APPLIED STATISTICS

TEXTBOOK

Doc. RNDr. Zdeněk Karpíšek, CSc.

Ing. Miloš Drdla, Dr., MBA

BRNO 2007

(Second Edition)

Doc. RNDr. Zdeněk Karpíšek, CSc.
Department of Statistics and Optimization
Institute of Mathematics
Faculty of Mechanical Engineering
Brno University of Technology
Technická 2, 616 69 Brno
Phone: +420 541 142 529
E-mail: karpisek@fme.vutbr.cz

Ing. Miloš Drdla, Dr., MBA
Brno International Business School
Lidická 81, 602 00 Brno
Phone: +420 545 570 111
E-mail: drdla@bibs.cz

CONTENT

PREFACE

Modern times have created modern problems and many of those problems involve data. Marketing studies, product testing and quality control are typical application areas that require an intelligent analysis of data. In particular, the business and manufacture now demand employers as well as employees that are better prepared to use statistics.

This textbook is introduction into statistics and consists of five chapters, statistical tables, and bibliography.

Chapter No1: Fundamentals of Probability

Probability plays a special role in all our lives, because we use it to measure uncertainty. We are continually faced with decisions leading to uncertain outcomes and we rely on probability to help us make our choice. A probability is a numerical value that measures the uncertainty that a particular event will occur. The probability of an event ordinarily represents the proportion of times under identical circumstances that the event can be expected to occur.

Chapter No2: Descriptive Statistics

The emphasis on the decision-making aspects of statistics is a recent one. In its early years, the study of statistics consisted of methodology for summarizing or describing numerical data. This area of study has become known as descriptive statistics because it is concerned with summary calculations and graphical displays. These methods are in contrast to the modern statistical approach in which generalizations are made about the whole (called the population) by investigating a portion (referred to as the sample).

Chapter No3: Analysis of Time Series

When data for evaluations are collected at regular intervals from monthly, quarterly, or annual reports, they are referred to as time-series data. In each case, values of the variable being predicted are available for several past periods of time. Such data are called time-series. Statistical procedures that use such values are called time-series analysis.

Chapter No4: Index Numbers

Index numbers are used to facilitate comparisons of time-series data from different periods. Price indexes are especially useful in expressing price changes over time, thereby gauging inflation, a condition of rising prices that permeates the modern world.

Chapter No5: Mathematical Statistics

Mathematical statistics is a body of methods and theory that it applied to numerical evidence when making decisions in the face of uncertainty. This specification treats mathematical statistics as a relatively independent academic field. Using the methods of mathematical statistics, we can describe quantities of random character from observed values. Most frequently, we try to establish the properties of the probability distribution of a random variable - estimating parameters or quantitative characteristics, testing the hypotheses that they have certain properties, analyzing the relationships between them etc.

In writing this introductory statistics textbook for students of business and management, our overriding goal has been to enliven statistics, to make it more interesting and relevant and easier to learn. To illustrate that statistics is neither boring nor irrelevant this textbook treats it as essentially a decision-making tool and includes some relevant concepts and possible applications. We are greatly indebted to RNDr. Karel Mikulášek Ph.D. who has assisted us in preparing and translating this textbook.

Brno — Doc. RNDr. Zdeněk Karpíšek, CSc.

October 2007 — Ing. Miloš Drdla, Dr., MBA

1. FUNDAMENTALS OF PROBABILITY

Random events

An experiment is any well-defined situation or procedure that results in one of two or more possible outcomes. An outcome is a particular result of an experiment. A random experiment is an experiment where there is uncertainty about which of the possible outcomes will occur. We will assume that a random experiment may be repeated a sufficiently large number of times while the conditions that define the random experiment remain constant.

An elementary event $\{\omega\}$ expresses individual outcomes in their simplest terms, that is, in the most elementary form. The set Ω of all possible outcomes for an experiment is called the sample space. A random event (or event) A is a subset of Ω. If $\{\omega\} \in A$, and $\{\omega\}$ is the outcome of a random experiment, we say that a random event A has occurred as a result of the experiment. We will denote random events by A, B, A_1, A_i, … . An event that unavoidably occurs for every random experiment is called certain (sure). It is equivalent to Ω. An event that cannot occur for any random experiment is called impossible and is equivalent to the empty set $\varnothing$.

Relationships between random events are expressed in terms of set inclusions:

a) $A \subseteq B$ means that the occurrence of random event A results in the occurrence of random event B.

b) A = B expresses the equality of random events A and B.

Operations with random events are expressed in terms of set operations:

a) The union $A \cup B$ of random events A and B occurs if at least one of the random events A and B, that is, A and/or B occurs. In a similar way, we define $\bigcup_{i=1}^{n} A_i$ and $\bigcup_{i=1}^{\infty} A_i$.

b) The intersection $A \cap B$ of random events A and B occurs when both events occur. Similarly, we define $\bigcap_{i=1}^{n} A_i$ and $\bigcap_{i=1}^{\infty} A_i$.

c) The difference A – B of random events A and B occurs when A occurs and B does not occur.

d) The complementary event to random event A is the event $\overline{A} = \Omega - A$.

e) Random events A and B are called exclusive if $A \cap B = \varnothing$.

A field of events Σ is a set of random events (a system of subsets of the sample space Ω) with the following properties:

$$\varnothing, \Omega \in \Sigma,$$

$$A \in \Sigma \Rightarrow \overline{A} \in \Sigma,$$

$$A_i \in \Sigma,\ i = 1, 2, \ldots \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \Sigma.$$

Example 2.1

A random experiment is made by throwing a dice from homogeneous material with faces numbered from 1 to 6. Event A occurs when an even number comes up and event B occurs if a number greater than 4 comes up. Determine Ω, $\overline{A}, \overline{B}$, $A \cup B$, $A \cap B$, A – B, B – A, Σ.

