are equal can be checked by help of the F test (see section 13.5). If it is known that the variances are different the *Welch test* is applied (see section 13.4).

Example 13.4 Eleven piglings were randomly divided into two groups. The first group contained 6 piglings and they were fed by diet A. In the second group were 5 piglings on diet B. Average daily increments after 6 months are given in tab. 13.2. It should be established if both diets are equally effective.

Table 13.2: Average daily increments in dkg

Diet A	62	54	55	60	53	58
Diet B	52	56	49	50	51	

Data are plotted in fig. 13.5 and we apply two-sample t test.

```
da<-c(62,54,55,60,53,58) # data
db<-c(52,56,49,50,51)
boxplot(da, db, names=c("da", "db"), boxwex=0.3, las=1) # graph
t.test(da,db,var.equal=T) # test
```

```
        Two Sample t-test
data:  da and db
t = 2.7712, df = 9, p-value = 0.02171
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9919634 9.8080366
sample estimates:
mean of x mean of y
     57.0      51.6
```

Since p-value is smaller than 0.05, we reject hypothesis that both diets are equally effective. ♢

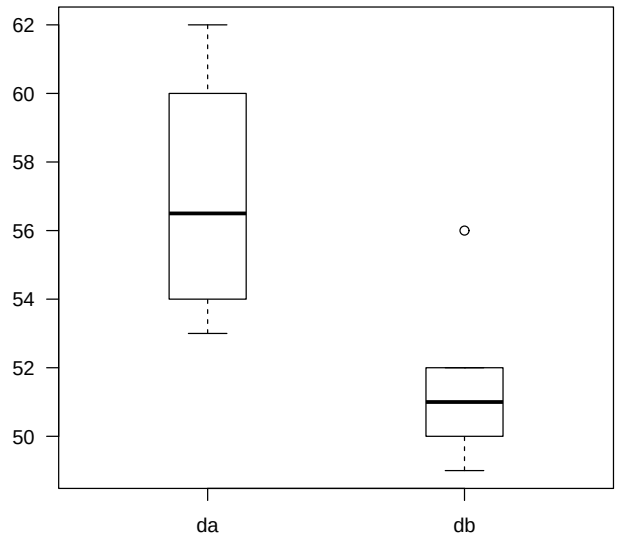

Figure 13.5: Boxplots for diet a and b

13.4 Two-sample Welch test

Let $X_1, \ldots, X_{n_1}$ be a sample from $\mathsf{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_{n_2}$ an independent sample from $\mathsf{N}(\mu_2, \sigma_2^2)$. Assume that $\sigma_1 > 0$, $\sigma_2 > 0$, $n_1 \geq 2$, $n_2 \geq 2$. Define $f_1 = n_1 - 1$, $f_2 = n_2 - 1$,

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i, \qquad S_1^2 = \frac{1}{f_1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{f_2}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2.$$

Welch test (see Welch 1938) was derived for testing the hypothesis $H_0 : \mu_1 = \mu_2$ against one-sided or two-sided alternative. If variances σ_1^2 and σ_2^2 were known, we would use the test statistic

$$\xi = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}, \tag{13.3}$$

which has distribution $\mathsf{N}(0, 1)$ under H_0. If the variances σ_1^2 and σ_2^2 are not known, it is natural to substitute them by their estimators S_1^2 a S_2^2. Thus we will investigate the test statistic

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}. \tag{13.4}$$

The statistics can be written in the form

$$t = \frac{\dfrac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}{\sqrt{\underbrace{\dfrac{f_1 S_1^2}{\sigma_1^2}}_{\eta_1}\dfrac{\sigma_1^2}{n_1 f_1} + \underbrace{\dfrac{f_2 S_2^2}{\sigma_2^2}}_{\eta_2}\dfrac{\sigma_2^2}{n_2 f_2}}}.$$

Define

$$\eta_1 = \frac{f_1 S_1^2}{\sigma_1^2}, \qquad \eta_2 = \frac{f_2 S_2^2}{\sigma_2^2}.$$

We know that

$$\eta_1 =\sim \chi^2_{f_1}, \qquad \eta_2 =\sim \chi^2_{f_2}.$$

Inserting from (13.3) we get

$$t = \frac{\xi}{\sqrt{\underbrace{\frac{\frac{\sigma_1^2}{n_1 f_1}}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}_{a}\eta_1 + \underbrace{\frac{\frac{\sigma_2^2}{n_2 f_2}}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}_{b}\eta_2}}.$$

Let

$$a = \frac{\frac{\sigma_1^2}{n_1 f_1}}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \qquad b = \frac{\frac{\sigma_2^2}{n_2 f_2}}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, \qquad w = a\eta_1 + b\eta_2.$$

Then

$$t = \frac{\xi}{\sqrt{w}}.$$

To calculate exact distribution of the variable t would be difficult. Welch suggested the following approximation of the distribution of the variable w:

$$\mathcal{L}(w) \doteq \mathcal{L}(gZ),$$

where g is an appropriate constant and Z is a random variable having distribution χ^2_f. The number of degrees f must be determined. The constants g and f can be calculated using moment method from conditions

$$\begin{aligned} \mathsf{E}a\eta_1 + \mathsf{E}b\eta_2 &= \mathsf{E}gZ, \\ \mathsf{var}\, a\eta_1 + \mathsf{var}\, b\eta_2 &= \mathsf{var}\, gZ. \end{aligned}$$

Thus we obtain system of equations

$$\begin{aligned} af_1 + bf_2 &= gf, \\ 2(a^2 f_1 + b^2 f_2) &= 2g^2 f. \end{aligned}$$

The solution is

$$g = \frac{a^2 f_1 + b^2 f_2}{af_1 + bf_2}, \qquad f = \frac{(af_1 + bf_2)^2}{a^2 f_1 + b^2 f_2}.$$

In the frame of the used approximation we have

$$\frac{w}{g} \sim \chi^2_f.$$

Thus

$$\frac{\xi}{\sqrt{\frac{w}{g}\cdot\frac{1}{f}}} \sim \mathsf{t}_f.$$

A simple derivation gives

$$gf = af_1 + bf_2 = 1.$$

It implies

$$\frac{\xi}{\sqrt{w}} \sim \mathsf{t}_f,$$

where

$$f = \frac{(af_1+bf_2)^2}{a^2f_1+b^2f_2} = \frac{1}{a^2f_1+b^2f_2} = \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2f_1}+\frac{\sigma_2^4}{n_2^2f_2}}.$$

In this formula the unknown parameters σ_1^2 a σ_2^2 are substituted by estimators S_1^2 a S_2^2. Instead of f the degrees of freedom are taken as

$$f^* = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2f_1}+\frac{S_2^4}{n_2^2f_2}}$$

and the approximation $t \sim \mathsf{t}_{f^*}$ is used.

It can be derived that

$$\min\{f_1, f_2\} \le f^* \le f_1 + f_2.$$

Example 13.5 We use data from example 13.4. We get

```
da<-c(62,54,55,60,53,58) db<-c(52,56,49,50,51) t.test(da,db)
        Welch Two Sample t-test
data:  da and db t = 2.8487, df = 8.947, p-value = 0.01924
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.107981 9.692019
sample estimates: mean of x mean of y
     57.0      51.6
```

Since the p-value is smaller than 0.05, we reject the hypothesis that the true difference in means is equal to 0. $\diamondsuit$

13.5 Test of equality of variances

Theorem 13.6 *Let $X_1, \ldots, X_n$ be a sample from $\mathsf{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m$ a sample from $\mathsf{N}(\mu_2, \sigma_2^2)$. Let these two samples be independent. Assume that $n \ge 2$, $m \ge 2$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$. Let $\bar{X}$, S_X^2 and $\bar{Y}$, S_Y^2 be characteristics of the samples. If the equality $\sigma_1^2 = \sigma_2^2$ holds, then $Z = S_X^2/S_Y^2 \sim \mathsf{F}_{n-1,m-1}$.*

Proof. The assertion follows from theorem 3.11 on p. 40. □

We shall test hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$. We reject H_0 if $S_X^2/S_Y^2 \leq k_1$, or $S_X^2/S_Y^2 \geq k_2$. The constants k_1 and k_2 are chosen so that under H_0

$$\mathsf{P}(S_X^2/S_Y^2 \leq k_1) = \alpha/2, \qquad \mathsf{P}(S_X^2/S_Y^2 \geq k_2) = \alpha/2. \tag{13.5}$$

Then the probability of error of the first kind is α. Theorem 13.6 gives that (13.5) holds in the case

$$k_1 = F_{n-1,m-1}\left(1 - \frac{\alpha}{2}\right) = \frac{1}{F_{m-1,n-1}\left(\frac{\alpha}{2}\right)}, \qquad k_2 = F_{n-1,m-1}\left(\frac{\alpha}{2}\right), \tag{13.6}$$

where $F_{u,v}(\alpha)$ is the critical value of F distribution. It is recommended to introduce such denotation that $S_X^2 \geq S_Y^2$. Then it suffices to see if the inequality $S_X^2/S_Y^2 \geq F_{n-1,m-1}\left(\frac{\alpha}{2}\right)$ holds, and it is not needed to calculate the inverse of critical values.

If the variances are σ_1^2 and σ_2^2 and (13.6) holds, then

$$\mathsf{P}\left\{\frac{S_X^2}{S_Y^2}\frac{1}{k_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_X^2}{S_Y^2}\frac{1}{k_1}\right\} = 1 - \alpha.$$

Then

$$\left(\frac{S_X^2}{S_Y^2}\frac{1}{F_{n-1,m-1}\left(\frac{\alpha}{2}\right)}, \frac{S_X^2}{S_Y^2}F_{m-1,n-1}\left(\frac{\alpha}{2}\right)\right)$$

is two-sided confidence interval for the ratio σ_1^2/σ_2^2 with confidence coefficient $1 - \alpha$.

Since F-test is sensitive to assumption about equality of variances, the *Levene test* is often used instead (see section 15.6, p. 143).

Example 13.7 Consider again data introduced in example 13.4. We have

```
da<-c(62,54,55,60,53,58)
db<-c(52,56,49,50,51)
var.test(da,db)
```

```
        F test to compare two variances
data:  da and db
F = 1.7534, num df = 5, denom df = 4, p-value = 0.6063
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.1872423 12.9541010
sample estimates: ratio of variances
          1.753425
```

Since p-value is larger than 0.05, the hypothesis that the true ratio of variances is equal to 1 is not rejected. Now, we show Levene test.

```
prirust <- c(da, db)
dieta <- factor(c(rep("A",6), rep("B",5)))
library(car)
leveneTest(prirust, dieta)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.3439 0.2762
       9
```

Since p-value 0.2762 is larger than 0.05, Levene test also does not reject hypothesis that true ratio of variances is equal to 1.

Let us remark that in the case of data prepared like in this example the classical two-sample t-test will be calculated using

```
t.test(prirust ~ dieta, var.equal=T)
```

and Welch two-sample t-test using

```
t.test(prirust ~ dieta).
```

$\diamondsuit$

13.6 Two-sample Wilcoxon test

Let $X_1, \ldots, X_m$ be a random sample from a continuous distribution with a distribution function F and $Y_1, \ldots, Y_n$ an independent random sample with a distribution function G.

We want to test the hypothesis $H_0 : F = G$ against the alternative $H_1 : F \neq G$. All $m + n$ variables $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ form a *pooled sample*. Let T_1 be sum of ranks $X_1, \ldots, X_m$ and T_2 sum of ranks $Y_1, \ldots, Y_n$. It is clear that

$$T_1 + T_2 = \frac{1}{2}(m+n)(m+n+1).$$

First we investigate general properties of tests of this type. Let $N = m + n$ and let R_i be the rank of the ith variable. Finally, let $a(i)$ be a function defined for $i = 1, \ldots, N$. The variable

$$S = \sum_{i=1}^{N} c_i\, a(R_i)$$

is called *simple linear rank statistic*.

The numbers $c_1, \ldots, c_N$ are *regression constants* and $a(i)$ are *scores*. Define

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a(i), \qquad \bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i,$$
$$\sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N}[a(i) - \bar{a}]^2, \qquad \sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \bar{c})^2.$$

Theorem 13.8 *If H_0 holds, then*

$$\mathsf{E}S = N\bar{a}\bar{c}, \qquad \mathsf{var}\, S = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2.$$

Proof. If H_0 holds, then R_i is a random variable which is equal to every value $1, \ldots, N$ with probability $1/N$. Thus

$$\mathsf{E}a(R_i) = \sum_{t=1}^{N} a(t)\frac{1}{N} = \bar{a}, \tag{13.7}$$

so that

$$\mathsf{E}S = \sum_{i=1}^{N} c_i \,\mathsf{E}a(R_i) = \sum_{i=1}^{N} c_i \bar{a} = N\bar{a}\bar{c}.$$

If $i \neq j$, then under H_0 we have

$$\mathsf{P}(R_i = s,\, R_j = t) = \frac{1}{N(N-1)} \qquad \text{for} \quad 1 \leq s \neq t \leq N.$$

Using (13.7) we get

$$\mathsf{var}\, a(R_i) = \mathsf{E}[a(R_i) - \mathsf{E}a(R_i)]^2 = \mathsf{E}[a(R_i) - \bar{a}]^2 = \frac{1}{N}\sum_{t=1}^{N}[a(t) - \bar{a}]^2 = \sigma_a^2$$

and then for $i \neq j$

$$\begin{aligned}
\mathsf{cov}[a(R_i), a(R_j)] &= \mathsf{E}[a(R_i) - \bar{a}][a(R_j) - \bar{a}] \\
&= \frac{1}{N(N-1)} \sum_{s \neq t} \sum [a(s) - \bar{a}][a(t) - \bar{a}] \\
&= \frac{1}{N(N-1)} \left\{ \sum_{s} \sum_{t} [a(s) - \bar{a}][a(t) - \bar{a}] - \sum_{s} [a(s) - \bar{a}]^2 \right\} \\
&= -\frac{1}{N(N-1)} \sum_{s} [a(s) - \bar{a}]^2 = -\frac{1}{N-1}\sigma_a^2.
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathsf{var}\, S &= \sum_i c_i^2 \,\mathsf{var}\; a(R_i) + \sum_{i \neq j} \sum c_i c_j \,\mathsf{cov}[a(R_i), a(R_j)] \\
&= \sigma_a^2 \left(\sum_i c_i^2 - \frac{1}{N-1} \sum_{i \neq j} \sum c_i c_j \right) \\
&= \sigma_a^2 \left(\sum_i c_i^2 - \frac{1}{N-1} \sum_i \sum_j c_i c_j + \frac{1}{N-1} \sum_i c_i^2 \right) \\
&= \frac{\sigma_a^2}{N-1} \left[N \sum_i c_i^2 - \Big(\sum c_i \Big)^2 \right] = \frac{N}{N-1}\sigma_a^2 \sum_i (c_i - \bar{c})^2 \\
&= \frac{N^2}{N-1}\sigma_a^2 \sigma_c^2. \qquad \square
\end{aligned}$$

Theorem 13.9 *If H_0 holds, then*

$$\mathsf{E}T_1 = \frac{1}{2}\, m(m+n+1), \qquad \mathsf{var}\, T_1 = \frac{1}{12}\, mn(m+n+1).$$

Proof. Variable T_1 is special case of S, if we insert $a(i) = i$ and

$$c_i = \begin{cases} 1 & \text{pro} \quad i = 1, \ldots, m, \\ 0 & \text{pro} \quad i = m+1, \ldots, m+n. \end{cases}$$

Define again $N = m + n$. We have

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2}, \qquad \sigma_a^2 = \frac{1}{N}\sum_{i=1}^{N} i^2 - \bar{a}^2 = \frac{1}{12}(N+1)(N-1),$$
$$\bar{c} = \frac{m}{N}, \qquad \sigma_c^2 = \frac{1}{N}\sum_{i=1}^{N} c_i^2 - \bar{c}^2 = \frac{mn}{N^2}.$$

It follows from theorem 13.8 that

$$\mathsf{E}T_1 = N\bar{a}\bar{c} = \frac{m(m+n+1)}{2}, \qquad \mathsf{var}\, T_1 = \frac{N^2}{N-1}\sigma_a^2\sigma_c^2 = \frac{mn(m+n+1)}{12}. \qquad \square$$

Instead of T_1 the variable $U_1 = mn + \frac{1}{2}m(m+1) - T_1$ is used. The test based on U_1 is called *Mann-Whitney test*. Let $U_2 = mn + \frac{1}{2}n(n+1) - T_2$. Then $U_1 + U_2 = mn$. The statistic U_1 gives number of cases when $X_i < Y_j$. Similarly, U_2 gives number of cases when $X_i > Y_j$ holds. If $\min(U_1, U_2)$ is less or equal to the critical value then the hypothesis H_0 is rejected. The samples are denoted so that $m \geq n$. If m and n are large, the following procedure is used. It follows from theorem 13.9 that

$$\mathsf{E}U_1 = \frac{1}{2}mn, \qquad \mathsf{var}\, U_1 = \frac{1}{12}mn(m+n+1). \tag{13.8}$$

Since $U_2 = mn - U_1$, we have $\mathsf{E}U_2 = \mathsf{E}U_1$, $\mathsf{var}\, U_2 = \mathsf{var}\, U_1$. It is proved that for $m \to \infty$ and $n \to \infty$ the variable U_1 (as well as the variable T_1) has asymptotically normal distribution. We obtain

$$U = \frac{U_1 - EU_1}{\sqrt{\mathsf{var}\, U_1}}, \tag{13.9}$$

where $\mathsf{E}U_1$ and $\mathsf{var}\, U_1$ are given in (13.8). If $|U| \geq u(\frac{\alpha}{2})$, H_0 is rejected on the level which tends to α. The test based on (13.9) can be used for $m > 10$, $n > 10$.

If some variables are equal, we have *ties*. Then such a variable gets corresponding average rank. If the number of ties is large, instead of U introduced in formula (13.9) the variable

$$z = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn}{S(S-1)}\left(\frac{S^3-S}{12} - \sum_{i=1}^{r}\frac{t_i^3 - t_i}{12}\right)}},$$

is used, where $S = m + n$, r is number of ties and t_i is multiplicity of the i-th tie.

