3 The basic concepts of mathematical statistics

Probability theory and mathematical statistics have the subject of their interest in common. Describing the reality they both respect the impact of stable random influences. In contrast to probability theory, mathematical statistics handles with more unfamiliarity with the considered reality - we have to handle with a family of in advance acceptable probability measures P, not only with one measure, as probability theory does. We do not know which measure is appropriate in advance. On the basis of observed and analyzed statistical data we try at lest approximately identify the best fitting probability measure.

A population consists of 10 balls about which we know only they are black or white. We do not know the exact number of black or white balls and we can not find it out. But we can draw one ball with replacement many times. On the basis of the results of drawing we may estimate the unknown number of black balls. This estimation is plausible, if the number of drawing is sufficiently large.
Let us imagine we have drawn 100 times one ball (with replacement) and there have been only 9 black ball results. It is highly probable that the number of black balls is less then the number of white balls. And we can assert more. The numbers 1,2,…,9 are the candidates for the unknown quantity of black balls. It seems, that according to the results, the number one is the most plausible candidate.

Mathematical statistics is concerned with (among others):

a) *theory of estimating parameters*, (e.g. the quantity of black balls in the sample).

b) *theory of hypothesis testing*, (e.g. the hypothesis test whether there are c black balls).

Both procedures are based on statistical data organized in data sets. This procedures give plausible results if the drawing fulfill some assumptions. The following term of random sample is related to the idea of proper collection of data. [Usually, the data processed by statisticians are taken from a population. The portion drawn is called a sample. If the sample is representative of the population, then inferences and conclusions made from the sample can be extended to the population as a whole.]

Definition 3.1

(i.) Let $X_1, \dots, X_n$ be independent and identically distributed random variables, $X_i \sim L(\vartheta),\ i = 1, \dots, n$, where $L(\vartheta)$ is a probability distribution. Then $X_1, \dots, X_n$ is referred to as *random sample* of size n from the probability distribution $L(\vartheta)$. The realizations $x_1, \dots, x_n$ organized into the column vector form data set.

(ii.) Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent and identically distributed random vectors, $(X_i, Y_i) \sim L_2(\vartheta),\ i = 1, \dots, n$; where $L_2(\vartheta)$ is a joint probability distribution. Then $(X_1, Y_1), \dots, (X_n, Y_n)$ is referred to as *random sample* of size n from the two-dimensional probability distribution $L_2(\vartheta)$. The realization $(x_1, y_1), \dots, (x_n, y_n)$ organized into the matrix $n \times 2$ form the data set.

(iii.) Analogously random sample of size n from the $p-$dimensional probability distribution $L_p(\vartheta)$, $p \geq 3$ can be defined.

(iv.) Any function T of a random sample (or a number of random samples), where the function itself is independent of the sample's distribution, is called *statistic*. The term is used both for the function and for the value of the function on a given sample.

Remark
A statistic is an observable random variable, which differentiates it from a parameter, a generally unobservable quantity describing a property of a statistical population.

Corollary 3.2
Let $X_1, \dots, X_n$ be a random sample from distribution function $F(x)$.
Let $F_*(\mathbf{x}) = F_*(x_1, \dots, x_n)$ be a joint distribution function of a random vector $(X_1, \dots, X_n)$. Then the following assertion holds:
$F_*(\mathbf{x}) = F(x_1) \cdot F(x_2) \cdot \ldots \cdot F(x_n)$

The following definition lists essential often used statistics (plural of statistic).

Definition 3.3

(i.) Let $X_1, \dots, X_n$ be a random sample, $n \geq 2$

- Statistic $M = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ is called *sample mean.* [Instead of M it can be denoted $\bar{X}$.]
- Statistic $S^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - M)^2$ is called *sample variance.*
- Statistic $S = \sqrt{S^2}$ is called *sample standard deviation.*
- Statistic $F_n(x) = \frac{1}{n} \cdot card\{i,\ X_i \leq x\},\ x \in \mathbf{R}$ is called *the value of sample distribution function in a point* x. [for any fixed real number $x:\ card\{i,\ X_i \leq x\}$ is a quantity of those realizations of a random vector which are less than or equal to x.]