Solution:

The sample space Ω = {1, 2, 3, 4, 5, 6} with elementary events {1}, {2}, {3}, {4}, {5}, {6}. Next we have A = {2, 4, 6} a B = {5, 6} so that

$\overline{A}$ = {1, 3, 5} ... an odd number comes up,

$\overline{B}$ = {1, 2, 3, 4} ... a number less than 5 comes up,

$A \cup B = \{2, 4, 6\} \cup \{5, 6\} = \{2, 4, 5, 6\}$ ... numbers 1 a 3 do not come up,

$A \cap B = \{2, 4, 6\} \cap \{5, 6\} = \{6\}$ ... number 6 comes up,

$A - B = \{2, 4, 6\} - \{5, 6\} = \{2, 4\}$ ... number 2 or 4 comes up,

B – A = {5, 6} – {2, 4, 6} = {5} ... number 5 comes up.

Since no restrictions are imposed on random events, we can consider the maximal field of events (the system of all subsets of Ω):

Σ = {$\varnothing$, {1}, {2}, {3}, {4}, {5}, {6}, {1,2}, {1,3},…,{5,6},…, {2, 3, 4, 5, 6}, Ω }

which contains 2^6 = 64 random events.

Axiomatic definition of probability and its properties

The probability P(A) of a random event $A \in \Sigma$ is a non-negative function (defined on Σ) such that P(Ω) = 1 and, for each sequence of mutually exclusive random events $A_i \in \Sigma$, i = 1, 2, …, we have $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

It holds:

a) $P(\overline{A}) = 1 - P(A)$; P($\varnothing$) = 0; $0 \le P(A) \le 1$.

b) $A \subseteq B \Rightarrow P(A) \le P(B)$ and P(B – A) = P(B) – P(A).

c) $P(A_1 \cup ... \cup A_n) = 1 - P(\overline{A}_1 \cap ... \cap \overline{A}_n) =$

$$= \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{n} P(A_i \cap A_j) + ... + (-1)^{n-1} P(A_1 \cap ... \cap A_n) \ .$$

This means particularly that $P(A \cup B) = 1 - P(\overline{A} \cap \overline{B}) = P(A) + P(B) - P(A \cap B)$.

For a finite or countable sample space Ω (i.e. its elementary events $\{\omega\}$ may be listed in a sequence) we get

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

Particularly, for a sample space Ω with n equally probable elementary events we have

$$P(A) = \frac{m}{n},$$

where m is the number of elementary events $\{\omega\}$ of which random event A consists. We say that "m is the number of favourable outcomes of the experiment" and "n is

the number of outcomes of the experiment". This is the so-called classical definition of probability.

Example 2.2

Calculate the probabilities P(A), P(B), $P(\overline{A})$, $P(\overline{B})$, $P(A \cup B)$, $P(A \cap B)$, P(A – B), P(B – A) of the random events from Example 2.1.

Solution:

Due to the symmetry and homogeneity of the cube, all the elementary events have the same probability $P(\{\omega\}) = 1/6$ and n = 6. This yields the following probabilities:

$P(A) = 3/6 = 1/2,$

$P(B) = 2/6 = 1/3,$

$P(\overline{A}) = 3/6 = 1/2,$

$P(\overline{B}) = 4/6 = 2/3,$

$P(A \cup B) = 4/6 = 2/3,$

$P(A \cap B) = 1/6,$

$P(A - B) = 2/6 = 1/3,$

$P(B - A) = 1/6.$

Using the properties of probability we can, for example, calculate

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2},$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}.$$

Example 2.3

In a supply of 100 shafts, 10 items do not comply with the standard diameter, 20 items have not the required length and 5 items comply neither with the length nor with the diameter requirement. Calculate the probability that a shaft selected at random has both the required length and diameter.

S o l u t i o n:

Denoting by A and B the event that the selected shaft does not comply with the required diameter and length respectively, the probability that the selected shaft has both the required length and diameter

$$P(\overline{A} \cap \overline{B}) = 1 - P\left(\overline{\overline{A} \cap \overline{B}}\right) = 1 - P(A \cup B) = 1 - \left[P(A) + P(B) - P(A \cap B)\right] =$$
$$= 1 - (0.10 + 0.20 - 0.05) = 0.75 .$$

Conditioned probability and independent events

The probability of event $A \in \Sigma$, given event $B \in \Sigma$, $P(B) \neq 0$, is called the conditional probability of event A

$$P(A / B) = \frac{P(A \cap B)}{P(B)} .$$

It holds:

a) $P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2/A_1) \ldots P(A_n/A_1 \cap \ldots \cap A_{n-1})$,

particularly $P(A \cap B) = P(A)P(B/A) = P(B)P(A/B)$,

b) For a random event $A \subseteq \bigcup_{i=1}^{n} B_i$ where B_i are mutually exclusive random events, $i = 1, \ldots, n$, the probability

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A / B_i)$$

is calculated by the so-called formula of total probability and, for $P(A) \neq 0$, we can derive Bayes' formulas for the probability of hypotheses or Bayes' theorem

$$P(B_j / A) = \frac{P(B_j)P(A / B_j)}{\sum_{i=1}^{n} P(B_i)P(A / B_i)}, \qquad j = 1, \ldots, n.$$

E x a m p l e 2.4

Ten products out of a total of 100 are defective. We choose 3 products at random without replacement. The probability that the first product chosen is defective - random event A_1, the second product chosen is defective - random event A_2, and the third product chosen is not defective – random event $\overline{A}_3$, is calculated below:

$$P(A_1 \cap A_2 \cap \overline{A}_3) = P(A_1)P(A_2 / A_1)P(\overline{A}_3 / A_1 \cap A_2) = \\ = (90/100)(89/99)(10/98) \cong 0.08256 .$$

E x a m p l e 2.5

A grocery receives supplies of croissants from 3 bakehouses in the following quantities: 500, 1000, and 1500 croissants daily. The percentages of defective croissants for the suppliers are 5%, 4%, and 3%. In the grocery, the croissants from individual suppliers are mixed together and sold. A croissant is bought at random. Calculate the probability that

a) it is defective,

b) it has been supplied by the second bakehouse.