Although the Wilcoxon test is formulated as a test against a general alternative, it is sensitive especially to a *shift* $H_1' : G(x) = F(x - \Delta)$, $\Delta \neq 0$. For other alternatives, *Kolmogorov-Smirnov test* described in section 13.2 on p. 108 is recommended.

Example 13.10 We shall analyze data introduced in example 13.2, p. 109. The numbers are ordered. The values X_i are underlined and their rank is introduced (see tab. 13.3).

Thus

$$T_1 = 70, \qquad U_1 = 6, \qquad T_2 = 21, \qquad U_2 = 34.$$

Critical value is 6. Since $\min(U_1, U_2) = 6 \leq 6$, we reject the hypothesis. The asymptotic procedure calculated in (13.9) gives $U = -2.049$. We have $|U| \geq u(0.025) = 1.96$, and so we would reject the hypothesis, that both diets are equally effective, also by this procedure.

Program `R` gives

Table 13.3: Yields of wheat and their ranks

Yields	4.2	4.3	4.4	4.5	5.0	5.1	5.2	5.4	5.5	5.6	5.7	5.8	5.9
Ranks		2				6	7		9	10	11	12	13

```
new <- c(5.7,5.5,4.3,5.9,5.2,5.6,5.8,5.1)
old <- c(5.0,4.5,4.2,5.4,4.4)
wilcox.test(new, old, alternative = "two.sided", exact=T, correct=F)
```

```
        Wilcoxon rank sum test
data:  new and old
W = 34, p-value = 0.04507
alternative hypothesis: true mu is not equal to 0
```

The hypothesis is also rejected. ♢

Chapter 14

Discrete two-sample problem

14.1 Testing homogeneity of two binomial distributions

Let p_1 be probability that the event A occurs in an experiment. Assume that in m independent experiments the event A occurred X times. Then the experiment is repeated under different conditions so that the event A occurs with probability p_2. Assume that in those n additional experiments the event A occurred Y times. Using these data we wont to test hypothesis $H_0 : p_1 = p_2$ against alternative $H_1 : p_1 \neq p_2$. The hypothesis H_0 is called *hypothesis of homogeneity of two binomial distributions*, since $X \sim \mathsf{Bi}(m, p_1)$ and $Y \sim \mathsf{Bi}(n, p_2)$.

This is one of the oldest statistical problems which occurs frequently till now.

Denote $x = X/m$, $y = Y/n$. We shall assume that $x(1-x) + y(1-y) \neq 0$. Central limit theorem gives that for large values m and n the approximation

$$x \sim \mathsf{N}\left[p_1, \frac{p_1(1-p_1)}{m}\right], \qquad y \sim \mathsf{N}\left[p_2, \frac{p_2(1-p_2)}{n}\right]$$

can be used. Since x a y are independent variables, in the frame of this approximation we have result

$$\frac{x - y - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{m} + \dfrac{p_2(1-p_2)}{n}}} \sim \mathsf{N}(0,1).$$

If H_0 holds, then we insert in numerator simply $p_1 - p_2 = 0$. However, in denominator the unknown values p_1 and p_2 remain. They can be substituted by their estimators. It can be proved that the limiting distribution remains the same, namely $\mathsf{N}(0,1)$. Two kinds of estimators of the parameters p_1 and p_2 can be used. Thus we have two variants of the test of homogeneity.

In the first case we use x as estimator of the parameter p_1 and y as estimator of p_2. The strong law of large numbers ensures that $x \to p_1$ and $y \to p_2$ almost surely. To test the hypothesis H_0 we calculate variable

$$U_a = \frac{x - y}{\sqrt{\dfrac{x(1-x)}{m} + \dfrac{y(1-y)}{n}}}. \tag{14.1}$$

If $|U_a| \geq u(\frac{\alpha}{2})$, we reject H_0. The tests of H_0 against one-sided alternatives are similar.

In the second case we use the fact that under H_0 we have $p_1 = p_2$. This value can be estimated by

$$z = \frac{X+Y}{m+n} = \frac{mx+ny}{m+n}.$$

Thus we calculate

$$U_b = \frac{x-y}{\sqrt{z(1-z)\left(\frac{1}{m}+\frac{1}{n}\right)}}. \tag{14.2}$$

If $|U_b| \geq u(\frac{\alpha}{2})$, we reject H_0. Similar procedure could be used for one-sided alternatives.

Some information about properties of U_a and U_b is contained in the following assertion published in the paper Eberhardt, Fliegner (1977).

Theorem 14.1 *If $m = n$, then $|U_b| \leq |U_a|$ and the equality holds if and only if $x = y$.*

Proof. Function $f(x) = x(1-x)$ is concave for $x \in [0,1]$. Thus for all $\gamma \in [0,1]$ and for all numbers $a, b \in [0,1]$ it holds

$$f[\gamma a + (1-\gamma)b] \geq \gamma f(a) + (1-\gamma) f(b).$$

Choose $\gamma = \frac{1}{2}$, $a = x$, $b = y$. Then we get

$$\frac{1}{2}(x+y)\left[1 - \frac{1}{2}(x+y)\right] \geq \frac{1}{2}x(1-x) + \frac{1}{2}y(1-y). \tag{14.3}$$

For $m = n$ we have $z = (x+y)/2$. Numerator of the fraction (14.1) has under the root

$$\frac{2}{m}\left[\frac{1}{2}x(1-x) + \frac{1}{2}y(1-y)\right]$$

and denominator of the fraction (14.2) has under the root

$$\frac{2}{m}\left[\frac{x+y}{2}\left(1 - \frac{x+y}{2}\right)\right].$$

Inequality (14.3) gives that $|U_b| \leq |U_a|$. The equality in (14.3) holds if and only if $x = y$. □

We remark that for $m \neq n$ there is no simple relation between U_a and U_b.

Seemingly for $m = n$ it would be better to use the test which is based on U_a, since it is more powerful. However, if the size of the sample is small, the probability of the error of the first kind is larger than α. For example, for $\alpha = 0.05$ and $m = n = 20$ the probability of the error of the first kind is 0.081 and for $m = 20$, $n = 40$ even 0.085.

Data can be written in the form of a *contingency table* (see tab. 14.1). If we use in this table the denotation of frequencies by n_{ij} introduced in section 10.5, we get

$$U_b^2 = \frac{\left(\frac{n_{11}}{n_{1.}} - \frac{n_{21}}{n_{2.}}\right)^2}{\frac{n_{.1}}{n}\frac{n_{.2}}{n}\frac{n}{n_{1.}n_{2.}}} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

Table 14.1: Table of successes and failures

	Successes	Failures	Total
1. experiment	X	$m-X$	m
2. experiment	Y	$n-Y$	n
Total	$X+Y$	$m+n-X-Y$	$m+n$

The result is the formula for the variable χ^2 in 2×2 *contingency table* (see formula (10.19)). We can say that frequently used statistic is U_b, whereas U_a is rare.

Comparison of probabilities p_1 and p_2 in small samples can be found in the papers Santer, Snell (1980), Gart, Nam (1990), D'Agostino et al. (1988).

Example 14.2 It is not recommended to change quickly hot and cold meals because then the teeth have temperature shock. In an experiment 50 teeth were plunged into boiled and cold water. Other 50 teeth were in boiled water without temperature shocks. Finally, all teeth were crushed. From 50 teeth with temperature shocks 21 teeth were broken. From 50 teeth without shocks were only 11 broken. It should be tested if the temperature shocks influence mechanical resistance of teeth. (Osborn 1979).

We have $m = n = 50$, $X = 21$, $Y = 11$, $x = 0{,}42$, $y = 0{,}22$, $z = 0{,}32$. This gives

$$U_a = 2{,}195, \qquad U_b = 2{,}144.$$

If we test the hypothesis of homogeneity against two-sided alternative, we would compare the calculated results with the critical value $u(0.025) = 1.96$. The result would be statistically significant. It means that the effect of temperature shocks is statistically proved.

However, our example leads to the test against one-sided alternative. Shocks cannot fasten the teeth. Thus the test statistics should be compared with the critical value $u(0.05) = 1.645$. The results larger than this critical values are significant. In the introduced example the hypothesis H_0 must be rejected.

In program `R` we get

```
zlomene.zuby <- c(21,11)
zuby.celkem <- c(50,50)
prop.test(zlomene.zuby, zuby.celkem)
        2-sample test for equality of proportions with continuity correction
data:  zlomene.zuby out of zuby.celkem X-squared = 3.7224, df = 1, p-value = 0.05369
alternative hypothesis: two.sided 95 percent
confidence interval:
 0.001395761 0.398604239
sample estimates: prop 1 prop 2
  0.42   0.22
```

Since we used the continuity correction, the result is not significant. If we do not use the correction, we would have

```
prop.test(zlomene.zuby, zuby.celkem, correct = F)
 2-sample test for equality of proportions without continuity correction
```

```
data:  zlomene.zuby out of zuby.celkem X-squared = 4.5956, df = 1, p-value = 0.03205
alternative hypothesis: two.sided 95 percent
confidence interval:
 0.02139576 0.37860424
sample estimates: prop 1 prop 2
  0.42   0.22
```

It gives significant result. One-sided test is

```
prop.test(zlomene.zuby, zuby.celkem, correct = F,
alternative="greater")
```

which gives significant result

```
X-squared = 4.5956, df = 1, p-value = 0.01603
```

♢

14.2 Confidence interval for difference of probabilities

Let $X \sim \mathsf{Bi}(m, p_1)$ and $Y \sim \mathsf{Bi}(n, p_2)$ be independent random variables. The problem is to find confidence interval for $\Delta = p_1 - p_2$. A review of 11 methods described in papers for a construction of this interval can be found in Newcombe (1998). Here we introduce only three of them.

Denote $\hat{p}_1 = X/m$, $\hat{p}_2 = Y/n$, $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$, $q_i = 1 - p_i$, $\hat{q}_i = 1 - \hat{p}_i$ $(i = 1, 2)$. Let u_α be the critical value of the distribution $\mathsf{N}(0, 1)$ on the level α. Then $\hat{p}_1\hat{q}_1/m + \hat{p}_2\hat{q}_2/n$ is estimator of $\mathsf{var}\,\Delta$. Since the random variable

$$T = (\hat{\Delta} - \Delta) \Big/ \sqrt{\hat{p}_1\hat{q}_1/m + \hat{p}_2\hat{q}_2/n}$$

has asymptotically distribution $\mathsf{N}(0, 1)$, *Wald confidence interval* is

$$\hat{\Delta} \mp u_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/m + \hat{p}_2\hat{q}_2/n}.$$

Denote $u = u_{\alpha/2}$. Let (l_i, u_i) be the *Wilson confidence interval* for p_i. As we know, endpoints of this interval are roots of quadratic equation

$$(\hat{p}_i - p_i)^2 = u^2 p_i(1 - p_i)/n_i.$$

Then *Newcomb confidence interval* is

$$\left(\hat{\Delta} - u\sqrt{\frac{l_1(1-l_1)}{m} + \frac{u_2(1-u_2)}{n}},\ \hat{\Delta} + u\sqrt{\frac{u_1(1-u_1)}{m} + \frac{l_2(1-l_2)}{n}}\right).$$

Confidence interval based on score test is obtained by inverting test

$$\frac{\hat{p}_1 - \hat{p}_2 - \Delta}{\hat{\sigma}} = u_\alpha,$$

where

$$\hat\sigma^2 = \frac{\hat p_1(1-\hat p_1)}{m} + \frac{(\hat p_1-\Delta)(1-\hat p_1+\Delta)}{n},$$

and $\hat p_1$ is maximum likelihood estimator of the probability p_1 under the condition $\hat p_1 - \hat p_2 = \Delta$. Explicit solution is not known, numerical methods can be used.

The following conclusions follow from numerical studies (see Brown, Li 2003):

- If the sizes m and n are small, then the Wald confidence interval has confidence coefficient smaller than its nominal value. However, its length is short.
- Score and Newcomb intervals behave similarly and have good properties.
- If $\min(m,n) \ge 50$, then all confidence intervals give good results.

Thus Brown, Li (2003) apart from the test suggested in their paper and called *re-centred* recommend to use score and Newcomb intervals. Because there exists explicit formula for the Newcomb interval, its use can be especially recommended.

14.3 Confidence interval for ratio of probabilities

Let $X \sim \mathsf{Bi}(n_1, p_1)$ and $Y \sim \mathsf{Bi}(n_2, p_2)$. Assume that X and Y are independent variables. Denote

$$\theta = \frac{p_1}{p_2}, \qquad q_1 = 1-p_1, \qquad q_2 = 1-p_2, \qquad x = \frac{X}{n_1}, \qquad y = \frac{Y}{n_2}.$$

It is known that

$$\mathsf{E}x = p_1, \qquad \mathsf{E}y = p_2, \qquad \mathsf{var}\, x = \frac{p_1 q_1}{n_1}, \qquad \mathsf{var}\, y = \frac{p_2 q_2}{n_2}.$$

Introduce function $g(u,v) = u/v$. The value $g(x,y) = x/y$ can be expressed using Taylor expansion around the point (p_1, p_2). If we neglect the remaining term we have

$$g(x,y) = g(p_1,p_2) + \frac{1}{p_2}(x-p_1) - \frac{p_1}{p_2^2}(y-p_2).$$

In the frame of this approximation we have

$$\mathsf{E}\,\frac{x}{y} = \frac{p_1}{p_2},$$

$$\mathsf{E}\,\frac{x^2}{y^2} = \frac{p_1^2}{p_2^2} + \frac{1}{p_2^2}\frac{p_1 q_1}{n_1} + \frac{p_1^2}{p_2^4}\frac{p_2 q_2}{n_2},$$

$$\mathsf{var}\,\frac{x}{y} = \mathsf{E}\frac{x^2}{y^2} - \left(\mathsf{E}\frac{x}{y}\right)^2 = \frac{p_1^2}{p_2^2}\left(\frac{q_1}{p_1 n_1} + \frac{q_2}{p_2 n_2}\right).$$

It can be proved that for $n_1 \to \infty$, $n_2 \to \infty$ the ratio x/y has asymptotically normal distribution and

$$U^* = \frac{\dfrac{x}{y} - \dfrac{p_1}{p_2}}{\dfrac{p_1}{p_2}\sqrt{\dfrac{q_1}{p_1 n_1} + \dfrac{q_2}{p_2 n_2}}}$$

has asymptotically $\mathsf{N}(0,1)$ distribution. An exact proof of this assertion can be performed using *delta method*. In the denominator p_1 is approximated using x, then p_2 using y, q_1 using $1-x$ and q_2 using $1-y$. Then

$$U = \frac{\dfrac{x}{y} - \theta}{\dfrac{x}{y}\sqrt{\dfrac{1-x}{xn_1} + \dfrac{1-y}{yn_2}}}$$

has asymptotically $\mathsf{N}(0,1)$ distribution, so that the event

$$\frac{\left|\dfrac{x}{y} - \theta\right|}{\dfrac{x}{y}\sqrt{\dfrac{1-x}{xn_1} + \dfrac{1-y}{yn_2}}} \le u_{\alpha/2}$$

has asymptotically probability $1-\alpha$. This implies that

$$\frac{x}{y}\left(1 \pm u_{\alpha/2}\sqrt{\frac{1-x}{xn_1} + \frac{1-y}{yn_2}}\right)$$

are endpoints of the confidence interval for $\theta = p_1/p_2$ with confidence coefficient which is asymptotically equal to $1-\alpha$. A detailed derivation of this confidence interval and some other intervals can be found in the paper Noether (1957).

14.4 Test of hypothesis $\lambda_1/\lambda_2 = r$

Let X_1 and X_2 be independent random variables such that $X_1 \sim \mathsf{Po}(\lambda_1)$, $X_2 \sim \mathsf{Po}(\lambda_2)$. Denote $\gamma = \lambda_1/\lambda_2$. Consider a test of hypothesis $H_0 : \gamma = r$ against several alternatives.

We have $\lambda_1 = \gamma\lambda_2$. Since $X_1 + X_2 \sim \mathsf{Po}(\lambda_1 + \lambda_2) = \mathsf{Po}(\gamma\lambda_2 + \lambda_2)$, we obtain

$$\begin{aligned}\mathsf{P}(X_1 = x_1, X_2 = x_2 | X_1 + X_2 = x_1 + x_2) &= \frac{\dfrac{(\gamma\lambda_2)^{x_1}}{x_1!}\mathrm{e}^{-\gamma\lambda_2}\dfrac{\lambda_2^{x_2}}{x_2!}\mathrm{e}^{-\gamma_2}}{\dfrac{(\gamma\lambda_2+\lambda_2)^{x_1+x_2}}{(x_1+x_2)!}\mathrm{e}^{-(\gamma\lambda_2+\lambda_2)}} \\ &= \binom{x_1+x_2}{x_1}\left(\frac{\gamma}{\gamma+1}\right)^{x_1}\left(1-\frac{\gamma}{\gamma+1}\right)^{x_2}.\end{aligned}$$

Conditional distribution is binomial. When testing H_0 against the alternative $H_1 : \gamma > r$, then H_0 will be rejected in the case

$$\sum_{i=x_1}^{x_1+x_2}\binom{x_1+x_2}{i}\left(\frac{r}{r+1}\right)^{i}\left(1-\frac{r}{r+1}\right)^{x_1+x_2-i} \le \alpha. \tag{14.4}$$

Similarly, when testing H_0 against $H_1' : \gamma < r$ we reject H_0 in the case that

$$\sum_{i=0}^{x_1}\binom{x_1+x_2}{i}\left(\frac{r}{r+1}\right)^{i}\left(1-\frac{r}{r+1}\right)^{x_1+x_2-i} \le \alpha. \tag{14.5}$$

Test H_0 against $H_1^* : \gamma \neq r$ is usually carried out so that it is found out if at least one of inequalities (14.4) and (14.5) holds when α is substituted by the number $\alpha/2$. There exist also other variants of the two-sided test.