(ii.) Let $X_{11}, \dots, X_{1n_1}; \dots; X_{p1}, \dots, X_{pn_p}$ be a sequence of p independent random samples of sizes $n_1 \geq 2, \dots, n_p \geq 2$,. The total size is $n = \sum\limits_{j=1}^{p} n_j$. Let us denote sample means as $M_1, \dots, M_p$ and sample variances of particular samples as $S_1^2, \dots, S_p^2$. Let $c_1, \dots, c_p$ be real constants at least one of which is non-zero.

- Statistic $\sum\limits_{j=1}^{p} c_j M_j$ is called *linear combination of sample means*
- Statistic $S_*^2 = \frac{\sum\limits_{j=1}^{p}(n_j-1)S_j^2}{n-p}$ is called *weighted mean of sample variances*

(iii.) Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from two-dimensional distribution. Let us denote sample means as $M_1 = \frac{1}{n} \sum\limits_{i=1}^{n} X_i,\ M_2 = \frac{1}{n} \sum\limits_{i=1}^{n} Y_i$
and sample variances as $S_1^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - M_1)^2, S_2^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (Y_i - M_2)^2$.

- Statistic $S_{12} = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - M_1)(Y_i - M_2)$ is called *sample covariance*
- Statistic $R_{12} = \begin{cases} \frac{1}{n-1} \sum\limits_{i=1}^{n} \frac{(X_i - M_1)}{S_1} \frac{(Y_i - M_2)}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{for } S_1 S_2 \neq 0 \\ 0 & \text{otherwise} \end{cases}$
  is called *sample correlation coefficient*

[Transforming the random sample by particular function we obtain statistics $M, S^2, S, S_{12}, R_{12}$, thus these statistics are random variables and are denoted by capital letters. The numerical realization of a random sample leads to numerical realizations of previously mentioned statistics which are denoted by small letters $m, s^2, s, s_{12}, r_{12}$. These realisations are corresponding to characteristics known from descriptive statistics. But there is an important difference. In case of variance, covariance and correlation coefficient there is a constant $\frac{1}{n-1}$ in front of sum instead of $\frac{1}{n}$ as it is used in descriptive statistics.

Example 3.4
An unknown constant μ was mutually independently measured by 10 times. The results of measurement follow: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2. These results we can view as a realization of a random sample $X_1, \dots, X_{10}$. Calculate m, s^2 and the values of a sample distribution function $F_{10}(x)$.

Solution

$$\begin{aligned}
m &= \tfrac{1}{n}\sum_{i=1}^{n} x_i = \tfrac{1}{10}(2+1,8+\ldots+2,2) = 2,06\\
s^2 &= \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-m)^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2mx_i+m^2) = \tfrac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - 2m\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}m^2] =\\
&= \tfrac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - 2mnm + nm^2] = \tfrac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - nm^2] =\\
&= \tfrac{1}{9}(2^2+1,8^2+\ldots+2,2^2-10\cdot 2,06^2) = 0,0404\\
s &= \sqrt{s^2} = \sqrt{0,0404} = 0,2011
\end{aligned}$$

To make the calculation of $F_{10}(x)$ easier , the results of measurement will be in ascendent order:
1,8; 1,8; 1,9; 2; 2; 2,1; 2,1; 2,2; 2,3; 2,4;

for $x < 1,8:\ F_{10}(x) = 0$
for $1,8 \le x < 1,9:\ F_{10}(x) = 0,2$
for $1,9 \le x < 2:\ F_{10}(x) = 0,3$
for $2 \le x < 2,1:\ F_{10}(x) = 0,5$
for $2,1 \le x < 2,2:\ F_{10}(x) = 0,7$
for $2,2 \le x < 2,3:\ F_{10}(x) = 0,8$
for $2,3 \le x < 2,4:\ F_{10}(x) = 0,9$
for $x \ge 2,4:\ F_{10}(x) = 1$

Example 3.5

Consider 11 randomly drawn cars of a particular brand. A random variable X stands for an age of a car (in years) and Y stands for a price of a car (in Kč). The results are listed in the following tab:

X	5	4	6	5	5	5	6	6	2	7	7
Y	85	103	70	82	89	98	66	95	169	70	48

Calculate and interpret r_{12}.