S o l u t i o n:

Denote by A the random event that the croissant bought is defective and by B_i, i = 1, 2, 3, the event that the croissant has been supplied by the i-th bakehouse. We get the following probabilities

$$P(B_1) = \frac{500}{500+1000+1500} = \frac{1}{6}, \; P(A/B_1) = 0.05 ,$$

$$P(B_2) = \frac{1000}{500+1000+1500} = \frac{2}{6}, \; P(A/B_2) = 0.04 ,$$

$$P(B_3) = \frac{1500}{500+1000+1500} = \frac{3}{6}, \; P(A/B_3) = 0.03 .$$

Using the formula of total probability we calculate

$$P(A) = (0.05)\frac{1}{6} + (0.04)\frac{2}{6} + (0.03)\frac{3}{6} = \frac{0.22}{6} = 0.03\overline{6} \cong 0.03667 ,$$

so that, from the customer's point of view, the chance of buying a defective croissant is approximately 3.667%. Applying Bayes' theorem for j = 2, we have

$$P(B_2 / A) = \frac{(0.04)\frac{2}{6}}{\frac{0.22}{6}} = \frac{0.08}{0.22} = 0.\overline{36} \cong 0.36364 .$$

In a similar way, we can obtain $P(B_1 /A) \cong 0.22727$ and $P(B_3 /A) \cong 0.40909$, which means that the third bakehouse supplies the largest quantity even if their percentage of defective products is the lowest of the three. This is due to the fact that they supply the largest part of the croissants.

Random events A, B $\in \Sigma$ are called <u>independen*t*</u>, if P(A/B) = P(A) or P(B) = 0. Random events $A_1, \ldots, A_n \in \Sigma$ are <u>mutually independent</u>, if all the pairs of random events

A_i, A_j for $i \neq j$,

$A_i, A_j \cap A_k$ for $i \neq j, i \neq k$,

$A_i, A_j \cap A_k \cap A_m$ for $i \neq j, i \neq k$ and $i \neq m$,

etc.

are independent.

The following assertions are true:

a) A, B are independent if and only if $P(A \cap B) = P(A)P(B)$.

b) If $A_1,\ldots, A_n$ are mutually independent, then

- $P(A_1 \cap \ldots \cap A_n) = P(A_1) \ldots P(A_n)$,
- $P(A_1 \cup \ldots \cup A_n) = 1 - [1 - P(A_1)] \ldots [1 - P(A_n)]$,
- $B_1,\ldots, B_n$ are mutually independent for arbitrary variants $B_i = A_i, \overline{A}_i, \Omega$.

<u>E x a m p l e 2.6</u>

What is the probability that, when throwing a dice, an even number comes up in the first trial, (random event A) and, in the second trial, an odd number comes up (random event B)?

S o l u t i o n:

Random events A and B are independent and their probabilities are $P(A) = P(B) = 1/2$, so that $P(A \cap B) = (1/2)(1/2) = 1/4$.

E x a m p l e 2.7

A product is processed by three independent operations, where the chances of producing a defective product are $P(A_1) = 0.05$, $P(A_2) = 0.08$ and $P(A_3) = 0.03$. Calculate the probability that, after having been processed by all the three operations, the product is defective.

S o l u t i o n:

Since the operations are independent, the random events A_1, A_2, A_3 are mutually independent and the product is defective if at least one of them occurs so that

$$P(A_1 \cup A_2 \cup A_3) = 1 - [1 - P(A_1)][1 - P(A_2)][1 - P(A_3)] = 1 - (0.95)(0.92)(0.97) = 0.152220.$$

Random variable and its probability distributions

A random variable X is a variable that takes on real numbers x and whose outcomes occur by chance – for details see [1], [2], [3] and [4]. Its distribution function of probabilities (or just distribution function) is given by

$$F(x) = P(X < x) = P[X \in (-\infty; x)],\ x \in (-\infty; +\infty).$$

The distribution function has the following properties:

a) $0 \le F(x) \le 1$ for all $x \in (-\infty; +\infty)$,

b) $F(x)$ is non-decreasing and continuous on the left in $(-\infty; +\infty)$.

c) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$,

d) $P(a \le X < b) = F(b) - F(a)$ for arbitrary real numbers $a < b$,

e) $P(X = c) = \lim_{x \to c+} F(x) - F(c)$ for any real c.

A random variable X is said to be discrete with a discrete distribution of probabilities if it takes on at most a countable number of values $x = x_1, x_2, \ldots$. Its probability distribution is given in the form of a sequence

$$p(x) = P(X = x) > 0 \quad \text{for } x = x_1, x_2, \ldots.$$

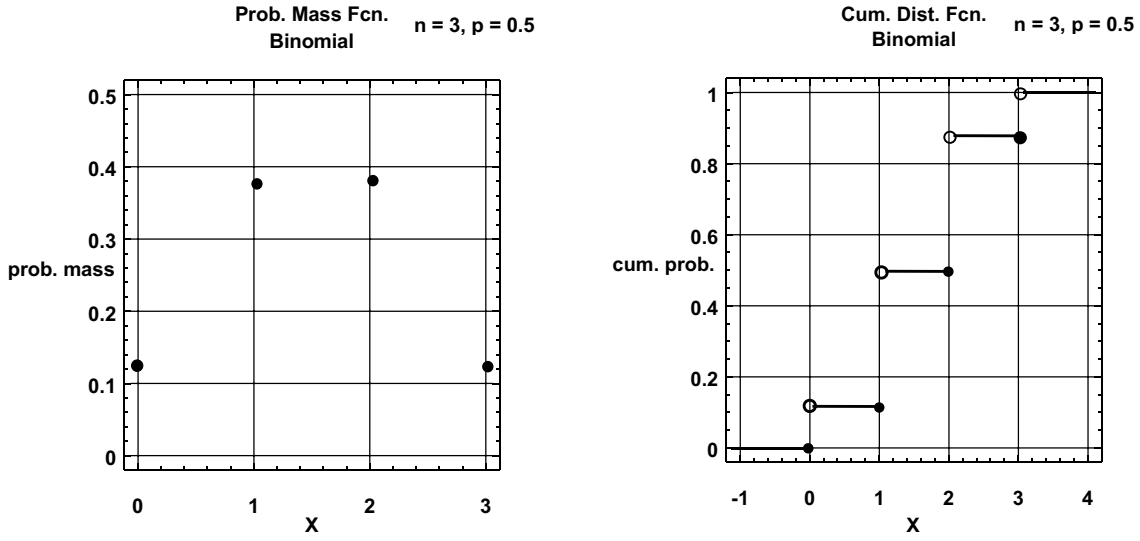
We have:

a) $\sum_x p(x) = 1,$

b) $F(x) = \sum_{t<x} p(t)$ for all $x \in (-\infty; +\infty)$,

c) $P(X \in M) = \sum_{x \in M} p(x)$ for an arbitrary set of real numbers M.