Example 14.3 We introduce an example which is presented in the book Hátle, Likeš (1972), p. 334. It is known that the number of failures of an equipment during 100 hours of operation is a random variable with Poisson distribution. Equipment A had 8 failures and equipment B 5 failures. If A $\sim$ Po(λ_1) and B $\sim$ Po(λ_2) and failures occur independently, we want to test $H_0 : \lambda = 1$ against $H_1 : \lambda > 1$. This can be done using program

```
library(stats)
poisson.test(c(8,5), r=1, alternative="greater")
```

```
       Comparison of Poisson rates
data:  c(8, 5)
time base: 1
count1 = 8, expected count1 = 6.5, p-value = 0.2905
alternative hypothesis: true rate ratio is greater than 1
95 percent confidence interval:
 0.5499053       Inf
sample estimates: rate ratio
       1.6
```

Right hand side confidence interval for λ is $(0.5499, \infty)$, p-value is 0.2905. Hypothesis $H_0 : \lambda = 1$ is not rejected.

Another program can be based on function `poisson.exact` in library `exactci`. One-sided tests in both cases give the same result, but the function `poisson.exact` in two-sided test presents three different variants. $\diamondsuit$

Chapter 15

Problem of k samples

15.1 Linear model

This section is an introduction and it serves as theoretical background for methods of analysis of variance described later in this chapter. The results will be also used in the chapter devoted to regression analysis.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be a random vector and $\boldsymbol{X}_{n\times k}$ a matrix of given numbers. Assume that $\boldsymbol{Y}$ follows *linear model*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{15.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown parameters and $\boldsymbol{e} = (e_1, \ldots, e_n)'$ is a vector of random variables satisfying conditions

$$\mathsf{E}\boldsymbol{e} = \boldsymbol{0}, \qquad \mathsf{var}\,\boldsymbol{e} = \sigma^2\boldsymbol{I}.$$

The parameter $\sigma^2 > 0$ is also not known. However, the vector $\boldsymbol{Y}$ is observable. It can be characterized as *vector of errors*. Under the term "errors" they are understood errors which follow from inaccuracies when the vector $\boldsymbol{Y}$ is measured, sometimes the errors include deviations from the exact linear dependence $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$. The vector $\boldsymbol{e}$ is not observable. The assumption $\mathsf{E}\boldsymbol{e} = \boldsymbol{0}$ reflect the fact that observations of vector $\boldsymbol{Y}$ do not include systematic errors. The relation $\mathsf{var}\,\boldsymbol{e} = \sigma^2\boldsymbol{I}$ says that the measurements of different components of the vector $\boldsymbol{Y}$ are made with equal precision, and the errors of measurements of different components of the vector $\boldsymbol{Y}$ are uncorrelated.

For simplicity we assume in this section that the rank of the matrix $\boldsymbol{X}$ is k and that $n > k$. Linear model (15.1) which fulfills these additional assumptions is called *regression model*.

It follows from our assumptions that $\mathsf{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$, $\mathsf{var}\,\boldsymbol{Y} = \sigma^2\boldsymbol{I}$. Vector $\boldsymbol{\beta}$ is estimated using *least squares method*, i.e. from the condition that the expression

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

is minimal. This estimator is denoted by $\boldsymbol{b} = (b_1, \ldots, b_k)'$.

It is known that for every matrix $\boldsymbol{A}$ it holds $h(\boldsymbol{A}) = h(\boldsymbol{A}\ \boldsymbol{A}') = h(\boldsymbol{A}'\boldsymbol{A})$, where $h(\boldsymbol{A})$ denotes the rank of he matrix $\boldsymbol{A}$. Since $h(\boldsymbol{X}) = k$ and the matrix $\boldsymbol{X}'\boldsymbol{X}$ is of type $k \times k$, we conclude that the matrix $\boldsymbol{X}'\boldsymbol{X}$ is regular.

Theorem 15.1 *It holds* $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

Proof. First, we verify that the vector $\boldsymbol{b}$ introduced in assertion of the theorem fulfills the condition $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}$. Then we write $S(\boldsymbol{\beta})$ in the form

$$S(\boldsymbol{\beta}) = [(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})]'[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{X}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta})].$$

Using the condition $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}$ we get

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + (\boldsymbol{b} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta}).$$

It follows from the assumption $h(\boldsymbol{X}) = k$ that the matrix $\boldsymbol{X}'\boldsymbol{X}$ is positive definite. Thus

$$(\boldsymbol{b} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{\beta}) \geq 0$$

and equality holds if and only if $\boldsymbol{\beta} = \boldsymbol{b}$. □

The vector $\boldsymbol{b}$ can be calculated from the *system of normal equations* $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$. The expression

$$R = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$$

is called *residual sum of squares* and it plays an important role in statistical tests. Quite often instead of R the denotation S_e is used. We shall also write $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}$.

Theorem 15.2 *The estimator* $\boldsymbol{b}$ *satisfies* $\mathsf{E}\boldsymbol{b} = \boldsymbol{\beta}$, $\mathsf{var}\,\boldsymbol{b} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Proof. We have

$$\begin{aligned}\mathsf{E}\boldsymbol{b} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},\\ \mathsf{var}\,\boldsymbol{b} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$ □

Theorem 15.3 *It holds*

$$R = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{Y}, \qquad R = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}.$$

Proof. Denote $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. We can write

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b} = \boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{Y}.$$

The matrix $\boldsymbol{M}$ is symmetric and idempotent, so that $\boldsymbol{M}^2 = \boldsymbol{M}$. Thus

$$R = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{Y}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}.$$

The first formula is proved. The second formula follows from the first one. □

Further we shall use the following auxiliary theorem.

Theorem 15.4 *Let the random vector* $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ *have finite second moments and let* $\mathsf{E}\boldsymbol{Z} = \boldsymbol{\mu}$, $\mathsf{var}\,\boldsymbol{Z} = \boldsymbol{V}$. *Then for arbitrary matrix* $\boldsymbol{A}_{n\times n}$ *we have* $\mathsf{E}\,\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} = \mathsf{Tr}\,\boldsymbol{A}\boldsymbol{V} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}$.

Proof. Since $\mathsf{var}\,\boldsymbol{Z} = \mathsf{E}(\boldsymbol{Z} - \boldsymbol{\mu})(\boldsymbol{Z} - \boldsymbol{\mu})' = \mathsf{E}\boldsymbol{Z}\boldsymbol{Z}' - \boldsymbol{\mu}\boldsymbol{\mu}'$, we obtain

$$\mathsf{E}\,\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} = \mathsf{E}\,\mathsf{Tr}\,\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}' = \mathsf{Tr}\,\boldsymbol{A}\mathsf{E}\,\boldsymbol{Z}\boldsymbol{Z}' = \mathsf{Tr}\,\boldsymbol{A}(\boldsymbol{V} + \boldsymbol{\mu}\boldsymbol{\mu}') = \mathsf{Tr}\,\boldsymbol{A}\boldsymbol{V} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$ □

Theorem 15.5 *Denote $s^2 = R/(n-k)$. Then $\mathsf{E}s^2 = \sigma^2$.*

Proof. We use theorems 15.3 and 15.4. Since $\mathsf{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$, we obtain

$$\mathsf{E}R = \mathsf{E}\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} = \mathsf{Tr}\,\boldsymbol{M}\sigma^2\boldsymbol{I} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} = \sigma^2\mathsf{Tr}\,\boldsymbol{M} = (n-k)\sigma^2. \qquad \square$$

Variable s^2 is called *residual variance*. We proved that s^2 is *unbiased estimator* of the parameter σ^2.

Now, we shall also assume that the vector $\boldsymbol{e}$ is normally distributed. This implies immediately that $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

Theorem 15.6 *Under assumption of normal distribution we have*

(a) $\boldsymbol{b} \sim \mathsf{N}[\boldsymbol{\beta},\, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}]$,

(b) $\dfrac{R}{\sigma^2} \sim \chi^2_{n-k}$,

(c) $\boldsymbol{b}$ *and* R *are independent.*

Proof. Assertion (a) follows from theorem 15.1 because linear transformation of normally distributed vector is again a vector with normal distribution.

Now, we prove (b). It is known that rank of an idempotent matrix is equal to its trace. The trace of matrix $\boldsymbol{A}$ will be denoted by $\mathsf{Tr}\,\boldsymbol{A}$. For arbitrary matrices $\boldsymbol{K}$ a $\boldsymbol{L}$, such that their product is a square matrix we have $\mathsf{Tr}\,\boldsymbol{K}\boldsymbol{L} = \mathsf{Tr}\,\boldsymbol{L}\boldsymbol{K}$. Define $\boldsymbol{\xi} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})/\sigma$. We insert for $\boldsymbol{b}$ and obtain $\boldsymbol{\xi} = \sigma^{-1}\boldsymbol{M}\boldsymbol{Y}$, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'$. Thus $\boldsymbol{\xi} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{M})$. Matrix $\boldsymbol{M}$ is idempotent and its rank equals to

$$\begin{aligned} h(\boldsymbol{M}) &= \mathsf{Tr}\,[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'] = \mathsf{Tr}\,\boldsymbol{I} - \mathsf{Tr}\,\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ &= n - \mathsf{Tr}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = n - k. \end{aligned}$$

Since $R/\sigma^2 = \boldsymbol{\xi}'\boldsymbol{\xi}$, property (b) follows from theorem 3.8 on p. 39.

Finally, we calculate

$$\mathsf{cov}(\boldsymbol{b}, \boldsymbol{\xi}) = \mathsf{cov}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \sigma^{-1}\boldsymbol{M}\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\sigma^{-1}\boldsymbol{M} = \boldsymbol{0}.$$

Vectors $\boldsymbol{b}$ and $\boldsymbol{\xi}$ are uncorrelated. Since

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\xi} \end{pmatrix} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\ \sigma^{-1}\boldsymbol{M} \end{bmatrix} \boldsymbol{Y},$$

they have simultaneous normal distribution. Thus $\boldsymbol{b}$ and $\sigma^2\boldsymbol{\xi}'\boldsymbol{\xi} = R$ are also independent. $\square$

Theorem 15.7 *Let v_{ij} be (i,j)-th element of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then for every $i = 1, \ldots, k$ the random variable*

$$T_i = \frac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}}$$

has the distribution t_{n-k}.

Proof. It follows from theorem 15.6 (a) that $b_i \sim \mathsf{N}(\beta_i, \sigma^2 v_{ii})$. Define

$$U_i = \frac{b_i - \beta_i}{\sigma\sqrt{v_{ii}}}.$$

We can see that $U_i \sim \mathsf{N}(0,1)$. Theorem 15.6 (b) gives that the variable $Z = (n-k)s^2/\sigma^2$ has distribution χ^2_{n-k}. Finally, assertion (c) from theorem 15.6 ensures independence U_i and Z. We know from theorem 3.10 that

$$T_i = \frac{U_i}{\sqrt{Z/(n-k)}} \sim \mathsf{t}_{n-k}. \qquad \square$$

Theorem 15.8 *Let $\boldsymbol{c} = (c_1, \ldots, c_k)' \neq \boldsymbol{0}$ be a given vector. Then $\mathsf{E}\boldsymbol{c}'\boldsymbol{b} = \boldsymbol{c}'\boldsymbol{\beta}$ and*

$$T = \frac{\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathsf{t}_{n-k}.$$

Proof. The unbiasednes of the estimator $\boldsymbol{c}'\boldsymbol{b}$ is obvious. From theorem 15.2 we obtain

$$\mathsf{var}\,\boldsymbol{c}'\boldsymbol{b} = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}.$$

Thus

$$U = \frac{\boldsymbol{c}'\boldsymbol{b} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathsf{N}(0,1).$$

Since U depends only on $\boldsymbol{b}$, it does not depend on $Z = (n-k)s^2/\sigma^2$. It is easy to see that

$$T = \frac{U}{\sqrt{Z/(n-k)}}.$$

Theorem 3.10 ensures that $T \sim \mathsf{t}_{n-k}$. $\square$

Let us remark that theorem 15.7 is special case of theorem 15.8. It suffices to choose vector $\boldsymbol{c}$ in such a way that its i-th component equals to 1 and the remaining components are all 0.

Vector $\boldsymbol{u} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is called *vector of residuals*. It is easy to check that

$$\boldsymbol{u} = \boldsymbol{M}\boldsymbol{Y}, \qquad \mathsf{E}\,\boldsymbol{u} = \boldsymbol{0}, \qquad \mathsf{var}\,\boldsymbol{u} = \sigma^2\boldsymbol{M}.$$

Let m_{ij} be elements of the matrix $\boldsymbol{M}$ and h_{ij} elements of the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. We see that the matrix $\mathsf{var}\,\boldsymbol{e}$ is diagonal with equal elements on the diagonal, the matrix $\mathsf{var}\,\boldsymbol{u}$ has not this property. From this reason *normed residuals*

$$v_t = \frac{u_t}{s\sqrt{m_{tt}}} = \frac{u_t}{s\sqrt{1-h_{tt}}}$$

are used. The observation Y_i is influential if its small change leads to rather large change of vector $\hat{\boldsymbol{Y}}$. The influential observations can be detected by diagonal elements h_{ii} of the matrix $\boldsymbol{H}$. The elements h_{ii} are called *leverages*, i.e. *influences*. It can be proved that $\sum h_{ii} = k$, $h_{ii} \geq 1/n$ for every i. If $h_{ii} > 2k/n$ then such observation is influential. The influence of individual observations is measured using *Cook statistics*

$$D_i = \frac{(\boldsymbol{b} - \boldsymbol{b}_{(i)})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{b} - \boldsymbol{b}_{(i)})}{ks^2},$$

where $\boldsymbol{b}_{(i)}$ is estimator of the vector $\boldsymbol{\beta}$ when the i-th observation is excluded.

15.2 Weighted average

In some cases it is necessary to apply the least squares method, although the observations have not the same variances. A model for such situation we investigate in this section.

Theorem 15.9 *Let $Y_1, \dots, Y_n$ be independent random variables. Assume that $\mathsf{E}Y_i = \beta$, $\mathsf{var}\, Y_i = \sigma_i^2 > 0$, where β is an unknown parameter and $\sigma_1^2, \dots, \sigma_n^2$ are known numbers. Then the estimator b of the parameter β by the least squares method is*

$$b = \frac{\sum_{j=1}^{n} \sigma_j^{-2} Y_j}{\sum_{i=1}^{n} \sigma_i^{-2}} \tag{15.2}$$

and it holds

$$\mathsf{E}b = \beta, \qquad \mathsf{var}\, b = \left(\sum_{i=1}^{n} \sigma_i^{-2} \right)^{-1}.$$

Proof. Variables Y_i correspond to the model

$$Y_i = \beta + e_i, \qquad i = 1, \dots, n, \tag{15.3}$$

where e_i are independent variables with moments $\mathsf{E}e_i = 0$, $\mathsf{var}\, e_i = \sigma_i^2$. Variables e_i have not equal variances so that it is not possible to use results derived in section 15.1. If we multiply (15.3) by expression σ_i^{-1}, we get

$$Z_i = \sigma_i^{-1}\beta + e_i^*, \qquad i = 1, \dots, n, \tag{15.4}$$

where $Z_i = \sigma_i^{-1} Y_i$ and $e_i^* = \sigma_i^{-1} e_i$. Now, the variables e_i^* are not only independent with vanishing expectations, but they have the same variances $\mathsf{var}\, e_i^* = 1$. Thus we can write (15.4) in the form $\boldsymbol{Z} = \boldsymbol{X}\beta + \boldsymbol{e}^*$, where

$$\boldsymbol{Z} = (Z_1, \dots, Z_n)', \qquad \boldsymbol{X} = (\sigma_1^{-1}, \dots, \sigma_n^{-1})', \qquad \boldsymbol{e}^* = (e_1^*, \dots, e_n^*)'.$$

From theorem 15.1 we get for arbitrary β least squares estimator

$$b = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z} = \frac{\sum_{j=1}^{n} \sigma_j^{-2} Y_j}{\sum_{i=1}^{n} \sigma_i^{-2}}.$$

Unbiasednes of the estimator b and the formula for $\mathsf{var}\, b$ follow from theorem 15.2. □

It can be proved that b is the *best linear unbiased estimator* —- BLUE — of the parameter β.

Theorem 15.10 *Let the variables Y_i fulfil assumptions of theorem 15.9 and, moreover, have normal distribution. Then the estimator b is also normally distributed and the random variable*

$$Q = \sum_{i=1}^{n} \frac{(Y_i - b)^2}{\sigma_i^2}$$

has χ^2_{n-1} distribution.

Proof. It is obvious that b has normal distribution, since b in view of (15.2) is a linear function of normally distributed vector $\boldsymbol{Y}$. Further we get

$$\begin{aligned} Q &= \sum_i \sigma_i^{-2} Y_i^2 - \left(\sum_i \sigma_i^{-2}\right)^{-1} \sum_j \sigma_j^{-2} Y_j \sum_k \sigma_k^{-2} Y_k \\ &= \sum_i Z_i^2 - \left(\sum_i \sigma_i^{-2}\right)^{-1} \sum_j \sigma_j^{-1} Z_j \sum_k \sigma_k^{-1} Z_k, \end{aligned} \tag{15.5}$$

where $Z_i = \sigma_i^{-1} Y_i$. We can see that $Q = \boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z}$, with

$$\boldsymbol{A} = \boldsymbol{I} - \left(\sum_i \sigma_i^{-2}\right)^{-1} \begin{pmatrix} \sigma_1^{-1} \\ \vdots \\ \sigma_n^{-1} \end{pmatrix} \left(\sigma_1^{-1}, \ldots, \sigma_n^{-1}\right).$$

Matrix $\boldsymbol{A}$ is idempotent and its rank (which is equal to its trace) equals to $n-1$. Expectation of the vector $\boldsymbol{Z}$ is

$$\mathsf{E}\boldsymbol{Z} = \left(\sigma_1^{-1}, \ldots, \sigma_n^{-1}\right)' \beta.$$

Vector $\boldsymbol{Z} - \mathsf{E}\boldsymbol{Z}$ has distribution $\mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$. It can be checked that

$$Q = \boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} = (\boldsymbol{Z} - \mathsf{E}\boldsymbol{Z})'\boldsymbol{A}(\boldsymbol{Z} - \mathsf{E}\boldsymbol{Z}).$$

Theorem 3.9 implies that $Q \sim \chi^2_{n-1}$. □

Theorem 15.10 can be used for testing hypothesis H_0, that all variables in model (15.3) have equal expectation. If H_0 does not hold, then Q is large. Thus H_0 is rejected when $Q \geq \chi^2_{n-1}(\alpha)$.