Solution

$$\begin{aligned}
m_1 &= \tfrac{1}{11}(5+4+\ldots+7) = 5,28\\
m_2 &= \tfrac{1}{11}(85+\ldots+48) = 88,63\\
s_1^2 &= \tfrac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - nm_1^2] = \tfrac{1}{10}(5^2+4^2+\ldots+7^2-11\cdot 5,28^2) = 2,02\\
s_2^2 &= \tfrac{1}{n-1}[\sum_{i=1}^{n}y_i^2 - nm_2^2] = \tfrac{1}{10}(85^2+103^2+\ldots+48^2-11\cdot 88,63^2) = 970,85\\
s_{12} &= \tfrac{1}{n-1}[\sum_{i=1}^{n}(x_i-m_1)(y_i-m_2)] = \tfrac{1}{n-1}[\sum_{i=1}^{n}x_iy_i - nm_1m_2] =\\
&= \tfrac{1}{10}(5\cdot 85+4\cdot 103+\ldots+7\cdot 48-11\cdot 5,28\cdot 88,63) = -40,89\\
r_{12} &= \tfrac{s_{12}}{s_1s_2} = \tfrac{-40,89}{\sqrt{2,02}\sqrt{970,85}} = -0,92
\end{aligned}$$

There is a strong decreasing linear relationship between variables X and Y: the older car, the lower price.

The essential properties of frequently used statistics are listed in the following theorem. The properties mentioned in the first paragraph will be derived in seminar.

Theorem 3.6

1.) Let $X_1,\ldots,X_n$ be a random sample from a distribution with expected value μ, variance σ^2 and distribution function $F(x)$. Then:

$E(M) = \mu \qquad D(M) = \frac{\sigma^2}{n},\ n \ge 2 \qquad E(S^2) = \sigma^2$
for any $x \in \mathbf{R}:\quad E[F_n(x)] = F(x), \qquad D[F_n(x)] = \frac{F(x)(1-F(x))}{n}$

2.) Let $X_{11},\ldots,X_{1n_1};\ldots;X_{p1},\ldots,X_{pn_p}$ be a sequence of p independent random samples with mean values $\mu_1,\ldots,\mu_p$ and identical variance σ^2 for each of p samples. Let us denote the total size as $n = \sum_{j=1}^{p} n_j$. Further let $c_1,\ldots,c_p$ be real constants at least one of which is non-zero. Then:

$E(\sum\limits_{j=1}^{p} c_j M_j) = \sum\limits_{j=1}^{p} c_j \mu_j \qquad E(S_*^2) = \sigma^2$

3.) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from two-dimensional distribution with the covariance σ_{12} and the correlation coefficient ρ. Then:

$E(S_{12}) = \sigma_{12},$ however $E(R_{12}) \approx \rho$ (Approximation is appropriate for $n \geq 30$)

Remark 3.7

Methods of mathematical statistics are often used for analyzing and interpreting the results of experiments. In order to do it correctly it is very important to design experiments in a right way. The list of basic patterns of the arrangement of experimental units into groups follows.

a) Simple observation: The random variable X is observed on equal terms. A random sample $X_1, \ldots, X_n$ correspondents to this arrangement of an experiment.

b) Dual observation: The random variable X is observed on two different terms. And there are two different strategies of this arrangement of an experiment:

- Two-sample comparing: The two independent samples $X_{11}, \ldots, X_{1n_1};\ X_{21}, \ldots, X_{2n_2}$, whose sizes may differ, correspondent to this design.
- Pair comparing: A random sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ from two-dimensional distribution correspondents to this design. In this case we transform the given sample into a random sample $Z_1, \ldots, Z_n$; where $Z_i = X_{i1} - X_{i2},\ i = 1, 2, \ldots, n$. This procedure leads to the simple observation.

c) Multiple observation: The random variable X is observed on $p \geq 3$ different terms. And there are two different strategies of this arrangement of an experiment:

- Multi-sample comparing: p independent samples $X_{11}, \ldots, X_{1n_1}; \ldots;\ X_{p1}, \ldots, X_{pn_p}$, whose sizes may differ, correspondent to this design.
- Block comparing: A random sample $(X_{11}, \ldots, X_{1p}), \ldots, (X_{n1}, \ldots, X_{np})$ from $p-$dimensional distribution correspondents to this design.