The distribution function for a discrete random variable is a "step-like line with jumps" – see Fig. 2.1.



(a) (b)

Fig. 2.1 The graphs of the distribution of probabilities (a) and the distribution function (b) of a discrete variable.

E x a m p l e 2.8

The probability of failure for each of three independently operating production lines is $0 < p < 1$. The discrete random variable X that expresses the number of production lines with a failure takes on the values x = 0, 1, 2, and 3 and the values of its distribution of probabilities are given below

$p(0) = (1 - p)^3$,

$p(1) = 3p(1 - p)^2$,

$p(2) = 3p^2(1 - p)$,

$p(3) = p^3$.

Its distribution function is the following

$F(x) = 0$ for $x \in (-\infty, 0\rangle$,

$F(x) = p(0) = (1 - p)^3$ for $x \in (0, 1\rangle$,

$F(x) = p(0) + p(1) = (1 + 2p)(1 - p)^2$ for $x \in (1, 2\rangle$,

$F(x) = p(0) + p(1) + p(2) = (1 + p + p^2)(1 - p) = 1 - p^3$ for $x \in (2, 3\rangle$,

$F(x) = p(0) + p(1) + p(2) + p(3) = 1$ for $x \in (3; \infty)$.

In Fig. 2.1 you can see the graphs of p(x) and F(x) for p = 0.5. The probability of a failure occurring in at least one of the production lines is given by

$$P(X \geq 1) = P(1 \leq X < +\infty) = F(+\infty) - F(1) = 1 - (1 - p)^3.$$

A random variable X is said to be continuous with a <u>continuous distribution of probabilities</u> if its distribution function is continuous (thus X takes on all the values of an interval, etc.). Its <u>density function</u>, is such a non-negative function f(x) that

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad \text{for all } x \in (-\infty;+\infty).$$

It has the following properties:

a) $\int_{-\infty}^{+\infty} f(x)dx = 1$,

b) $f(x) = F'(x)$, if the derivative exists,

c) F(x) is a continuous function for all $x \in (-\infty;+\infty)$,

d) $P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$

for arbitrary real numbers $a \leq b$,

e) P(X = c) = 0 for any real c.

E x a m p l e 2.9

A real variable X has a density function f(x) = cx for $x \in \langle 0; 2\rangle$ and 0 for $x \notin \langle 0; 2\rangle$. Using the properties of a continuous random variable, we can derive the following results. We have

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{2} cxdx + \int_{2}^{+\infty} 0dx = \ldots = 2c = 1,$$

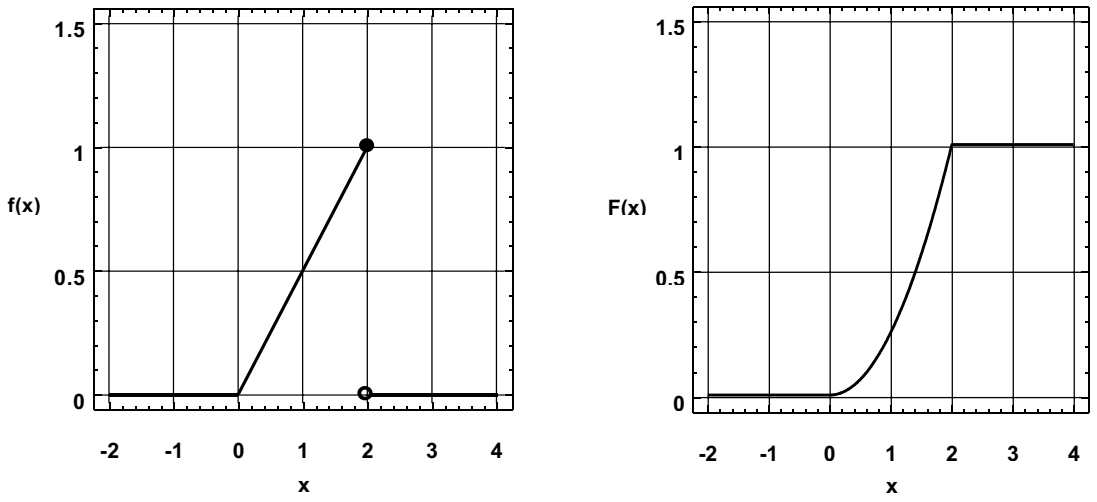
which yields c = 1/2. The distribution function of X is given by

$$F(x) = \int_{-\infty}^{x} 0dt = 0 \quad \text{for } x \in (-\infty; 0\rangle,$$

$$F(x) = \int_{-\infty}^{0} 0dt + \int_{0}^{x} \frac{t}{2}dt = \ldots = \frac{x^2}{4} \quad \text{for } x \in \langle 0; 2\rangle,$$

$$F(x) = \int_{-\infty}^{0} 0dt + \int_{0}^{2} \frac{t}{2}dt + \int_{2}^{x} 0dt = \ldots = 1 \quad \text{for } x \in \langle 2;+\infty).$$

The graphs of f(x) and F(x) are shown in Fig. 2.2. The probability of the random variable taking on a value $x \in \langle 1; 3\rangle$ is $P(1 \le X \le 3) = F(3) - F(1) = 1 - (1^2/4) = 0{,}75$.



(a) (b)

Fig. 2.2 Graphs of the density function (a) and the distribution function (b) of a continuous random variable.

Numerical characteristics of random variables

Numerical characteristics of random variables are real numbers that, in a concentrated way, express their important properties.

The central tendency is characterized by the expected value or mathematical expectation of a random variable X

$$E(X) = \sum_x xp(x) \quad \text{for discrete random variable X,}$$

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad \text{for continuous random variable X,}$$

provided that the sum or the integral is absolutely convergent. The expected value has the following properties:

a) $E(aX + b) = aE(X) + b$ for arbitrary real numbers a, b,

b) $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$ for random variables $X_1,\ldots,X_n$.