15.3 Extrapolation in linear model

Assume that the model (15.1) holds. We expect that a new observation Y_0 will be realized, which will correspond to a new row of the matrix $\boldsymbol{X}$, say $\boldsymbol{x}_0' = (x_{01}, \ldots, x_{0k})$. The new variable Y_0 follows the model $Y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + e_0$, where e_0 is a random variable with vanishing mean and with variance σ^2, which is independent of the vector of errors $\boldsymbol{e}$. We want to estimate the variable Y_0, i.e. to calculate its *extrapolation*. We intend to calculate a point estimator and an interval estimator. Obviously, $\hat{Y}_0 = \boldsymbol{x}_0'\boldsymbol{b}$ can serve as the point estimator. The interval estimator will be calculated under assumption of normality. Thus let $\boldsymbol{Y}$ be normal and $e_0 \sim \mathsf{N}(0, \sigma^2)$. Then $Y_0 \sim \mathsf{N}(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2)$. We are interested in the difference

$$\hat{Y}_0 - Y_0 = \boldsymbol{x}_0'\boldsymbol{b} - \boldsymbol{x}_0'\boldsymbol{\beta} - e_0 = \boldsymbol{x}_0'(\boldsymbol{b} - \boldsymbol{\beta}) - e_0.$$

Calculate its expectation and variance. We have

$$\begin{aligned} \mathsf{E}(\hat{Y}_0 - Y_0) &= \mathsf{E}[\boldsymbol{x}_0'(\boldsymbol{b} - \boldsymbol{\beta}) - e_0] = 0, \\ \mathsf{var}(\hat{Y}_0 - Y_0) &= \mathsf{var}[\boldsymbol{x}_0'(\boldsymbol{b} - \boldsymbol{\beta}) - e_0] = \mathsf{var}\,\boldsymbol{x}_0'(\boldsymbol{b} - \boldsymbol{\beta}) - 2\mathsf{cov}[\boldsymbol{x}_0'(\boldsymbol{b} - \boldsymbol{\beta}), e_0] + \mathsf{var}\,e_0. \end{aligned}$$

Since $\boldsymbol{b}$ and e_0 are independent, we have $\mathsf{cov}[\boldsymbol{x}_0'(\boldsymbol{b}-\boldsymbol{\beta}), e_0] = 0$. This gives

$$\mathsf{var}(\hat{Y}_0 - Y_0) = \boldsymbol{x}_0'(\mathsf{var}\,\boldsymbol{b})\boldsymbol{x}_0 + \sigma^2 = \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \sigma^2.$$

Denote

$$\xi = \frac{\hat{Y}_0 - Y_0}{\sigma\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1}}, \qquad \eta = \frac{(n-k)s^2}{\sigma^2}.$$

It was proved that $\xi \sim \mathsf{N}(0,1)$, $\eta \sim \chi^2_{n-k}$ and that ξ and η are independent. Thus

$$T = \frac{\xi}{\sqrt{\eta/(n-k)}} \sim \mathsf{t}_{n-k}.$$

This can be written in the form

$$T = \frac{\hat{Y}_0 - Y_0}{s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1}}$$

and it follows from formula $\mathsf{P}\{|T| \le t_{n-k}(\alpha)\} = 1-\alpha$ that

$$\mathsf{P}\{\hat{Y}_0 - t_{n-k}(\alpha)s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1} \le Y_0 \le \hat{Y}_0 + t_{n-k}(\alpha)s\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1}\} = 1-\alpha. \tag{15.6}$$

Thus we derived confidence interval for Y_0 with confidence coefficient $1-\alpha$.

15.4 Submodel of the linear model

We say that the linear model

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{e} \tag{15.7}$$

is *submodel* of the model (15.1), if the columns of the matrix $\boldsymbol{U}_{n\times l}$ are linear combinations of the columns of the matrix $\boldsymbol{X}$ and in the same time $h(\boldsymbol{U}) = l < k$. Also here we shall deal only with such matrices $\boldsymbol{U}$, which have full column rank. It happens quite often that the submodel (15.7) arises from the model (15.1) by the omitting some columns of the matrix $\boldsymbol{X}$. It corresponds to the procedure where we define the corresponding components of the parameter $\boldsymbol{\beta}$ as zero. We show generally that linear bindings among the components of the vector $\boldsymbol{\beta}$ always lead to the submodel (15.7).

Theorem 15.11 *Let the components of the vector $\boldsymbol{\beta}$ fulfil the condition $\boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{0}$, where $h(\boldsymbol{G}_{t\times k}) = t < k$. Define $l = k - t$. Then there exists such a matrix $\boldsymbol{Q}_{l\times k}$ of rank l and such a matrix $\boldsymbol{U}_{n\times l}$ of rank l, that the columns of the matrix $\boldsymbol{U}$ are linear combinations of the columns of the matrix $\boldsymbol{X}$ and it holds $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{U}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \boldsymbol{Q}\boldsymbol{\beta}$.*

Proof. Since $h(\boldsymbol{G}_{t\times k}) = t$, there exists such a matrix $\boldsymbol{Q}_{l\times k}$ that the matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{Q} \\ \boldsymbol{G} \end{pmatrix}$$

is regular. We shall write the matrix $\boldsymbol{X}\boldsymbol{A}^{-1}$ in the form $\boldsymbol{X}\boldsymbol{A}^{-1} = (\boldsymbol{U}, \boldsymbol{T})$, where $\boldsymbol{U}$ is of type $n \times l$ and $\boldsymbol{T}$ is of type $n \times t$. We have

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{A}^{-1}\begin{pmatrix} \boldsymbol{Q}\boldsymbol{\beta} \\ \boldsymbol{G}\boldsymbol{\beta} \end{pmatrix} = (\boldsymbol{U}, \boldsymbol{T})\begin{pmatrix} \boldsymbol{Q}\boldsymbol{\beta} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{U}\boldsymbol{\gamma}. \qquad \square$$

It is important to realize that if the submodel (15.7) holds, then the model (15.1) holds, too. The estimator of the vector $\boldsymbol{\gamma}$ using the least squares method is

$$\boldsymbol{g} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Y}$$

and the residual variance R_1 in case of the submodel (15.7) is

$$R_1 = (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{g})'(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{g}) = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}']\boldsymbol{Y}.$$

At the same time it holds $R_1 \geq R$, since R is the minimum of the function $S(\boldsymbol{\beta})$ without any restriction on the vector $\boldsymbol{\beta}$, whereas R_1 is minimum of the function $S(\boldsymbol{\beta})$ under some subsidiary conditions (e.g. under condition $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{0}$). Test of the hypothesis that the vector $\boldsymbol{\beta}$ fufills the subsidiary conditions, is the test of hypothesis that the model (15.1) can be reduced to the submodel (15.7). The test is based on the following theorem.

Theorem 15.12 *If the submodel (15.7) holds (i.e. when the vector $\boldsymbol{\beta}$ is tied by $t = k - l$ linearly independent relations) then the random variable*

$$F = \frac{(n-k)(R_1 - R)}{tR}$$

has $\mathsf{F}_{t,n-k}$ distribution.

Proof. Again, denote $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{\xi} = \sigma^{-1}\boldsymbol{M}\boldsymbol{Y}$. In the proof of theorem 15.6 (c) it was derived that $\sigma^2\boldsymbol{\xi}'\boldsymbol{\xi} = R$, and it follows from theorem 15.6 (b) that $\boldsymbol{\xi}'\boldsymbol{\xi} \sim \chi^2_{n-k}$.

Denote $\boldsymbol{B} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'$. Then $R_1 - R = \boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}$. Let $\boldsymbol{\eta} = \sigma^{-1}\boldsymbol{B}\boldsymbol{Y}$. Since the columns of the matrix $\boldsymbol{U}$ are linear combinations of the columns of matrix $\boldsymbol{X}$, there exists a matrix $\boldsymbol{K}_{k\times l}$ such that $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{K}$. Matrix $\boldsymbol{B}$ is symmetric and idempotent. If (15.7) holds, we have $\boldsymbol{\eta} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{B})$. The rank of the matrix $\boldsymbol{B}$ is equal to its trace $k - l = t$. The trace can be derived similarly as in the proof of theorem 15.6 (b). Theorem 3.8 on p. 39 implies that $\boldsymbol{\eta}'\boldsymbol{\eta} \sim \chi^2_t$. We have

$$\mathsf{cov}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \mathsf{cov}(\sigma^{-1}\boldsymbol{M}\boldsymbol{Y}, \sigma^{-1}\boldsymbol{B}\boldsymbol{Y}) = \sigma^{-2}\boldsymbol{M}\sigma^2\boldsymbol{I}\boldsymbol{B} = \boldsymbol{M}\boldsymbol{B} = \boldsymbol{0}.$$

Thus the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are uncorrelated. Because thy are normally distributed, they are independent. Then the variables $\boldsymbol{\eta}'\boldsymbol{\eta}$ a $\boldsymbol{\xi}'\boldsymbol{\xi}$ are also independent. Using theorem 3.11 on p. 40 we obtain

$$\frac{\boldsymbol{\eta}'\boldsymbol{\eta}/t}{\boldsymbol{\xi}'\boldsymbol{\xi}/(n-k)} \sim \mathsf{F}_{t,n-k}. \qquad \square$$

15.5 One-way analysis of variance

The problem is similar as in the case of two-sample t test, but now it is more complicated because we have more samples. Assume that

$$Y_{11}, \quad \ldots, \quad Y_{1n_1} \qquad \text{is the sample from } \mathsf{N}(\mu_1, \sigma^2)$$

etc. and

$$Y_{I1}, \quad \ldots, \quad Y_{In_I} \qquad \text{is the sample from } \mathsf{N}(\mu_I, \sigma^2).$$

Let the samples be independent. Such a model is called *one-way analysis of variance*. We want to test hypothesis $H_0 : \mu_1 = \cdots = \mu_I$ against the alternative H_1, that some of the means $\mu_1, \ldots, \mu_I$ are different. Let us remark that if H_0 is rejected it is necessary to find all pairs μ_i, μ_j, which caused this rejection, i.e. for which it can be statistically proved that $\mu_i \neq \mu_j$.

Seemingly it could be done so that all two-sample t tests are calculated. However, the number of such tests is $I(I-1)/2$. If each of them has the level α, the level of the complete test would be substantially larger than α. The level of each test could be smaller, say $2\alpha/[I(I-1)]$, which would ensure that the overall level would not be larger than α. The power of such test would nod be large enough and its real level would be considerably smaller than α. Thus another procedure is chosen. We form one vector $\boldsymbol{Y}$ from all variables Y_{ij} so that first we give all members of the first sample, then of the second and finally the last. Let

$$n = n_1 + \cdots + n_I.$$

Our assumptions lead to the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{15.8}$$

where

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{matrix} \} \, n_1 \\ \\ \} \, n_2 \\ \\ \} \, n_I \end{matrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}. \tag{15.9}$$

The vector $\boldsymbol{e}$ has n components and it holds $\boldsymbol{e} \sim \mathsf{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Denote

$$Y_{i.} = Y_{i1} + \cdots + Y_{in_i}, \qquad y_{i.} = \frac{Y_{i.}}{n_i} \quad \text{for} \quad i = 1, \ldots, I$$

and

$$Y_{..} = Y_{1.} + \cdots + Y_{I.} = \sum\sum Y_{ij}, \qquad \bar{Y} = y_{..} = \frac{Y_{..}}{n}.$$

Since

$$\boldsymbol{X}'\boldsymbol{X} = \mathsf{Diag}\{n_1, \ldots, n_I\}, \qquad \boldsymbol{X}'\boldsymbol{Y} = (Y_{1.}, \ldots, Y_{I.})',$$

we have

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (y_{1.}, \ldots, y_{I.})'.$$

Theorem 15.3 gives that the residual sum of squares is

$$R = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y} = \sum_i \sum_j Y_{ij}^2 - \sum_i y_{i.}Y_{i.} = \sum_i \sum_j Y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i}.$$

If H_0 holds, we have the submodel

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{e},$$

where $\boldsymbol{U}_{n\times 1} = (1, \ldots, 1)'$ and $\boldsymbol{\gamma}$ is of type 1×1. This time $\boldsymbol{U}'\boldsymbol{U} = n$, $\boldsymbol{U}'\boldsymbol{Y} = Y_{..}$, and the estimator $\boldsymbol{g}$ of the parameter $\boldsymbol{\gamma}$ using method of least squares is

$$\boldsymbol{g} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{Y} = y_{..} \ .$$

Residual sum of squares is

$$R_1 = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{g}'\boldsymbol{U}'\boldsymbol{Y} = \sum_i \sum_j Y_{ij}^2 - y_{..}Y_{..} = \sum_i \sum_j Y_{ij}^2 - \frac{Y_{..}^2}{n}.$$

Thus

$$R_1 - R = \sum_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n}.$$

The rank of the matrix $\boldsymbol{X}$ is I, the rank of the matrix $\boldsymbol{U}$ is 1. If H_0 holds, we get from theorem 15.12, that

$$F_A = \frac{(n-I)(R_1 - R)}{(I-1)R} \sim \mathsf{F}_{I-1,n-I}\,.$$

In practical calculations the *total sum of squares* is calculated first

$$S_T = \sum_i \sum_j Y_{ij}^2 - \frac{Y_{..}^2}{n}.$$

Then the sum of squares $R_1 - R$ is calculated. It is

$$S_A = \sum_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n}.$$

Residual sum of squares is usually denoted S_e instead of R and it is calculated from the formula

$$S_e = S_T - S_A.$$

The variable $s^2 = S_e/(n-I)$ is *residual variance*.

The results are written in the *table of analysis of variance* (see tab. 15.1). Sum of squares is denoted SS, degrees of freedom are *df* and the ratio (mean square) is MS.

In the case that $F_A \geq F_{I-1,n-I}(\alpha)$ we reject the hypothesis H_0. Then it is necessary to decide which pairs of indices satisfy $\mu_i \neq \mu_j$. Since $y_{i.}$ is an estimator of μ_i, the table of differences $y_{i.} - y_{j.}$ is calculated (see tab. 15.2).

Here we introduce only *Tukey method of multiple comparisons*, since it is one of the most sensitive.

Table 15.1: One-way analysis of variance

Source	Sum squares SS	Degrees of freedom df	MS $MS = \frac{SS}{df}$	Test statistic $F = \frac{MS}{s^2}$
Groups	S_A	$f_A = I - 1$	S_A/f_A	F_A
Residual	S_e	$f_e = n - I$	$s^2 = S_e/f_e$	–
Total	S_T	$f_T = n - 1$	–	–

Table 15.2: Differences of means

	j			
i	2	3	...	I
1	$y_{1.} - y_{2.}$	$y_{1.} - y_{3.}$	$\cdots$	$y_{1.} - y_{I.}$
2		$y_{2.} - y_{3.}$	$\cdots$	$y_{2.} - y_{I.}$
.....	.....			
$I-1$				$y_{I-1,.} - y_{I.}$

Lemma 15.13 *Let $X_1, \ldots, X_m$ be random sample from the distribution $\mathsf{N}(\mu, \sigma^2)$, where $\sigma > 0$. Denote $R = \max X_i - \min X_i$ the* range. *Let s^2 be an independent estimator for σ^2 with ν degrees of freedom; it means that $\nu s^2/\sigma^2 \sim \chi^2_\nu$ and that s^2 and $\boldsymbol{X} = (X_1, \ldots, X_m)'$ are independent. Denote $Q = R/s$ random variable called* studentized range. *Then the distribution of the random variable Q, which is denoted by symbol $q_{m,\nu}$, does not depend on μ and σ.*

Proof. It holds

$$Q = \frac{\max(X_i - \mu)/\sigma - \min(X_i - \mu)/\sigma}{s/\sigma}. \tag{15.10}$$

Variables $(X_i - \mu)/\sigma$ $(i = 1, \ldots, m)$ are independent with distribution $\mathsf{N}(0, 1)$. Thus the distribution of the expression in the numerator of formula (15.10) does not depend on μ and σ. The distribution of the variable s/σ does not depend on μ and σ, too. Since numerator and denominator of the formula (15.10) are independent variables, the distribution of the variable Q does not depend on μ and σ. □

We will not derive the density of the variable Q here. The critical value $q_{m,\nu}(\alpha)$ is defined as the number satisfying $\mathsf{P}\{Q > q_{m,\nu}(\alpha)\} = \alpha$.

Theorem 15.14 (Tukey) *Let $X_1, \ldots, X_m$ be independent random variables and let $X_i \sim \mathsf{N}(\mu_i, b^2\sigma^2)$, $i = 1, \ldots, m$, where b is a known positive constant. Let s^2 be independent estimator for σ^2 with ν degrees of freedom; it means that $\nu s^2/\sigma^2 \sim \chi^2_\nu$ and that s^2 and $\boldsymbol{X}$ are independent. Then probability that*

$$X_i - X_j - bsq_{m,\nu}(\alpha) \le \mu_i - \mu_j \le X_i - X_j + bsq_{m,\nu}(\alpha)$$

holds for all pairs (i, j) simultaneously, equals to $1 - \alpha$.