A measure of the variation of a random variable X about its expected value E(X) is expressed by its variance (dispersion)

$$D(X) = E\left[(X - E(X))^2\right].$$

The variance has the following properties:

a) $D(X) = \sum_x (x - E(X))^2 p(x) = \sum_x x^2 p(x) - (E(X))^2$ for discrete random variable X,

b) $D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - (E(X))^2$ for continuous random variable X, provided that the sum or the integral converges.

c) $D(X) \geq 0$,

d) $D(aX + b) = a^2 D(X)$ for arbitrary real numbers a, b,

e) $D\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} D(X_i)$ for independent random variables $X_1,\ldots, X_n$.

The standard deviation of a random variable X is defined as $\sigma(X) = \sqrt{D(X)}$. The standard deviation has the following properties:

a) $\sigma(X) \geq 0$;

b) $\sigma(aX + b) = |a|\,\sigma(X)$ for arbitrary real numbers a, b.

The mathematical expectation, is one of the so-called moments about the origin while the variance is a special case of the so-called central moment. You can learn more about moment characteristics such as the coefficient of variation, coefficient of skewness, coefficient of pointedness in [1], [2], [3] and [4].

For a $0 < P < 1$, the P-quantile or the 100P%-quantile (percentile) of a random variable is defined as $x_P = \max\{x;\ F(x) \leq P\}$. For a continuous random variable X with an increasing distribution function we have $F(x_P) = P$. The median of a random variable X is its quantile $x_{0,5}$ and it is characteristic of its position. Further quantile characteristics can be found in [2], [3]. The mode $\bar{x}$ of a random variable X is its value for which the probability distribution or the density function assumes a maximum value or a supremum.

E x a m p l e 2.10

The random variable X from Example 2.9 has the expected value

$$E(X) = \int_{-\infty}^{0} x \cdot 0\,dx + \int_{0}^{2} x \frac{x}{2}\,dx + \int_{2}^{+\infty} x \cdot 0\,dx = \ldots = \frac{4}{3} \cong 1.33333,$$

the variance

$$D(X) = \int_{-\infty}^{0} x^2 \cdot 0\,dx + \int_{0}^{2} x^2 \frac{x}{2}\,dx + \int_{2}^{+\infty} x^2 \cdot 0\,dx - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = \frac{2}{9} \cong 0.22222,$$

and the standard deviation

$$\sigma(X) = \sqrt{\frac{2}{9}} \cong 0.47140\ .$$

The P-quantile x_P is the root of the equation $\frac{x^2}{4} = P$ that lies in the interval $\langle 0;\ 2\rangle$, which means that $x_P = 2\sqrt{P}$ so that the median of X is $x_{0,5} = 2\sqrt{0.5} \cong 1.41421$. From the graph of f(x) in Fig. 2.2, we can see that the mode of X is $\bar{x} = 2$.

Random vectors and their probability distributions

A two-dimensional random vector (a bivariate random variable) is an ordered pair of random variables (X,Y). In [1], [2], and [3], you can learn about n-dimensional random vectors in general . The joint distribution function of a random vector (X,Y) is defined as

$$F(x,y) = P(X< x, Y< y) = P[(X,Y)\in(-\infty; x)\times(-\infty; y)],\ (x,y)\in(-\infty;+\infty)^2.$$

It has the following properties:

a) $0 \le F(x,y) \le 1$ for all pairs $(x,y) \in(-\infty;+\infty)^2$,

b) $\lim_{x\to-\infty} F(x,y) = F(-\infty,y) = \lim_{y\to-\infty} F(x,y) = F(x,-\infty) = 0$,

c) $\lim_{(x,y)\to(+\infty,+\infty)} F(x,y) = F(+\infty,+\infty) = 1$,

d) F(x,y) is non-decreasing and continuous on the left for each variable x and y.

A random vector (X,Y) is said to be discrete with a discrete joint distribution function, if both its constituents X and Y are discrete and therefore it assumes at most a countable number of values $(x,y) = (x_1,y_1), (x_2,y_2),\ldots$. The joint distribution of probabilities of a discrete random vector is the sequence $p(x,y) = P(X = x,Y = y) > 0$. We have:

a) $\sum_x \sum_y p(x,y) = 1$,

b) $F(x,y) = \sum_{u<x} \sum_{v<y} p(u,v)$ for all pairs $(x,y) \in(-\infty;+\infty)^2$,

c) $P\big((X,Y) \in M\big) = \sum_{(x,y)\in M} \sum p(x,y)$ for $M \subseteq (-\infty;+\infty)^2$.

A random vector (X,Y) is said to be continuous with a continuous joint distribution function if both its constituents X, Y are continuous. The joint density function of a continuous random vector is such a function f(x,y) that

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)dudv \quad \text{for all pairs } (x,y) \in(-\infty;+\infty)^2.$$

We have:

a) $\int_{-\infty-\infty}^{+\infty+\infty} f(x,y)dxdy = 1,$

b) $f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$ provided that the partial derivative exists,

c) $P((X,Y) \in M) = \int\int_M f(x,y)dxdy$ for $M \subseteq (-\infty;+\infty)^2$.

If, in a random vector (X,Y), we leave out its constituent X or Y (that is, we assume that this random variable takes on any value), we get marginal probability distributions for Y or X. For marginal distributions of probability, density functions, and distribution functions we have

a) $p_X(x) = \sum_y p(x,y), \quad p_Y(y) = \sum_x p(x,y)$ for a discrete random vector,

b) $f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx$ for a continuous random vector,

c) $F_X(x) = F(x,+\infty), \quad F_Y(y) = F(+\infty,y)$ for both a discrete and continuous random vector.

Further we use conditioned probability distributions for one random variable given that the other random variable takes on an arbitrary but fixed value. In this way we get the conditioned probability distribution of a random variable X or Y given that Y = y or X = x respectively. For conditioned distributions of probability $p_X(x/y) = P(X = x / Y = y)$ and the like, and conditioned density functions we have

a) $p_X(x/y) = \frac{p(x,y)}{p_Y(y)}, \quad p_Y(y/x) = \frac{p(x,y)}{p_X(x)}$ for a discrete random vector,

b) $f_X(x/y) = \frac{f(x,y)}{f_Y(y)}, \quad f_Y(y/x) = \frac{f(x,y)}{f_X(x)}$ for a continuous random vector.