Proof. Denote $Z_i = X_i - \mu_i$. Variables Z_i are independent and each of them has distribution $\mathsf{N}(0, b^2\sigma^2)$. Vector $(Z_1, \dots, Z_m)'$ does not depend on s^2. It is clear that b^2s^2 is independent estimator for the variance $b^2\sigma^2$ with ν degrees of freedom, since $\nu b^2 s^2/b^2\sigma^2 \sim \chi^2_\nu$. It follows from lemma 15.13 that

$$\mathsf{P}\left\{\frac{\max Z_i - \min Z_i}{bs} \le q_{m,\nu}(\alpha)\right\} = 1 - \alpha.$$

This is equivalent to the formula

$$\mathsf{P}\{|Z_i - Z_j| \le bsq_{m,\nu}(\alpha) \quad \text{for all pairs } (i,j)\} = 1 - \alpha.$$

Inserting for Z_i and Z_j we get the assertion of the theorem. □

Tukey theorem is used in connection with testing hypothesis $H_0 : \mu_1 = \cdots = \mu_m$. If H_0 holds, then all intervals $[X_i - X_j - bsq_{m,\nu}(\alpha), X_i - X_j + bsq_{m,\nu}(\alpha)]$ overlap zero with probability $1 - \alpha$. If some of them does not cover zero, i.e. if

$$|X_{i_0} - X_{j_0}| > bsq_{m,\nu}(\alpha) \tag{15.11}$$

holds for a pair (or for some pairs) (i_0, j_0), then it leads to the conclusion that we have $\mu_{i_0} \neq \mu_{j_0}$. This procedure has an advantage that probability of the error of the first kind equals α and the rejection of H_0 shows directly to the pair or pairs responsible for this rejection.

Tukey method can be applied in the one-way analysis of variance if $y_{1.}, \dots, y_{I.}$ have equal variances. It comes in the case that $n_1 = \cdots = n_I$. If this condition is satisfied, we say, that the model is *balanced*. In such a case we denote size of each sample by P (i.e., $P = n_1 = \cdots = n_I$). Then $y_{i.} \sim \mathsf{N}(\mu_i, \sigma^2/P)$. In theorem 15.14 we have $b^2 = 1/P$. According to theorem 15.11 the equality of expectations of i-th and j-th sample is rejected, if

$$|y_{i.} - y_{j.}| > \frac{s}{\sqrt{P}} q_{I,n-I}(\alpha).$$

Tukey method can be modified to the case of unbalanced classification. It was proved (see Hayter 1984), that

$$\mathsf{P}\left\{|y_{i.} - y_{j.}| < sq_{I,n-I}(\alpha)\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \qquad \text{for all} \quad i,j\right\} \ge 1 - \alpha.$$

If we get

$$|y_{i.} - y_{j.}| > sq_{I,n-I}(\alpha)\sqrt{\frac{1}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

then we reject the hypothesis about equality $\mu_i = \mu_j$. This modification in programs is called *Tukey HSD*. The abbreviation HSD means *honest significant difference*.

In tab. 15.2 to each difference which is statistically significant on the level 0.05, a star as the upper index is added. If significance is on level 0.01, two stars are added and for significance on level 0.001 three stars. This holds also for other tests.

At the beginning of this section we introduced the model of one-way analysis of variance in the form (15.8) and (15.9). Writing it in components we get

$$Y_{ij} = \mu_i + e_{ij}, \qquad i = 1, \dots, I, \quad j = 1, \dots, n_i.$$

This form simplifies the derivation of the procedure, since the matrix $\boldsymbol{X}$ in (15.9) has full rank. However, usually this model is applied in the form

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$

New matrix $\boldsymbol{X}$, which now corresponds to parameters $\mu, \alpha_1, \ldots, \alpha_I$, is of type $n \times (I+1)$, but it has rank I. Null hypothesis was $\mu_1 = \cdots = \mu_I$, but now it has the form $\alpha_1 = 0, \ldots, \alpha_I = 0$. Both models are equivalent and lead to the same formulas. Only the methods of derivations are different.

Example 15.15 Four sorts of potatoes, namely A, B, C, and D, were cultivated. Each sort was cultivated on 7 fields of equal area. The yields in ton/hectare are introduced in tab. 15.3.

Table 15.3: Yields of potatoes in t/ha

Sort	Yield							Average
A	19.3	18.0	21.6	22.4	20.9	20.1	24.0	20.9
B	23.1	26.5	25.2	25.0	24.3	21.4	26.7	24.6
C	23.7	20.8	19.8	24.1	22.2	22.6	22.9	22.3
D	17.2	16.6	16.9	17.7	21.3	15.2	19.0	17.7

The results of analysis of variance are in tab. 15.4.

Table 15.4: Analysis of variance

Source	Sum of squares SS	Degrees of freedom df	Ratio $MS = \frac{SS}{df}$	Test statistic $F = \frac{MS}{s^2}$
Sorts	174.9125	3	58.3042	17.0148*
Residual	82.2400	24	3.4267	–
Total	257.1525	27	–	–

Since $F_A = 17.0148 \geq F_{3,24}(0.05) = 3.01$, we reject the hypothesis that the sorts do not influence the yield of potatoes. The differences among means are compared with the critical value

$$\sqrt{3.4267}\, q_{4,24}(0.05) \sqrt{\frac{1}{2}\left(\frac{1}{7} + \frac{1}{7}\right)} = 2.73.$$

We prepare table of differences of means (tab. 15.5).

Then the means are ordered and with the continuous line are underlined the groups of means the members of which are not significantly different. We get tab. 15.6

We can see that the yield of the sort D differs from the yield of all other sorts. Further it is proved that the yields of the sorts A and B are different. On the other side it was not proved that the yields of the sorts A and C and sorts C and B are different.

Program `R` gives

Table 15.5: Differences of means

	B	C	D
A	-3.7^*	-1.4	3.2^*
B		2.3	6.9^*
C			4.6^*

Table 15.6: Sorted means

D	A	C	B
17.7	20.9	22.3	24.6

```
vynos <- c(19.3, 18.0, 21.6, 22.4, 20.9, 20.1, 24.0,
           23.1, 26.5, 25.2, 25.0, 24.3, 21.4, 26.7,
           23.7, 20.8, 19.8, 24.1, 22.2, 22.6, 22.9,
           17.2, 16.6, 16.9, 17.7, 21.3, 15.2, 19.0)
odruda <- factor(c(rep("A",7), rep("B",7), rep("C",7), rep("D",7)))
tapply(vynos, odruda, mean) # calculation of averages
   A    B    C    D
20.9 24.6 22.3 17.7
bram.aov <- aov(vynos ~ odruda) # analysis of variance
summary(bram.aov)
            Df Sum Sq Mean Sq F value    Pr(>F)
odruda       3 174.91  58.304  17.015 3.872e-06 ***
Residuals   24  82.24   3.427
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(tk <- TukeyHSD(bram.aov))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = vynos ~ odruda) $odruda
    diff        lwr        upr     p adj
B-A  3.7  0.9704439  6.4295561 0.0052375
C-A  1.4 -1.3295561 4.1295561 0.5026381
D-A -3.2 -5.9295561 -0.4704439 0.0173519
C-B -2.3 -5.0295561  0.4295561 0.1204298
D-B -6.9 -9.6295561 -4.1704439 0.0000019
D-C -4.6 -7.3295561 -1.8704439 0.0005518
plot(tk) # figure for Tukey method
bartlett.test(vynos, odruda)
        Bartlett test of homogeneity of variances
data:  vynos and odruda
Bartlett's K-squared = 0.4537, df = 3, p-value = 0.9289
library(car)
leveneTest(vynos, odruda)
Levene's Test for Homogeneity of Variance (center = median)
```

```
      Df F value Pr(>F)
group  3  0.1209 0.9469
      24
leveneTest(vynos, odruda, center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  3   0.126 0.9438
      24
```

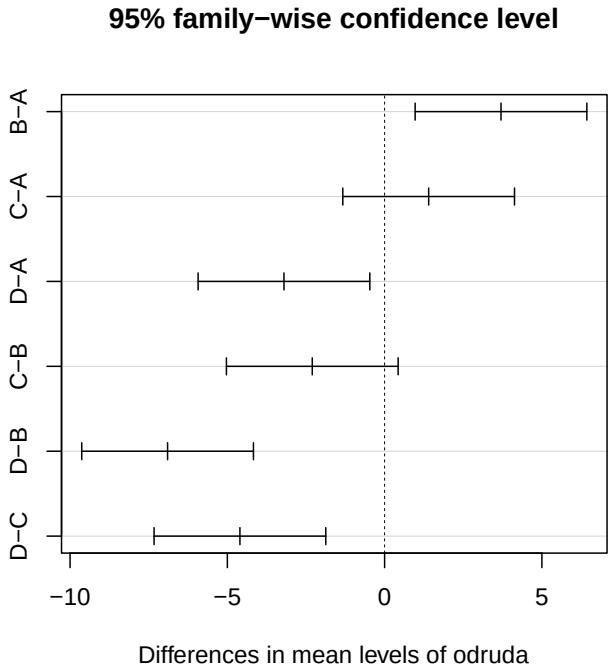

Figure 15.1: Tukey method

When applying Tukey method we look which interval does not contain zero. It corresponds to the significant difference. This is easy using the graph (see fig. 15.1). Also Bartlett test and Levene test were performed. These tests will be explained in the next section. ♢

15.6 Tests of homogeneity of variances

One of assumptions which we used in derivation of formulas for one-way analysis of variance was that variances of all I normal distributions are the same. If we admit that $Y_{i1}, \ldots, Y_{in_i}$ is the sample from the distribution $\mathsf{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, I$, the problem arises how to test the hypothesis

$$H_0 : \sigma_1^2 = \cdots = \sigma_I^2$$

against alternative H_1, that H_0 does not hold.

Several tests of the hypothesis H_0 has been derived. First, we describe *Bartlett test*. The denotation will be the same as in sec. 15.3.

Define

$$s_i^2 = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} Y_{ij}^2 - n_i y_{i.}^2\right),$$

$$s^2 = \frac{1}{n - I}\sum_{i=1}^{I}(n_i - 1)s_i^2,$$

$$C = 1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{n_i - 1} - \frac{1}{n - I}\right),$$

$$B = \frac{1}{C}\left[(n - I)\ln s^2 - \sum_{i=1}^{I}(n_i - 1)\ln s_i^2\right].$$

We have $s^2 = S_e/f_e$, where S_e and f_e are introduced in tab. 15.1. If H_0 holds, then the random variable B has approximately χ^2_{I-1} distribution. Thus H_0 is rejected in the case that $B \geq \chi^2_{I-1}(\alpha)$. However, this result can be used only in the case that the sizes $n_1, \ldots, n_I$ are sufficiently large. In textbooks is usually introduce that it must hold $n_i > 6$ for all $i = 1, \ldots, I$. Glaser (1976)published exact critical values. In the case $I = 2$ Bartlett test can be used as the test for homogeneity of two variances instead common F test. Manoukian a kol. (1986) published exact critical values for this case. Some other results about Bartlett test can be found in paper Nagarsenker (1984).

Unfortunately, Bartlett test is sensitive to violation of the assumption of normality (Box 1953, Box, Andersen 1955). The robust tests were suggested instead of it, see Layard 1973, Brown, Forsythe 1974.

Probably most often the *Levene test* (Levene 1960) is used for testing homogeneity of variances. The I groups of random variables introduced in rows of tab. 15.7 is created.

Table 15.7: Levene test

$\lvert Y_{11} - y_{1.}\rvert, \ldots, \lvert Y_{1n_1} - y_{1.}\rvert$
$\cdots\cdots\cdots\cdots\cdots\cdots$
$\lvert Y_{I1} - y_{I.}\rvert, \ldots, \lvert Y_{In_I} - y_{I.}\rvert$

Analysis of variance or *Kruskal-Wallis test* are applied to the variables in tab. 15.7. Nowadays in tab. 15.7 instead of means the medians are subtracted. Levene test can be found in the library `car` as the function `leveneTest`.

If we get a significant result, we reject the hypothesis about homogeneity of variances. This test is only approximate, since the assumptions of analysis of variance or its nonparametric variant are not fulfilled. The variables $Y_{i1} - y_{i.}, \ldots, Y_{in_i} - y_{i.}$ are not independent and their absolute values have not a normal distribution. In spite of this Levene test is one of the best. It is shown in the paper Conover et al. (1981).

Example 15.16 We use data from example 15.15. We get

$$s_1^2 = 3{,}9933, \quad s_2^2 = 3{,}5200, \quad s_3^2 = 2{,}3600, \quad s_4^2 = 3{,}8333,$$
$$s^2 = 3{,}4266, \quad C = 1{,}0694, \quad B = 0{,}4537.$$

Since $B < \chi^2_3(0{,}05) = 7{,}81$, Bartlett test does not reject hypothesis of homogeneity of variances in the fourth partial samples.

Levene test based on Kruskal-Wallis test gives test statistic $Q = 0.426$ with three degrees of freedom, which asymptotically corresponds to the level $p = 0.935$. Levene test based on analysis of variance when medians are subtracted gives F statistic 0.121 and the level of the test is $p = 0.947$. When the averages are subtracted, then we get $F = 0.126$ and p-value 0.9438. None of the tests leads to the rejection of the hypothesis about homogeneity of variances. $\diamondsuit$

15.7 Analysis when variances are not equal

Let $Y_{i1}, \dots, Y_{in_i}$ be a sample from $\mathsf{N}(\mu_i, \sigma_i^2)$, $i = 1, \dots, I$. Assume that the samples are independent. Let $n_i \geq 2$, $\sigma_i^2 > 0$. Denote

$$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \qquad f_i = n_i - 1, \qquad s_i^2 = \frac{1}{f_i} \sum_{j=1}^{n_i} (Y_{ij} - y_i)^2, \qquad \lambda_i = \frac{1}{n_i}.$$

We know that $y_i \sim \mathsf{N}(\mu_i, \lambda_i \sigma_i^2)$. We shall test the hypothesis $H_0 : \mu_1 = \cdots = \mu_I$. The common (but unknown) value of parameters $\mu_1, \dots, \mu_I$ will be denoted μ. We use the results derived in section about weighted mean (see sec. 15.2 on p. 133). If H_0 holds, then NNLO $\hat{\mu}$ of the parameter μ equals to

$$\hat{\mu} = \frac{1}{\sum_i \frac{1}{\lambda_i \sigma_i^2}} \sum_j \frac{1}{\lambda_j \sigma_j^2} y_j$$

and it holds

$$Z^* = \sum_i \frac{1}{\lambda_i \sigma_i^2} (y_i - \hat{\mu})^2 \sim \chi^2_{I-1}.$$

Usually, the variances $\sigma_1^2, \dots, \sigma_I^2$ are not known. Thus we substitute σ_i^2 by unbiased estimator s_i^2. Define

$$w_i = \frac{1}{\lambda_i s_i^2}, \qquad \hat{y} = \frac{1}{\sum_i w_i} \sum_j w_j y_j, \qquad Z = \sum_i w_i (y_i - \hat{y})^2.$$

Then Z can be used as the test statistic for testing H_0. James (1951) approximated the critical value of the variable Z by

$$h(w) = \chi^2_{I-1}(\alpha) \left[1 + \frac{3\chi^2_{I-1}(\alpha) + I + 1}{2(I^2 - 1)} \sum_i \frac{1}{f_i} \left(1 - \frac{w_i}{\sum_j w_j} \right)^2 \right].$$

Thus it holds (approximately)

$$\mathsf{P}\{Z > h(w)\} = \alpha.$$

Welch (1951) denoted

$$\hat{f}_1 = I - 1, \qquad \hat{f}_2 = \left[\frac{3}{I^2 - 1} \sum_i \frac{1}{f_i} \left(1 - \frac{w_i}{\sum_j w_j} \right)^2 \right]^{-1}$$

and proposed to use

$$v^2 = \frac{(I-1)^{-1}\sum_i w_i(y_i - \hat{y})^2}{1 + 2(I-2)(I^2-1)^{-1}\sum_i \frac{1}{f_i}\left(1 - \frac{w_i}{\sum_j w_j}\right)^2}$$

as the test statistic. He proved that when H_0 holds then the variable v^2 has approximately distribution $\mathsf{F}_{\hat{f}_1,\hat{f}_2}$. His method is used in program `R` in function `oneway.test`.

Example 15.17 We use data introduced in example 15.15. We get

```
oneway.test(vynos ~ odruda)
        One-way analysis of means (not assuming equal variances)
data:  vynos and odruda
F = 14.5298, num df = 3.000, denom df = 13.256, p-value = 0.0001764
```

It means that the difference is significant. ♢

15.8 Kruskal-Wallis test

Kruskal-Wallis test is a nonparametric analogue of one-way analysis of variance and it is a generalization of two-sample Wilcoxon test. It is used mainly in the situations when the samples are from very non-normal populations.

Let $Y_{i1}, \ldots, Y_{in_i}$ be a sample from a distribution with a continuous distribution function F_i, $i = 1, \ldots, I$. Let all the samples be independent. We are going to test the hypothesis

$$H_0 : F_1(x) = \cdots = F_I(x) \qquad \text{for all } x$$

against the alternative H_1, that H_0 does not hold. All variables Y_{ij} together form the *pooled sample* with size $n = n_1 + \cdots + n_I$. They are ordered into increasing sequence and the rank R_{ij} of each variable Y_{ij} in the pooled sample is determined. This rank can be written in tab. 15.8.