Random variables X and Y are said to be independent if, for all pairs (x,y),

$p(x,y) = p_X(x)p_Y(y)$ or $f(x,y) = f_X(x)f_Y(y)$, and $F(x,y) = F_X(x)F_Y(y)$.

Numerical characteristics of random vectors

A numerical characteristic of a random vector expresses its important properties. The central tendency of a joint probability distribution is characterized by the centre (E(X), E(Y)) of a random vector (X,Y) where E(X) and E(Y) are the expected values of X and Y. We can write

a) $E(X)=\sum_x xp_X(x)=\sum_x\sum_y xp(x,y)$, $E(Y)=\sum_y yp_Y(y)=\sum_x\sum_y yp(x,y)$ for a discrete random vector,

b) $E(X)=\int_{-\infty}^{+\infty} xf_X(x)dx=\int_{-\infty-\infty}^{+\infty+\infty} xf(x,y)dxdy$, $E(Y)=\int_{-\infty}^{+\infty} yf_Y(y)dy=\int_{-\infty-\infty}^{+\infty+\infty} yf(x,y)dxdy$ for a continuous random vector, provided that the sums or integrals absolutely converge.

The relationship between the random variables X and Y of a random vector (X,Y) is given by their covariance

$$\operatorname{cov}(X,Y)=E\left[(x-E(X))(y-E(Y))\right].$$

The covariance has the following property:

a) $\operatorname{cov}(X,Y)=\sum_x\sum_y (x-E(X))(y-E(Y))p(x,y)=\sum_x\sum_y xyp(x,y)-E(X)E(Y)$ for a discrete random vector,

$$\operatorname{cov}(X,Y)=\int_{-\infty-\infty}^{+\infty+\infty}(x-E(X))(y-E(Y))f(x,y)dxdy=\int_{-\infty-\infty}^{+\infty+\infty}xyf(x,y)dxdy-E(X)E(Y)$$

for a continuous random vector, provided that the sums or integrals absolutely converge.

b) cov(X,Y) = cov(Y,X),

c) cov(X,X) = D(X),

d) D(X + Y) = D(X) + D(Y) + 2cov(X,Y),

e) X, Y independent $\Rightarrow$ cov (X,Y) = 0 and E(X,Y) = E(X)E(Y).

The covariances of a random vector (X,Y) can be displayed in the form of a symmetrical covariance matrix **cov**(X,Y) – see [1], [2], [3] and [4].

A measure of the linear independence between the random variables X and Y is their correlation coefficient (coefficient of correlation)

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)D(Y)}}.$$

The correlation coefficient has the following properties:

a) $\rho(X,Y) = \rho(Y,X)$,

b) $\rho(X,X) = \rho(Y,Y) = 1$,

c) $-1 \le \rho(X,Y) \le 1$,

d) $\rho(aX + b, cY + d) = \frac{ac}{|ac|}\rho(X,Y)$ for arbitrary real numbers a, b, c, d, $ac \neq 0$,

e) $Y = aX + b \Leftrightarrow |\rho(X,Y)| = 1$ where a, b are real numbers, $a \neq 0$,

f) X, Y independent $\Rightarrow \rho(X,Y) = 0$.

If $\rho(X,Y) = 0$, we say that the random variables X and Y are not correlated.

Independent random variables are not correlated but two random variables that are not correlated are not necessarily independent. However, if they are both normally distributed, they are also independent. The correlation coefficients of a random vector (X,Y) can be displayed in the form of a symmetrical correlation matrix $\rho(X,Y)$ - see [1], [2], [3] and [4].

Example 2.11

A discrete random vector (X,Y) is given by the below table:

y \ x	0	1	2	3
-1	2c	c	0	0
0	c	2c	c	0
1	0	0	2c	c

Calculate c, F(2;0), $p_X(1)$, $F_Y(1)$, $p_X(x/y)$ for (x,y) = (1;0), E(X), E(Y), D(X), D(Y), cov(X,Y), $\rho(X,Y)$ and determine if the random variables X,Y are independent.

S o l u t i o n:

$2c + c + ... + 2c + c = 1 \Rightarrow 10c = 1 \Rightarrow c = 0.1$;

$F(2;0) = p(0;-1) + p(1;-1) = 0.3$;

$p_X(1) = p(1;-1) + p(1;0) + p(1;1) = 0.1 + 0.2 + 0 = 0.3$;

$F_Y(1) = p_Y(-1) + p_Y(0) = 0.3 + 0.4 = 0.7$;

$p_X(1/0) = p(1;0)/p_Y(0) = 0.2/0.4 = 0.5$;

$E(X) = (0)(0.2) + (0)(0.1) + ... + (3)(0) + (3)(0.1) = 1.2$;

$E(Y) = (-1)(0.2) + (-1)(0.1) + ... + (1)(0.2) + (1)(0.1) = 0$;

$D(X) = (0^2)(0.2) + (0^2)(0.1) + ... + (3^2)(0) + (3^2)(0.1) – 1.2^2 = 2.4 - 1.44 = 0.96$;

$D(Y) = (-1)^2(0.2) + (-1)^2(0.1) + ... + 1^2(0.2) + 1^2(0.1) - 0^2 = 0.6 - 0 = 0.6$;

cov (X,Y) = (0)(-1)(0.2) + (1)(-1)(0.1) +...+ (2)(1)(0.2) + (3)(1)(0.1) – (1.2)(0) =
= 0.6 - 0 = 0.6 ;

$\rho(X,Y) = \dfrac{0.6}{\sqrt{(0.96)(0.6)}} \cong 0.79057 \neq 0$, which means that X, Y are not independent.