Table 15.8: Kruskal-Wallis test

Sample	Rank of variables in pooled random sample				Sum of ranks
1	R_{11}	R_{12}	...	R_{1n_1}	T_1
2	R_{21}	R_{22}	...	R_{2n_2}	T_2
.....	.....				.....
I	R_{I1}	R_{I2}	...	R_{In_I}	T_I

Proposed test statistic is

$$Q^* = \sum_{i=1}^{I} \frac{1}{n_i}(T_i - \mathsf{E}T_i)^2.$$

Since

$$\mathsf{E}R_{ij} = \frac{1}{N}\sum_{k=1}^{N} k = \frac{N+1}{2},$$

we have

$$\mathsf{E}T_i = \mathsf{E}\sum_{j=1}^{n_i} R_{ij} = n_i\frac{N+1}{2}.$$

From this formula we obtain

$$Q^* = \sum_{i=1}^{I}\frac{1}{n_i}T_i^2 - \frac{1}{4}N(N+1)^2.$$

Define $\sigma_a^2 = \frac{1}{12}(N+1)(N-1)$. Instead of Q^* the test statistic

$$Q = \frac{N-1}{N\sigma_a^2}Q^* = \frac{12}{n(n+1)}\sum_{i=1}^{I}\frac{T_i^2}{n_i} - 3(n+1)$$

is often used. It can be proved that $\mathsf{E}Q = I-1$.

Instead of Q the symbol H is written and one speaks about *H test*. If data contain more than 25 % ties, we should calculate *corrected statistic*

$$Q_{korig} = \frac{Q}{1-(n^3-n)^{-1}\sum(t_i^3-t_i)},$$

where t_1, $t_2, \ldots$ are numbers of ties in groups with the same value.

It can be proved (see Hájek, Šidák 1967), that under H_0 the hypothesis Q has asymptotically χ^2 distribution, when all n_i tend to infinity. Since $\mathsf{E}Q = I-1$, we will have asymptotically χ^2 distribution with $I-1$ degrees of freedom. Thus we reject hypothesis H_0 when $Q \geq \chi^2_{I-1}(\alpha)$.

Kruskal-Wallis test is sensitive especially in the cases when the distribution functions differs by shift. If we reject H_0, it is worth to decide which pairs of samples significantly differ. In analysis of variance Tukey method was used. Kruskal-Wallis test can be supplemented as follows (see Miller 1966). Denote $t_i = T_i/n_i$, $i = 1, \ldots, I$. Let $h_{I-1}(\alpha)$ be critical value of Kruskal-Wallis test on level α. If the sizes of samples are small, $h_{I-1}(\alpha)$ can be found in special tables and in greater sizes the approximation $h_{I-1}(\alpha) \doteq \chi^2_{I-1}(\alpha)$ can be used. We decide that the distribution functions of the i-th and the j-th sample significantly differ if it holds

$$|t_i - t_j| > \sqrt{\frac{1}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)n(n+1)\,h_{I-1}(\alpha)}. \tag{15.12}$$

Probability that at least two of $I(I-1)/2$ distribution functions F_i, F_j will be declared that they are significantly different, although H_0 holds, is not greater than α.

If the sizes of all samples are the same, say $n_1 = \cdots = n_I = m$, then procedures of multiple comparisons can be based on Tukey idea which was applied in one-way analysis of variance (see Neményi 1963 and Miller 1966). For small values m and I are critical values for $|T_i - T_j|$ tabulated. If m and I are large, we use the following method. Let

$q_{I,\infty}(\alpha)$ be critical value of the range of I independent random variables with distribution defined as follows. If $\xi_1, \ldots, \xi_I$ is the sample from $\mathsf{N}(0,1)$, we denote $R = \xi_{(I)} - \xi_{(1)}$ its *range*. Then $q_{I,\infty}(\alpha)$ is defined by the condition

$$\mathsf{P}[R \geq q_{I,\infty}(\alpha)] = \alpha.$$

We declare that F_i and F_j significantly differ if

$$|t_i - t_j| > q_{I,\infty}(\alpha)\sqrt{\frac{1}{12}I(Im+1)}. \tag{15.13}$$

If possible, we prefer Nemény method, since it is more sensitive.

Other tests, which can be used instead of Kruskal-Wallis test, are described in the paper Bhapkar, Deshpande (1968).

It is necessary to point out that assigning ranks is monotonous transformation, but nonlinear. The nonlinearity may lead to paradoxal results, as it is shown in the next example (see Haunsperger, Saari 1991).

Example 15.18 Two workshops are in a factory, in each of them 6 identical machines. However, 2 are from producer A, 2 from B and 2 from C. The production is described in tab. 15.9 – 15.11.

Table 15.9: Workshop No. 1

Producer	Productivity of machines		Rank		Sum of ranks
A	5.89	5.98	3	5	8
B	5.81	5.90	2	4	6
C	5.80	5.99	1	6	7

Table 15.10: Workshop No. 2

Producer	Productivity of machines		Rank		Sum of ranks
A	5.69	5.74	3	5	8
B	5.63	5.71	2	4	6
C	5.62	6.00	1	6	7

Table 15.11: Factory

Producer	Productivity of machines				Rank				Sum of ranks
A	5.89	5.98	5.69	5.74	8	10	3	5	26
B	5.81	5.90	5.63	5.71	7	9	2	4	22
C	5.80	5.99	5.62	6.00	6	11	1	12	30

V workshop 1 the arrangement is $A \succ C \succ B$. Also in the workshop 2 we have $A \succ C \succ B$. In the factory we have $C \succ A \succ B$, which is completely different. $\diamondsuit$

Table 15.12: Ranks of the yields of potatoes

Sort	Ranks						
A	8	6	15	17	12	10	22
B	20	27	26	25	24	14	28
C	21	11	9	23	16	18	19
D	4	2	3	5	13	1	7

Some paradoxes are described in papers Haunsperger (1992) and Saari (1989).

Example 15.19 We use data from example 15.15, p. 141. The numbers given in tab. 15.3 will be ordered and their ranks assigned. We get tab. 15.12.

We have $n_i = 7$ for $i = 1, 2, 3, 4$ and further we get

$$T_1 = 90, \qquad T_2 = 164, \qquad T_3 = 117, \qquad T_4 = 35, \qquad Q = 18.369.$$

Since $Q \geq \chi^2_3(0.05) = 7.81$, we reject the hypothesis that the samples arise from the distributions with identical distribution function. Now, we calculate multiple comparisons. We get

$$t_1 = 12{,}857, \qquad t_2 = 23.429, \qquad t_3 = 16.714, \qquad t_4 = 5.$$

Critical value is 12.3. In our case the samples have the same size and thus we can calculate also critical value using (15.13). The result is 11.3. We use this second critical value, because it is smaller. The differences $t_i - t_j$ are introduced in tab. 15.13. The significance is denoted by a star.

Table 15.13: Values $t_i - t_j$

	j		
i	2	3	4
1	−10.57	−3.86	7.86
2		6.72	18.43*
3			11.71*

The sorts are arranged according t_i and we underline by unbroken line the groups which do not differ significantly. We obtain tab. 15.14.

Table 15.14: Sorted means

D A C B

It is proved that D is significantly different from C and also from B. These results are weaker than those obtained from the analysis of variance in example 15.15.

Using program `R` we get

```
kruskal.test(vynos ~ odruda)
        Kruskal-Wallis rank sum test
data:  vynos by odruda
Kruskal-Wallis chi-squared = 18.3695, df = 3, p-value = 0.000369
```

We obtained significant result. $\diamondsuit$

Chapter 16

Discrete problem of k samples

16.1 Testing homogeneity by method χ^2

Test of homogeneity of two binomial distributions was described in sec. 14.1. It was proved that it corresponds to the test of homogeneity in 2×2 table. For test of homogeneity of r binomial distributions we can use χ^2 test for $r \times 2$ table.

Theorem 16.1 *Let $c = 2$. Then*

$$\chi^2 = \frac{n^2}{n_{.1}n_{.2}} \sum_{i=1}^{r} n_{i.} \left(\frac{n_{i1}}{n_{i.}} - \frac{n_{.1}}{n} \right)^2 . \tag{16.1}$$

Proof. We use formula (10.18) for calculating χ^2 in contingency tables. We insert $n_{i2} = n_{i.} - n_{i1}$, $n_{.2} = n - n_{.1}$, and we obtain

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{r} \left[\frac{\left(n_{i1} - \frac{n_{i.}n_{.1}}{n} \right)^2}{\frac{n_{i.}n_{.1}}{n}} + \frac{\left(n_{i2} - \frac{n_{i.}n_{.2}}{n} \right)^2}{\frac{n_{i.}n_{.2}}{n}} \right] \\ &= \frac{1}{n} \sum_{i=1}^{r} \left[\frac{(nn_{i1} - n_{i.}n_{.1})^2}{n_{i.}n_{.1}} + \frac{(nn_{i2} - n_{i.}n_{.2})^2}{n_{i.}n_{.2}} \right] . \end{aligned}$$

Since

$$nn_{i2} - n_{i.}n_{.2} = n(n_{i.} - n_{i1}) - n_{i.}n_{.2} = -nn_{i1} + (n_{.1} + n_{.2})n_{i.} - n_{i.}n_{.2} = -nn_{i1} + n_{i.}n_{.1},$$

we get

$$\begin{aligned} \chi^2 &= \frac{1}{n} \sum_{i=1}^{r} (nn_{i1} - n_{i.}n_{.1})^2 \left[\frac{1}{n_{i.}n_{.1}} + \frac{1}{n_{i.}n_{.2}} \right] \\ &= \frac{1}{n} \sum_{i=1}^{r} (nn_{i1} - n_{i.}n_{.1})^2 \frac{n_{.2} + n_{.1}}{n_{i.}n_{.1}n_{.2}} \\ &= \sum_{i=1}^{r} \frac{(nn_{i1} - n_{i.}n_{.1})^2}{n_{i.}n_{.1}n_{.2}} \\ &= \frac{n^2}{n_{.1}n_{.2}} \sum_{i=1}^{r} n_{i.} \left(\frac{n_{i1}}{n_{i.}} - \frac{n_{.1}}{n} \right)^2 . \end{aligned}$$

□

Theorem 16.2 *Let $c = 2$. Then*

$$\chi^2 = \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \sum_{i=1}^{r} \frac{n_{i1}^2}{n_{i\cdot}} - n\frac{n_{\cdot 1}}{n_{\cdot 2}}. \tag{16.2}$$

Proof. From formula (16.1) we obtain

$$\chi^2 = \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \sum_{i=1}^{r} n_{i\cdot} \left(\frac{n_{i1}}{n_{i\cdot}} - \frac{n_{\cdot 1}}{n} \right)^2 = \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \left(\sum_{i=1}^{r} \frac{n_{i1}^2}{n_{i\cdot}} - 2\frac{n_{\cdot 1}}{n} \sum_{i=1}^{r} n_{i1} + \frac{n_{\cdot 1}^2}{n^2} \sum_{i=1}^{r} n_{i\cdot} \right)$$
$$= \frac{n^2}{n_{\cdot 1} n_{\cdot 2}} \sum_{i=1}^{r} \frac{n_{i1}^2}{n_{i\cdot}} - n\frac{n_{\cdot 1}}{n_{\cdot 2}}. \qquad \square$$

Formula (16.1) can be used for calculating χ^2 also in the case that we are interested in testing independence in table $2 \times c$. Formulas for test of independence and test for homogeneity are identical.

16.2 Test based on weighted average

Now, we shall consider the problem of testing homogeneity from other aspect. Let $X_1 \sim \mathsf{Bi}(n_1, p_1), \ldots, X_k \sim \mathsf{Bi}(n_k, p_k)$ and let the variables $X_1, \ldots, X_k$ be independent. We intend to test *hypothesis of homogeneity* $H_0 : p_1 = \cdots = p_k$. Denote $x_i = X_i/n_i$ and $N = n_1 + \cdots + n_k$. We know that $\mathsf{E}x_i = p_i$, $\mathsf{var}\, x_i = p_i(1-p_i)/n_i$. Assume that H_0 holds. The common value of all probabilities p_i will be denoted by p, so that $p = p_1 = \cdots = p_k$. Theorem 15.9 gives that

$$\hat{p} = \frac{1}{N} \sum_i n_i x_i$$

is estimator for parameter p (see formula (15.2).

The central limit theorem gives that x_i has asymptotically distribution $\mathsf{N}\left[p, \frac{p(1-p)}{n_i}\right]$ if hypothesis of homogeneity holds. From theorem 15.10 we get that

$$Q = \frac{1}{p(1-p)} \sum_i n_i (x_i - \hat{p})^2$$

has asymptotically χ^2_{k-1} distribution. However, we do not know the parameter p. Thus we use the estimator $\hat{p}$ instead and we use $\tilde{Q}$ instead of Q, where

$$\tilde{Q} = \frac{1}{\hat{p}(1-\hat{p})} \sum_i n_i (x_i - \hat{p})^2. \tag{16.3}$$

It can be proved that the asymptotic distribution of the variable $\tilde{Q}$ remains the same as the asymptotic distribution of variable Q, namely χ^2_{k-1}. We reject H_0 when $\tilde{Q} \geq \chi^2_{k-1}(\alpha)$. Since the test is asymptotical, for all i it must hold $n_i\hat{p} > 5$. Formula (16.3) can be a little simplified.

Theorem 16.3 *Let $c = 2$. Then*

$$\tilde{Q} = \frac{1}{\hat{p}(1-\hat{p})} \sum_{i=1}^{k} n_i x_i^2 - N\frac{\hat{p}}{1-\hat{p}}. \tag{16.4}$$

This formula is called Brandt-Snedecor formula.

Proof. First, we write (16.3) in the form

$$\tilde{Q} = \frac{1}{\hat{p}(1-\hat{p})} \left(\sum_{i=1}^{k} n_i x_i^2 - 2\hat{p} \sum_{i=1}^{k} n_i x_i + \hat{p}^2 \sum_{i=1}^{k} n_i. \right)$$

Since $\sum n_i x_i = n\hat{p}$, the formula (16.4) follows. □

16.3 Remark on tests

Theorem 16.4 *If $k = 2$, then the test based on statistic $\tilde{Q}$ is identical with two-sided test which is based on U_b defined in formula (14.2).*

Proof. Since $\chi_1^2(\alpha) = \left[u\left(\frac{\alpha}{2}\right)\right]^2$, test based on U_b rejects homogeneity of two binomial distributions when $U_b^2 \geq \chi_1^2(\alpha)$. If we compare denotation used in (14.2) and in our theorem, we can see that

$$x = x_1, \qquad y = x_2, \qquad z = \hat{p}, \qquad m = n_1, \qquad n = n_2.$$

According to (16.3) and to (14.2) we have

$$\tilde{Q} = \frac{1}{z(1-z)}[m(x-z)^2 + n(y-z)^2].$$

Since

$$z = \frac{mx + ny}{m+n},$$

we obtain

$$\tilde{Q} = \frac{1}{z(1-z)} \frac{(x-y)^2}{\frac{1}{m} + \frac{1}{n}} = U_b^2. \qquad \square$$

Similarly, for arbitrary $k \geq 2$ it holds $\tilde{Q} = \chi^2$, where χ^2 is defined in formula (16.2) and $\tilde{Q}$ in (16.4).

Theorem 16.5 *Let $c = 2$. Then*

$$\tilde{Q} = \frac{n^2}{n_{.1}n_{.2}} \sum_{i=1}^{k} \frac{n_{i1}^2}{n_{i.}} - n\frac{n_{.1}}{n_{.2}}. \tag{16.5}$$

Proof. We have $\hat{p} = \frac{n_{.1}}{n}$, $1 - \hat{p} = \frac{n_{.2}}{n}$. We insert into (16.4) and obtain assertion of theorem. □

16.4 Example

Example 16.6 The Faculty of Mathematics and Physics organizes camp for students who enrolled into the first year of study. The camp is in Albeř in district Jindřichův Hradec. The students are divided into groups with respect to the branch of their study. It is Mathematics (M), Physics (F), Computer Science (I), and Teaching (U). Since 2004 the students write the same test from Mathematics. The test consists of 12 problems. Each correctly solved problem brings 1 point. The student having 9 or more points is successful. The unsuccessful solvers are recommended to attend a special introductory course in Praha. The results of the test in year 2010 are introduced in tab. 16.1.

Table 16.1: Results of test from Maths in 2010

		Branch			
Year		M	F	I	U
2010	Successful	104	55	52	7
	Unsuccessful	54	42	70	10

We will deal with the problem if in year 2010 were significant differences among the branches. Using program R we obtain

```
usp10 <- c(104, 55, 52, 7)
neu10 <- c(54, 42, 70, 10)
tbl10 <- rbind(usp10,neu10)
colnames(tbl10) <- c("M","F","I","U")
(tbl10.am <- addmargins(tbl10))
       M  F   I  U Sum
usp10 104 55  52  7 218
neu10  54 42  70 10 176
Sum   158 97 122 17 394
clk10 <- usp10+neu10
prop.test(usp10,clk10)
data:  usp10 out of clk10
X-squared = 16.4601, df = 3, p-value = 0.0009125
alternative hypothesis: two.sided sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.6582278 0.5670103 0.4262295 0.4117647
```

The difference in year 2010 is statistically significant. However, in other years the difference was not significant.

In case of significant result the statistician must decide which pairs are significantly different. This can be done using the following program.

```
pairwise.prop.test(usp10, clk10)
        Pairwise comparisons using Pairwise comparison of proportions
data:  usp10 out of clk10
  1     2     3
2 0.554 -     -
```

```
3 0.001 0.265 -
4 0.328 0.714 1.000
P value adjustment method: holm
```

The result is represented by the matrix of adjusted p-values, calculated by *Holm method* (see Holm 1979). It is one of the methods of multiple comparisons. We can see that there exists only one adjusted p-value, which is smaller than 0.05, namely 0.001. It corresponds to the comparison of the students of mathematics and informatics. These two groups are significantly different. No other differences were significantly different. ♢

Chapter 17

Calculation of power of test

17.1 One-sample test

Let $X_1, \ldots, X_n$ be a random sample from the distribution $\mathsf{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$. At the beginning we assume that the variance σ^2 is known. Test of $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ has critical set

$$\frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \geq u(\alpha).$$

If the true expectation is μ, then the *power of test* (probability of rejection H_0) is

$$\begin{aligned}\beta(\mu) &= \mathsf{P}\left\{ \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \geq u(\alpha) \middle| \mu \right\} \\ &= \mathsf{P}\left\{ \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \geq u(\alpha) + \frac{\mu_0 - \mu}{\sigma}\sqrt{n} \middle| \mu \right\}.\end{aligned}$$

Denote $\Delta = (\mu - \mu_0)/\sigma$. If we want that for the given Δ the test has power β, we must have

$$\mathsf{P}\left\{ \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \geq u(\alpha) + \frac{\mu_0 - \mu}{\sigma}\sqrt{n} \middle| \mu \right\} = \beta. \tag{17.1}$$

Consider size n of the sample which ensures that for the given Δ the power of test $\beta(\mu)$ will be equal to the given probability β. From (17.1) we have the condition

$$u(\alpha) + \Delta\sqrt{n} = u(\beta),$$

so that

$$n = \frac{[u(\beta) - u(\alpha)]^2}{\Delta^2}.$$

We solved the right-hand side test. The results for left-hand side test and for the two-sided test can be derived analogously. For details see Hátle, Likeš (1972), p. 287.