Some important probability distributions

Discrete probability distributions

a) The binomial probability distribution Bi(n, p) where n is a natural number and p is a real number, $0 < p < 1$:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n;$$

$$E(X) = np; \quad D(X) = np(1 - p); \quad (n + 1)p - 1 \leq \bar{x} \leq (n + 1)p.$$

This is the probability distribution of a random variable that expresses the number of occurrences of an observed event in the sequence of n mutually independent trials (such as the number x of defective products out of a total of n products if p is the probability of a defective product being manufactured). This probability distribution may be employed when performing a random sample with replacement such as checking n products from a supply and replacing each product after it has been checked. For $np(1 - p) > 9$ the binomial distribution may be approximated by a normal distribution where $\mu = np$, $\sigma^2 = np(1 - p)$. For $p < 0{,}1$ and $n > 30$ we can also approximate this probability distribution by the Poisson probability distribution with $\lambda = np$. The graphs of the distribution of probability and the distribution function of the binomial distribution for $n = 3$ and $p = 0{,}5$ are shown in Fig. 2.1.

E x a m p l e 2.12

A series of 50 products contains 5 defective ones. Three products are selected at random from the series. The number of defective products among those selected is a random variable X. Determine its type of probability distribution, its distribution of probabilities p(x), expected value E(X), variance D(X), standard deviation $\sigma(X)$, median $x_{0.5}$, mode $\hat{x}$, and the probability $P(1 < X \leq 3)$. Each product selected is replaced so that this is the so-called random sample with replacement.

S o l u t i o n:

The random variable X has the binomial distribution Bi(n,p) with n = 3 and p = 5/50 = = 0.1. X takes on the values x = 0, 1, 2, 3. The distribution of probabilities

$$p(x) = \binom{3}{x} 0{,}1^x 0{,}9^{3-x} \quad \text{for } x = 0, 1, 2, 3.$$

The expected value E(X) = np = (3)(0.1) = 0.3,

the variance D(X) = np(1 - p) = (3)(0.1)(0.9) = 0.27,

the standard deviation $\sigma(X) = \sqrt{D(X)} = \sqrt{0.27} \cong 0.51962$,

the median $x_{0.5} = 0$, since p(0) = 0.729,

the mode $\hat{x} = 0$, since (n + 1)p -1 = -0.6 and (n + 1)p = 0.4,

$P(1 < X \leq 3) = p(2) + p(3) = 0.027 + 0.001 = 0.028$.

b) The hypergeometric distribution H(N,M,n) where N, M, and n are natural numbers, $1 \leq n \leq N$, $1 \leq M \leq N$:

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = \max\{0, M - N + n\}, \ldots, \min\{M, N\};$$

$$E(X) = n\frac{M}{N}; \quad D(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}; \quad a - 1 \leq \bar{x} \leq a \text{ where } a = \frac{(M+1)(n+1)}{N+2}.$$

This probability distribution describes the so-called random sample without replacement where, for example, N is the total number of products to be checked, M is the number of defective products among them, n products are selected at random without replacement, and x is the number of defective products among those selected. For n/N < 0,1, the hypergeometric probability distribution may be approximated by the binomial probability distribution with p = M/N, or, for n/N < 0,1, M/N < 0,1 and n > 30, by the Poisson probability distribution with λ = nM/N.

Example 2.13

Among a total of 50 products there are 5 defective ones. Three products are drawn at random out of the total. The number of defective products among those selected is a random variable X. Determine the type of its probability distribution, its distribution of

probabilities p(x), expected value E(X), variance D(X), standard deviation $\sigma(X)$, median $x_{0,5}$, mode $\hat{x}$, and $P(1 < X \leq 3)$. Assume, as opposed to Example 2.12, that a product is not replaced once it has been selected so that this is a random sample without replacement.

S o l u t i o n:

The random selection X has the probability distribution H(N,M,n) with N = 50, M = 5 and n = 3. X takes on the values x = 0, 1, 2, and 3. The distribution of probabilities is given by

$$p(x) = \frac{\binom{5}{x}\binom{45}{3-x}}{\binom{50}{3}} \quad \text{for } x = 0, 1, 2, 3.$$

The expected value $E(X) = n\frac{M}{N} = (3)(0.1) = 0.3$,

the variance $D(X) = D(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1} = (3)(0.1)(0.9)\,(47/49) \approx 0.25898$,

the standard deviation $\sigma(X) = \sqrt{D(X)} = \sqrt{0.25898} \cong 0.50890$,

the median $x_{0.5} = 0$ since max $p(x) = p(0) \cong 0.724$,

the mode $\hat{x} = 0$ since $a = \frac{(M+1)(n+1)}{N+2} \cong 0.46154$, $a - 1 \cong -0.53846$,

$P(1 < X \leq 3) = p(2) + p(3) \cong 0.023 + 0.0005 = 0.0235$.

c) The <u>Poisson probability distribution</u> Po(λ) where λ is a real number, $\lambda > 0$:

$$p(x) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0, 1, \ldots; \quad E(X) = \lambda; \quad D(X) = \lambda; \quad \lambda - 1 \leq \hat{x} \leq \lambda.$$

This probability distribution is usually used to determine the probability of a number of occurrences of an observed event within a time interval (number of failures, accidents, disasters, defective products and the like) with a small probability of occurrence.

E x a m p l e 2.14

On the average, three customers enter a shop within a given minute. Determine the appropriate type of probability distribution of a random variable that expresses the number of customers that enter the shop within a given minute, the expected number of customers, the variance of this number, and the most likely number of customers that enter the shop within a given minute. Next calculate the probability that within that minute a) exactly one customer enters the shop, b) at least one customer enters the shop.

S o l u t i o n:

If we approximate the expected value of customers that enter the shop within one given minute by their average number, we can assume that the random variable X has the Poisson probability distribution Po(λ) with the distribution of probabilities given by

$$p(x) = \frac{3^x}{x!} e^{-3}, \quad x = 0, 1, \ldots$$

The expected value $E(X) = \lambda = 3$,

The variance $D(X) = \lambda = 3$,

for the mode we have $\lambda - 1 \leq \hat{x} \leq \lambda$, which yields $\hat{x} = 2$ a 3,

$$P(X = 1) = p(1) = \frac{3^1}{1!} e^{-3} \cong 0.14936,$$

$$P(X \geq 1) = p(1) + p(2) + \ldots = 1 - p(0) = 1 - \frac{3^0}{0!} e^{-3} \cong 1 - 0.04979 = 0.95021.$$

Information on further numerical characteristics of the above discrete probability distributions and multi-dimensional probability distributions can be found in [1], [2], [3] and [4].