If the variance σ^2 is not known, then the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ has critical set

$$T' = \frac{\bar{X} - \mu_0}{S}\sqrt{n} \geq t_{n-1}(\alpha).$$

If the mean value is μ, then the random variable $(\bar{X} - \mu_0)\sqrt{n}/\sigma$ has distribution $\mathsf{N}(\delta, 1)$, where $\delta = (\mu - \mu_0)\sqrt{n}/\sigma$. Thus T' has *non-central* t *distribution* $\mathsf{t}_{n-1,\delta}$. The power is

$$\beta(\mu) = \mathsf{P}[T' \geq t_{n-1}(2\alpha)] = 1 - F_{n-1,\delta}[t_{n-1}(2\alpha)],$$

where $F_{n-1,\delta}$ is distribution function of $\mathtt{t}_{n-1,\delta}$. The size n of sample ensuring that for a given $\mu > \mu_0$ the power of test is β can be obtained by solving equation $\beta(\mu) = \beta$. Similar procedure can be used for left-hand side and for two-sided test.

Example 17.1 We analyze data introduced in example 9.5, p. 74. We calculate power of two-sided test $H_0 : \mu = 0$ against alternative $H_1 : \mu = 1$. This number 1 is denoted `delta`. First, the data are `odch <- c(-3,2,-2,0,-1)`. Then we use `sd(odch)` and have standard deviation 1.923538. The unknown standard deviation σ will be taken as 2. Further we have

```
power.t.test(n=5, delta=1, sd=2, type="one.sample")
     One-sample t test power calculation
              n = 5
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.1384528
    alternative = two.sided
```

In this function we have `sig.level = 0.05`. The power of the test is only 0.14. If we want to have the power of test 0.9, the sample size is obtained as follows.

```
power.t.test(power=0.9, delta=1, sd=2, type="one.sample")
     One-sample t test power calculation
              n = 43.99552
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
```

The size of the sample would be $n = 44$. ♢

17.2 Paired t-test

It is only a variant of one-sample t-test and so we show only an example.

Example 17.2 We use data introduced in example 11.1, p. 99. We shall calculate power of test for the case that the difference of means is `delta=0.3`. We have

```
ppneu <- c(1.8,1.0,2.2,0.9,1.5,1.6)
lpneu <- c(1.5,1.1,2.0,1.1,1.4,1.4)
rozdil <- ppneu-lpneu
sd(rozdil)
0.194079
```

The unknown standard deviation σ is replaced by the number 0.2. We obtain

```
power.t.test(n=6, delta=0.3, sd=0.2, type="paired")
     Paired t test power calculation
              n = 6
          delta = 0.3
             sd = 0.2
      sig.level = 0.05
          power = 0.8325291
    alternative = two.sided
 NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

In this case the power of test is 0.83, which is rather large. ♢

17.3 Two-sample t-test

We show only a practical application of the test.

Example 17.3 Consider data from example 13.4, p. 112. The program can be applied only in the case that the sizes of both samples are equal. Assume that when the difference of means is 5, the power should be 0.9. We obtain

```
da<-c(62,54,55,60,53,58)
db<-c(52,56,49,50,51)
sd(da) [1]
3.577709
sd(db)
[1] 2.701851
```

Standard deviation of the first sample is 3.6, of the second 2.7. The unknown standard deviation σ will be taken as 3. We calculate

```
power.t.test(power=0.9, sd=3, delta=5)
     Two-sample t test power calculation
              n = 8.649245
          delta = 5
             sd = 3
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
 NOTE: n is number in *each* group
```

The required power of test would be reached in the case that the sample size in each group would be $n = 9$. ♢

17.4 Analysis of variance

Using function `power.anova.test` it is possible to calculate power of test in one-way analysis of variance when the classification is *balanced*, Function `power.anova.test` has arguments

- `groups` Number of groups
- `n` Number of observations (per group)
- `between.var` Between group variance
- `within.var` Within group variance
- `sig.level` Significance level (Type I error probability)
- `power` Power of test (1 minus Type II error probability)

Example 17.4 We calculate power of test for data in example 15.15, p. 141. Value of `within.var` can be estimated by residual variance 3.43, value of `between.var` can be estimated by variance of class averages as follows.

```
x <- c(20.9, 24.6, 22.3, 17.7) # class averages
var(x)
8.329167
```

Now, we calculate the power of the test:

```
power.anova.test(groups=4, n=7, between.var=8.33,
within.var=3.43)
     Balanced one-way analysis of variance power calculation
         groups = 4
              n = 7
    between.var = 8.33
     within.var = 3.43
      sig.level = 0.05
          power = 0.9999658
```

In this case the power is very high, which corresponds to small p-value from example 15.15. ♢

17.5 Test for homogeneity of two binomial distributions

It happens quite often that we have two binomial distributions and we want to test if probability of success is the same for both distributions. Let $X_1 \sim \mathsf{Bi}(n_1, p_1)$ and $X_2 \sim \mathsf{Bi}(n_2, p_2)$ be two independent random variables. If we have their realizations, we are interested in the power of test of hypothesis $H_0 : p_1 = p_2$ against the alternative that the true probabilities p_1 and p_2 are different. If we prepare the samples, we are interested in sample sizes which ensure that the power of test will be β. In the case $n_1 = n_2 = n$ we can use the function `power.prop.test`. We have

```
power.prop.test(n=NULL, p1=NULL, p2=NULL, sig.level=0.05, power=NULL,
  alternative=c("twosided","one.sided"), strict=FALSE)
```

If one of the parameters `n`, `p1`, `p2`, `power`, is unknown, then the function `sig.level` the parameter calculates. We introduce some examples.

Example 17.5 We restrict ourselves to the usual situation `sig.level=0.05`. If `n=50`, `p1=0.5`, `p2=0.6`, then we get

```
power.prop.test(n = 50, p1 = .5, p2 = .6)
     Two-sample comparison of proportions power calculation
              n = 50
             p1 = 0.5
             p2 = 0.6
      sig.level = 0.05
          power = 0.1685815
    alternative = two.sided
 NOTE: n is number in *each* group
```

The power of the test is only 0.1685815. If we want power $\beta = 0.9$, then the calculation

```
power.prop.test(p1 = .5, p2 = .6, power = .9)
     Two-sample comparison of proportions power calculation
              n = 518.0372
             p1 = 0.5
             p2 = 0.6
      sig.level = 0.05
          power = 0.9
    alternative = two.sided
 NOTE: n is number in *each* group
```

gives that we must have at least 518 variables in each group.

♢

Chapter 18

Linear regression models

18.1 Introduction

Linear regression models were described in section 15.1 on p. 129. In this chapter we introduce some special cases. At the beginning we shall assume that $e_1, \dots, e_n$ are independent random variables with distribution $\mathsf{N}(0, \sigma^2)$.

18.2 Basic regression models

18.2.1 Line with zero intercept

Consider the model

$$Y_i = \beta x_i + e_i, \qquad i = 1, \dots, n.$$

If we write this model in the form (15.1), then we can see that the vector $\boldsymbol{\beta}$ has only one component β and it holds $\boldsymbol{X} = (x_1, \dots, x_n)'$. It follows from theorem 15.1 on p. 130 that the estimator of the parameter β is

$$b = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

Using the second formula in theorem 15.3 on p. 130 we have

$$s^2 = \frac{R}{n-1} = \frac{\sum Y_i^2 - b \sum x_i Y_i}{n-1}.$$

Usually, we test the hypothesis $H_0 : \beta = 0$. We use theorem 15.7 on p. 131. If H_0 holds, then the variable

$$T = \frac{b}{s} \sqrt{\sum x_i^2}$$

has the distribution t_{n-1}. If $|T| \geq t_{n-1}(\alpha)$, we reject H_0.

Example 18.1 Bend Y_i (in mm$\times 10^{-2}$) of plastic material was measured in dependence on pressure x_i (in kp/cm^2). The results are introduced in tab. 18.1, see fig. 18.1.

It is known that in this range of values x_i the bend is linearly dependent on pressure. This linear function has zero intercept. We obtain

$$b = 7.857, \qquad s^2 = 4.7143, \qquad T = 69.041.$$

Table 18.1: Bend of material in dependence on pressure

Pressure x_i	2	4	6	8	10	12
Bend Y_i	14	35	48	61	80	93

The value of T exceeds the critical value $t_5(0.05) = 2.571$, and so we reject hypothesis H_0. The calculation is as follows.

```
tlak <- 2*1:6
pruhyb <- c(14, 35, 48, 61, 80, 93)
g <- lm(pruhyb ~ tlak -1)
summary(g)
Call:
lm(formula = pruhyb ~ tlak - 1)
Residuals:
      1       2       3       4       5       6
-1.7143  3.5714  0.8571 -1.8571  1.4286 -1.2857
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
tlak   7.8571     0.1138   69.04 1.21e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.171 on 5 degrees of freedom
Multiple R-squared: 0.999,      Adjusted R-squared: 0.9987
F-statistic:  4767 on 1 and 5 DF,  p-value: 1.207e-08
confint(g)
       2.5 %   97.5 %
tlak 7.564601 8.149685
plot(pruhyb ~ tlak, las=1,  xlab="pressure", ylab="bend")
abline(g)
```

$\diamondsuit$

18.2.2 Regression line

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, n. \tag{18.1}$$

We have $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \ldots & \ldots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \qquad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Denote

$$\bar{Y} = \frac{1}{n}\sum Y_i, \qquad \bar{x} = \frac{1}{n}\sum x_i.$$

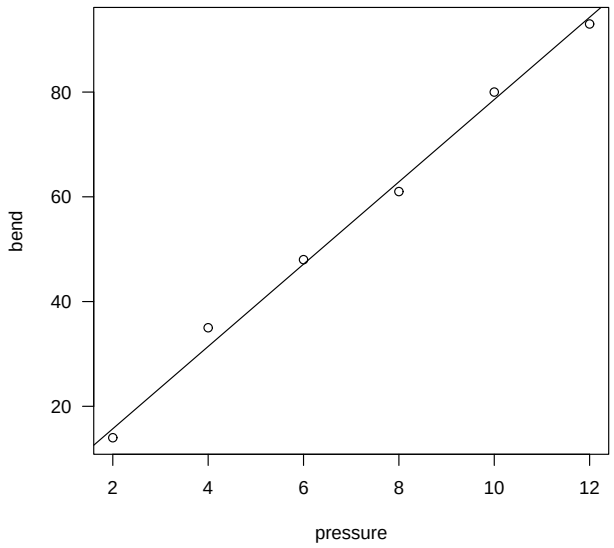

Figure 18.1: Bend of material in dependence on pressure

We obtain estimators

$$b_1 = \frac{\sum x_i Y_i - n\bar{x}\bar{Y}}{\sum x_i^2 - n\bar{x}^2}, \qquad b_0 = \bar{Y} - b_1\bar{x}, \qquad s^2 = \frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i}{n-2}.$$

We show an interpretation of b_1. Assume that $x_i \neq \bar{x}$ holds for all i. Then

$$b_1 = \frac{\sum\limits_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum\limits_j (x_j - \bar{x})^2} = \sum_i \frac{(x_i - \bar{x})^2}{\sum\limits_j (x_j - \bar{x})^2} \, \frac{Y_i - \bar{Y}}{x_i - \bar{x}} = \sum_i w_i \text{ tg } \alpha_i,$$

where weight w_i is

$$w_i = \frac{(x_i - \bar{x})^2}{\sum\limits_j (x_j - \bar{x})^2}$$

and α_i is the angle between horizontal line and the line joining points (x_i, Y_i) and $(\bar{x}, \bar{Y})$. Sum of the weights w_i is one. Thus the regression line is weighted average of all lines which go through points (x_i, Y_i) and their center of gravity $(\bar{x}, \bar{Y})$.

For testing $H_0 : \beta_1 = 0$ calculate

$$T_1 = \frac{b_1}{s}\sqrt{\sum x_i^2 - n\bar{x}^2}.$$

If $|T_1| \geq t_{n-2}(\alpha)$, then H_0 will be rejected.

Let $\boldsymbol{c} = (1, x)'$, where x is a given number. We verify that

$$\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c} = \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2} \, .$$

Theorem 15.8 on p. 132 implies that the interval with endpoints

$$b_0 + b_1 x \mp t_{n-2}(\alpha) s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \tag{18.2}$$

with probability $1-\alpha$ covers the value $\beta_0+\beta_1 x$ (and so it is *confidence interval* for $\beta_0+\beta_1 x$ with confidence coefficient $1-\alpha$). Formula (18.2) defines a hyperbola and thus a confidence band. The band ensures that a value $\beta_0+\beta_1 x$ will be covered with probability $1-\alpha$ but not the line $y=\beta_0+\beta_1 x$. Such a confidence band can be also derived. Instead of $t_{n-2}(\alpha)$ number $\sqrt{2F_{2,n-2}(\alpha)}$ must be used.

Now, we calculate *prediction interval* and *prediction confidence band*. Assume that a new observation Y_0 corresponds to the row $\boldsymbol{x}_0'=(1,x_0)$. We get

$$\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0+1=\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i^2-n\bar{x}^2}+1.$$

Formula (15.6) on p. 135 defines interval with endpoints

$$b_0+b_1x_0\mp t_{n-k}(\alpha)s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i^2-n\bar{x}^2}}.$$

It is confidence interval for Y_0 with confidence coefficient $1-\alpha$. If x varies continuously, we get *prediction confidence band*.

Example 18.2 (Continuation of example 6.3, p. 61.) Concentration of the milk acid in the blood of mothers (values x_i) and their newborn children (values Y_i) was measured. The results are introduced in tab. 6.1 and in tab. 18.2. We assume that the concentration of the milk acid in blood of new born children depends linearly on the concentration in blood of mother.

Table 18.2: Concentration of milk acid

x_i	40	64	34	15	57	45
Y_i	33	46	23	12	56	40

We insert data and calculate basic characteristics.

```
matky <- c(40,64,34,15,57,45)
deti <- c(33,46,23,12,56,40)
g <- lm(deti ~ matky)
summary(g)
Call:
lm(formula = deti ~ matky)
Residuals:
      1       2       3       4       5       6
 0.1358 -7.3677 -4.7384  0.4936  8.6125  2.8642

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3082     7.3615  -0.178  0.86759
matky         0.8543     0.1623   5.265  0.00623 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.312 on 4 degrees of freedom
Multiple R-Squared: 0.8739,     Adjusted R-squared: 0.8424
F-statistic: 27.72 on 1 and 4 DF,  p-value: 0.006232
confint(g)
                 2.5 %    97.5 %
(Intercept) -21.7469727 19.130521
matky         0.4038278  1.304795
```

The results are presented graphically in fig. 18.2.

```
pred.frame <- data.frame(matky=seq(15, 65, by=0.5))
pred.matky <- pred.frame$matky
pp <- predict(g, int="p", newdata=pred.frame)
pc <- predict(g, int="c", newdata=pred.frame)
plot(matky, deti, xlim=c(15,65), xlab="mothers",
  ylab="children", ylim=range(0,70,deti, pp, na.rm=T), pch=16, las=1)
abline(g)
matlines(pred.matky, pc, lty=c(1,2,2), col="black")
matlines(pred.matky, pp, lty=c(1,3,3), col="black")
```

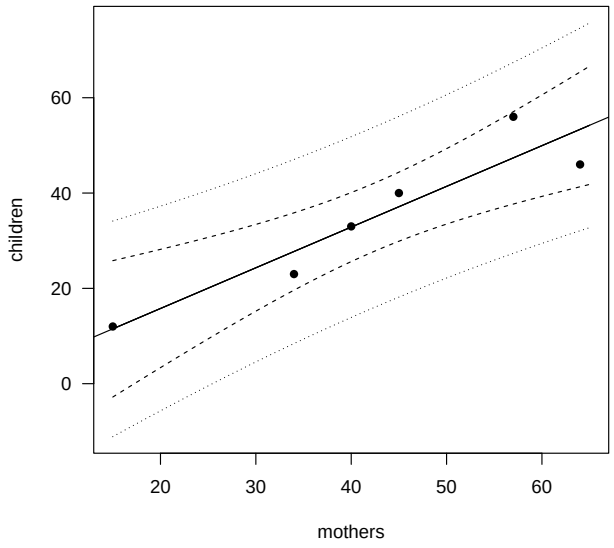

Figure 18.2: Concentration of milk acid

We have got $b_0 = -1.3082$, $b_1 = 0.8543$. Estimator b_1 is significant, its p-value is 0.00623. Confidence bands are presented in fig. 18.2. Confidence interval for β_1 is $(0.40, 1.30)$. $\diamondsuit$

Asymptotic properties of estimators can be derived without assumption of normality. It suffices to assume that $e_1, \ldots, e_n$ are random independent variables with zero expectation and equal variance $\sigma^2 > 0$.

Theorem 18.3 *Let exist numbers c_0, c_1 such that for $n \to \infty$ it holds*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to c_0, \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2 \to c_1,$$

and $c_1 - c_0^2 > 0$. Then b_0 and b_1 are consistent estimators of parameters β_0 and β_1. If additionally $\mathsf{E}(Y_i - \mathsf{E}Y_i)^4 \le D$ holds for some $D > 0$, then s^2 is consistent estimator of parameter σ^2.

Proof see Dupač, Hušková (1999), p. 141. □

Theorem 18.4 *Assume that the assumptions of theorem 18.3 are fulfilled. Moreover, assume that*

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} |x_i|^3 < \infty.$$

Then each of variables

$$\frac{b_0 - \beta_0}{\sqrt{\mathsf{var}\, b_0}}, \qquad \frac{b_1 - \beta_1}{\sqrt{\mathsf{var}\, b_1}}$$

has asymptotically distribution $\mathsf{N}(0,1)$.

Proof see Dupač, Hušková (1999), str. 143. □

18.2.3 Quadratic regression

Under this name we understand the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \qquad i = 1, \dots, n. \tag{18.3}$$

Here we have

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}, \ \boldsymbol{X'X} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}, \ \boldsymbol{X'Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \end{pmatrix}.$$

Estimator $\boldsymbol{b} = (b_0, b_1, b_2)'$ of vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ can be obtained by solving system of equations

$$\boldsymbol{X'Xb} = \boldsymbol{X'Y}.$$

Then we obtain

$$s^2 = \frac{1}{n-3}\left(\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i - b_2 \sum x_i^2 Y_i\right).$$

Denote

$$(\boldsymbol{X'X})^{-1} = \begin{pmatrix} v_{00} & v_{01} & v_{02} \\ v_{10} & v_{11} & v_{12} \\ v_{20} & v_{21} & v_{22} \end{pmatrix}.$$

Very often we test the hypothesis $H_0 : \beta_2 = 0$. This is a test if the dependence Y_i on x_i is linear. If H_0 holds then it follows from theorem 15.7 that

$$T_2 = \frac{b_2}{\sqrt{s^2 v_{22}}} \sim \mathsf{t}_{n-3}.$$

In case $|T_2| \geq t_{n-3}(\alpha)$ the hypothesis H_0 is rejected.

Consider test of hypothesis $H_0' : \beta_1 = \beta_2 = 0$. If H_0' holds, then Y_i does not depend on x_i. Alternative hypothesis is that Y_i depends on x_i either linearly or quadratically. For testing H_0' we use theorem 15.12. We know that residual sum of squares is $R = (n-3)s^2$. Here $k = 3$ (matrix $\boldsymbol{X}$ has three columns and rank 3). If H_0' holds, we have submodel

$$Y_i = \beta_0 + e_i.$$

Least squares estimator of the parameter β_0 is $\bar{Y}$, so that

$$R_1 = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2.$$

We have two independent linear bindings ($\beta_1 = 0$, $\beta_2 = 0$), so that $k - l = t = 2$. Thus

$$F = \frac{(R_1 - R)/2}{R/(n-3)} \sim \mathsf{F}_{2,n-3}.$$

If $F \geq F_{2,n-3}(\alpha)$, we reject H_0'.

Example 18.5 In tab. 18.3 we have measurements of density of water ρ in dependence on temperature. Temperature is introduced in degrees of Celsius, density in kg/dm^3. See fig. 18.3.

Table 18.3: Density of water

t	ρ	t	ρ	t	ρ
0	1,000	40	0,993	80	0,973
10	1,000	50	0,987	90	0,964
20	0,997	60	0,983	100	0,958
30	0,996	70	0,978		

We calculate

```
teplota <- 10*0:10
hustota <- c(1, 1, 0.997, 0.996, 0.993, 0.987,
0.983, 0.978, 0.973, 0.964, 0.958)
g <- lm(hustota ~ teplota + I(teplota^2))
summary(g)
Call:
lm(formula = hustota ~ teplota + I(teplota^2))
Residuals:
       Min         1Q     Median         3Q        Max
-1.114e-03 -5.753e-04 -6.853e-05  6.727e-04  1.052e-03
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)   1.000e+00  6.870e-04 1456.302  < 2e-16 ***
teplota      -6.312e-05  3.196e-05   -1.975   0.0837 .
I(teplota^2) -3.660e-06  3.078e-07  -11.888  2.3e-06 ***
```

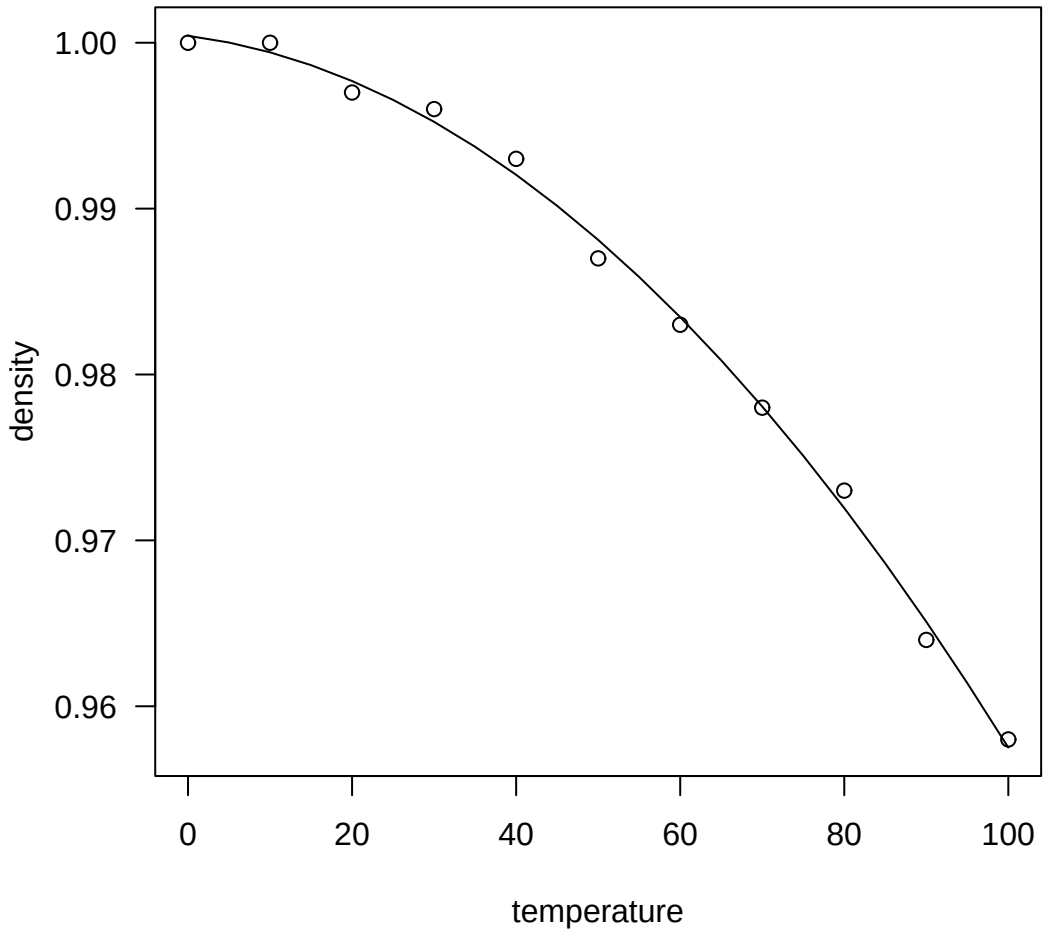

Figure 18.3: Density of water in dependence on temperature

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0009017 on 8 degrees of freedom
Multiple R-squared: 0.997,      Adjusted R-squared: 0.9962
F-statistic:  1316 on 2 and 8 DF,  p-value: 8.428e-11
```

Program for figure is

```
pred.frame <- data.frame(teplota=seq(0, 100, by=5))
pred.hustota <- predict(g, newdata=pred.frame)
new <- pred.frame$teplota
plot(teplota, hustota, ylim=range(hustota, pred.hustota), las=1)
matlines(new, pred.hustota)
```

Since F is significant, we proved that density of water depends on its temperature. The coefficient by quadratic component is significant, we reject hypothesis that the dependence is linear. $\diamondsuit$

18.2.4 Two independent variables

Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i, \qquad i = 1, \ldots, n,$$

so that

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ \cdots & \cdots & \cdots \\ 1 & x_n & z_n \end{pmatrix},$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum x_i & \sum z_i \\ \sum x_i & \sum x_i^2 & \sum x_i z_i \\ \sum z_i & \sum x_i z_i & \sum z_i^2 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum z_i Y_i \end{pmatrix}.$$

Estimator $\boldsymbol{b} = (b_0, b_1, b_2)'$ of vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ can be calculated from the system of normal equation

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}.$$

Using formulas

$$R = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i - b_2 \sum z_i Y_i, \qquad s^2 = \frac{R}{n-3}$$

we obtain residual sum of squares R and residual variance s^2. Elements of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ will be again denoted v_{ij}, $i, j = 0, 1, 2$. If we want to test $H_0 : \beta_2 = 0$, we calculate

$$T_2 = \frac{b_2}{\sqrt{s^2 v_{22}}}.$$

In the case that $|T_2| \geq t_{n-3}(\alpha)$ we reject H_0 and it is proved that Y_i depends on z_i. Similarly, for testing $H_0' : \beta_1 = 0$ calculate

$$T_1 = \frac{b_1}{\sqrt{s^2 v_{11}}}$$

and H_0' will be rejected, when $|T_1| \geq t_{n-3}(\alpha)$. If H_0' is rejected, the dependence Y_i on x_i is proved.

Consider testing hypothesis $H_0^* : \beta_1 = \beta_2 = 0$. We start with calculation

$$R_1 = \sum Y_i^2 - n\bar{Y}^2.$$

General theory gives that the variable

$$F = \frac{(R_1 - R)/2}{R/(n-3)}$$

has distribution $\mathsf{F}_{2,n-3}$ when H_0^* holds. If $F \geq F_{2,n-3}(\alpha)$, we reject H_0^*. This will prove that Y_i depends either on x_i or on z_i or on both x_i and z_i.

Example 18.6 (Continuation of example 6.3, p. 61.)

Remember that in year 1957 statisticians investigated expenses Y_i for food and drinks in households in dependence on number X_i of people in the household and on net earnings Z_i. The data concerning 7 randomly chosen households are given in Tab. 6.2. Assume that the dependence of expenses on remaining two variables is linear and calculate regression analysis. The program is

```
vyd <- c(4,3,4,1,6,4,5)
velik <- c(4,2,4,1,5,3,4)
prijem <- c(10,8,12,3,15,12,13)
potr <- data.frame(vyd,velik,prijem)
g <- lm(vyd ~ velik + prijem)
summary(g)
Call:
lm(formula = vyd ~ velik + prijem)
Residuals:
       1         2         3         4         5         6         7
 0.029903  0.252419 -0.536500 -0.003518  0.285840 -0.208443  0.180299
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.17414    0.40812  -0.427   0.6916
velik        0.32806    0.27125   1.209   0.2931
prijem       0.28320    0.09473   2.990   0.0404 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.3571 on 4 degrees of freedom
Multiple R-squared: 0.9657,     Adjusted R-squared: 0.9485
F-statistic: 56.25 on 2 and 4 DF,  p-value: 0.001179
```

Statistic F is significant. Expenses for food depend on number of people in the household and on income. However, influence of number of people is not significant. The number of data is very small.

♢

Bibliography

[1] Agresti A. (2002): Categorical Data Analysis. 2nd Ed. Wiley, Hoboken, New Jersey.

[2] Anděl J. (1978): Matematická statistika. SNTL, ALFA, Praha.

[3] Anděl J. (2005): Statistické metody. Matfyzpress, Praha.

[4] Anděl J. (2007): Základy matematické statistiky. Matfyzpress, Praha.

[5] Angus J. E. (1994): The probability integral transform and related results. *SIAM Review* **36**, 652–654.

[6] Barr D. R., Davison T. (1973): A Kolmogorov–Smirnov test for censored samples. *Technometrics* **15**, 739–757.

[7] Bennet B. M. (1962): On multivariate sign tests. *J. Roy. Statist. Soc.* **24**, 159–161.

[8] Campbell D. B., Oprian C. A. (1979): On the Kolmogorov–Smirnov test for the Poisson distribution with unknown mean. *Biom. J.* **21**, 17–24.

[9] Conover W. J. (1972): A Kolmogorov goodness-of-fit test for discontinuous distributions. *J. Amer. Statist. Assoc.* **67**, 591–596.

[10] Cramér H. (1946): Mathematical Methods of Statistics. Princeton Univ. Press, Princeton.

[11] Edwards A. W. F. (1963): The measure of association in a 2×2 table. *J. Roy. Statist. Soc.* **A 126**, 109–114.

[12] Farnsworth D. L. (2004): The most compact subdomain of a continuous probability distribution. *Teaching Statistics* **26**, 81*–83.

[13] Fisher R. A. (1935): The logic of inductive inference. *J. Roy. Statist. Soc.* **98**, 39–82.

[14] Gnedenko B. V. (1954): Kurs teorii verojatnostej. 2. vyd., Gos. izd., Moskva.

[15] Goodman L. A. (1964): Simultaneous confidence limits for cross-product ratios in contingency tables. *J. Roy. Statist. Soc.* **B 26**, 86–102.

[16] Haberman S. J. (1974): The Analysis of Frequency Data. Univ. of Chicago Press, Chicago.

[17] Hájek J., Šidák Z. (1967): Theory of Rank Tests. Academia, Praha.

[18] Haunsperger D. B. (1992): Dictionaries of paradoxes for statistical tests on k samples. *J. Amer. Statist. Assoc.* **87**, 149–155.

[19] Haunsperger D. B, Saari D. G. (1991): The lack of consistency for statistical decision procedures. *Amer. Statist.* **45**, 252–255.

[20] Hayter A. J. (1984): A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Ann. Statist.* **12**, 61–75.

[21] Hodges J. L., Lehmann E. L. (1963): Estimation of location based on ranks. *Ann. Math. Statist.* **34**, 598–611.

[22] Hollander M., Wolfe D. A. (1973): Nonparametric Statistical Methods. Wiley, New York.

[23] Hooker R. H. (1907): The correlation of the weather and crops. *J. Roy. Statist. Soc.* **70**, 1–42.

[24] Hyndman R. J., Fan Y. (1996): Sample quantiles in statistical packages. *Amer. Statistician* **50**, 361–365.

[25] Iman R. L. (1982): Graphs for use with the Lilliefors test for normal and exponential distributions. *Amer. Statist.* **36**, 109–112.

[26] Kendall M. G. (1962): Rank Correlation Methods, 3rd ed. Griffin, London.

[27] Kendall M. G., Stuart A. (1969, 1973, 1968): The Advanced Theory of Statistics. I 3. vyd., II 3. vyd., III 2. vyd. Griffin, London.

[28] Konishi S. (1981): Normalizing transformations of some statistics in multivariate analysis. *Biometrika* **68**, 647–651.

[29] Krejn M. T., Nudelman A. A. (1973): Problema momentov Markova i ekstremalnyje zadači. Izd. Nauka, Moskva.

[30] Likeš J., Laga J. (1978): Základní statistické tabulky. SNTL, Praha.

[31] Lilliefors H. W. (1967): On the Kolmogorov-Smirnov tests for normality with mean and variance unknown. *J. Amer. Statist. Assoc.* **62**, 399–402.

[32] Lilliefors H. W. (1969): On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *J. Amer. Statist. Assoc.* **64**, 387–389.

[33] Ling R. F. (1992): Just say no to binomial and other discrete distributions tables. *Amer. Statist.* **46**, 53–54.

[34] Mantel N. (1974): Comment and a suggestion to Conover. *J. Amer. Statist. Assoc.* **69**, 378–380.

[35] Mantel N., Haenszel W. (1959): Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719 – 748.

[36] McNemar Q. (1947): Note on sampling error of the differences between correlated proportions or percentages. *Psychometrika* **12**, 153–154.

[37] O'Cinneide C. A. (1990): The mean is within one standard deviation of any median. *Amer. Statist.* **44**, 292–293.

[38] Peizer D. B., Pratt J. W. (1968): A normal approximation for binomial, F, beta and other common, related tail probabilities. *J. Amer. Statist. Assoc.* **63**, 1416–1456.

[39] Pratt J. W. (1968): A normal approximation for binomial, F, beta and other common, related tail probabilities. II . *J. Amer. Statist. Assoc.* **63**, 1457–1483.

[40] Rao C. R. (1978): Lineární metody statistické indukce a jejich aplikace. Academia, Praha.

[41] Sachs L. (1974): Angewandte Statistik, 4. Auflage. Springer-Verlag, Berlin, Heidelberg, New York.

[42] Shao J. (2005): Mathematical Statistics: Exercises and Solutions. Springer Science + Business Media, Inc.

[43] Simonoff J. S. (2003): Analyzing Categorical Data. Springer, New York.

[44] Stoops G., Barr D. (1971): Moments of certain Cauchy order statistics. *Amer. Statist.* **25**, 51.

[45] Stuart A. (1955): A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* **42**, 412–416.

[46] van der Waerden B. L. (1957): Mathematische Statistik. Springer-Verlag, Berlin.

[47] Venables W. N., Ripley B. D. (2002): Modern Applied Statistics with S (4th ed.). Springer, New York.

[48] Verzani J. (2002): simpleR — Using R for Introductory Statistics. PDF, `www.math.csi.cuny.edu/Statistics/R/simpleR/Simple`.

[49] Walsh J. E. (1949): Some significance tests for the median which are valid under very general conditions. *Ann. Math. Statist.* **20**, 64–81.

[50] Williams D. (2001): Weighing the Odds. A Course in Probability and Statistics. Cambridge Univ. Press, Cambridge.

[51] Wilson E. B., Hilferty M. M. (1931): The distribution of chi-square. *Proc. Nat. Acad. Sci. USA* **17**, 684–688.

[52] Winterbottom A. (1979): A note on the derivation of Fisher's transformation of the correlation coefficient. *Amer. Statist.* **33**, 142–143.

[53] Yarnold J. K. (1970): The minimum expectation in χ^2 goodness of fit test and the accuracy of approximations for the null distribution. *J. Amer. Statist. Assoc.* **65**, 864–886.

Author Index

Topic Index