Continuous probability distributions

a) The uniform probability distribution $R(a, b)$ where $a < b$ are real numbers:

$$f(x) = \frac{1}{b-a} \quad \text{for } x \in \langle a;b \rangle,$$
$$= 0 \qquad \text{for } x \notin \langle a;b \rangle,$$

$$F(x) = 0 \qquad \text{for } x \in (-\infty;a),$$
$$= \frac{x-a}{b-a} \quad \text{for } x \in \langle a;b \rangle,$$
$$= 1 \qquad \text{for } x \in (b;+\infty),$$

$$E(X) = x_{0.5} = \frac{a+b}{2}; \quad D(X) = \frac{(b-a)^2}{12} \ .$$

The graphs of the density function and the distribution function for a = -1 and b = 2 are shown in Fig. 2.3. This probability distribution is mostly used to simulate real processes, in numerical calculations to implement the so-called Monte Carlo method on a computer, and for calculations using the so-called geometric probability.

E x a m p l e 2.15

An optical cable of a length of 500 m may be disrupted at any distance from its beginning. The probability of the random event that the cable will be disrupted in a given section is in direct proportion to the length of the section and is independent of its position. Determine the probability distribution of the random event X expressing the distance of a disruption from the beginning of the cable, the density function, the basic numerical characteristics, and the probability that the cable will be disrupted in the section beginning at 300 m and ending at 400 m.

S o l u t i o n:

Random variable X has the probability distribution R(a, b) with a = 0 and b = 500.

The density function is given by

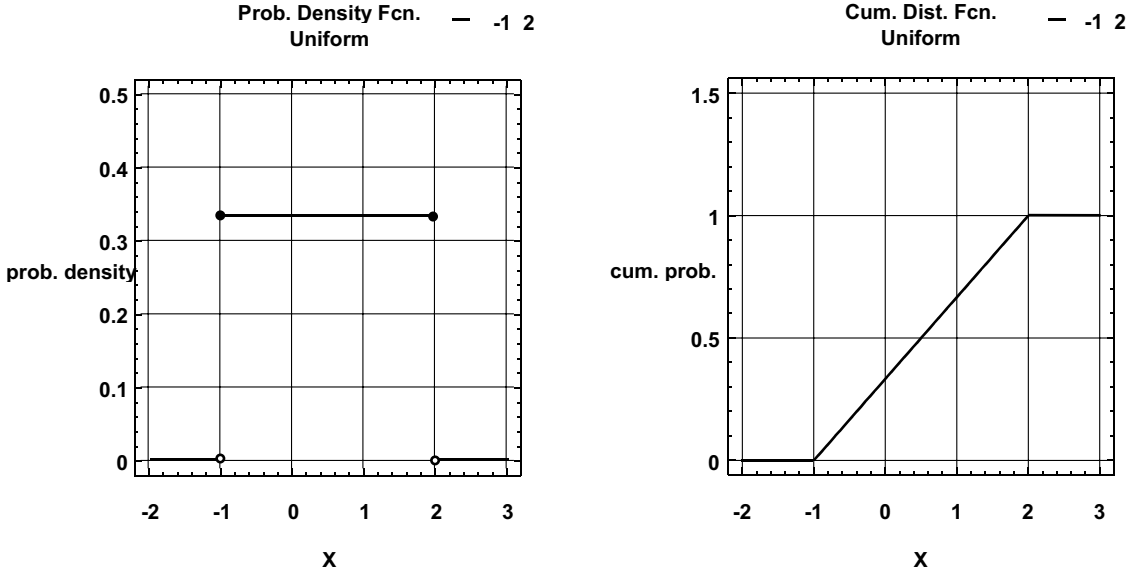
$$f(x) = \frac{1}{500} \quad \text{for } x \in \langle 0;500 \rangle,$$
$$= 0 \qquad \text{for } x \notin \langle 0;500 \rangle.$$

The expected distance and the median $E(X) = x_{0.5} = \frac{0+500}{2}$ = 250 m,

the variance $D(X) = \frac{(500-0)^2}{12} \cong 20\,833.3\ \text{m}^2$,

the standard deviation $\sigma(X) = \sqrt{D(X)} \cong \sqrt{20\,833.3} \cong$ 144.34 m,

the probability $P(300 \le X \le 400) = F(400) - F(300) = \frac{400}{500} - \frac{300}{500}$ = 0.2.



(a) (b)

Fig. 2.3 The graphs of the density function (a) and the distribution function (b) of a uniform probability distribution.

b) The <u>normal probability distribution</u> $N(\mu, \sigma^2)$ where μ, σ^2 are real numbers, $\sigma^2 > 0$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right], \quad x \in (-\infty, +\infty);$$

$$E(X) = x_{0.5} = \hat{x} = \mu, \quad D(X) = \sigma^2.$$

For μ = 1 and σ = 1, the graphs of the density function and the distribution function are plotted in Fig. 2.4. This is the most widely used probability distribution sometimes also called the <u>Gauss probability distribution</u> applied to random variables that can be

interpreted as the result of adding up a multitude independent influences (such as the error of a measurement, the size deviation of a product and the like). Using the transformation

$$U = \frac{X - \mu}{\sigma}$$

we get the standard normal probability distribution N(0;1) whose distribution function $\Phi(u)$ is tabulated (see Table T1) or its values are approximated. We have

$$\Phi(-u) = 1 - \Phi(u).$$

For a random variable X with the normal probability distribution $N(\mu, \sigma^2)$ we have

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

and, for example, $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \cong 0.9973$ (the so-called three-sigma rule).

E x a m p l e 2.16

What is the probability that a random variable X with the normal probability distribution N(20, 16) will take on a value a) less than 16, b) greater than 20, c) from 12 to 28, d) less than 12 or greater than 28 ?

S o l u t i o n:

Using the formula $F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$ and table T1 we get

a) $P(X < 16) = F(16) = \Phi((16 - 20) / 4) = \Phi(-1) = 1 - \Phi(1) = 1 - 0.84135 = 0.15865$;

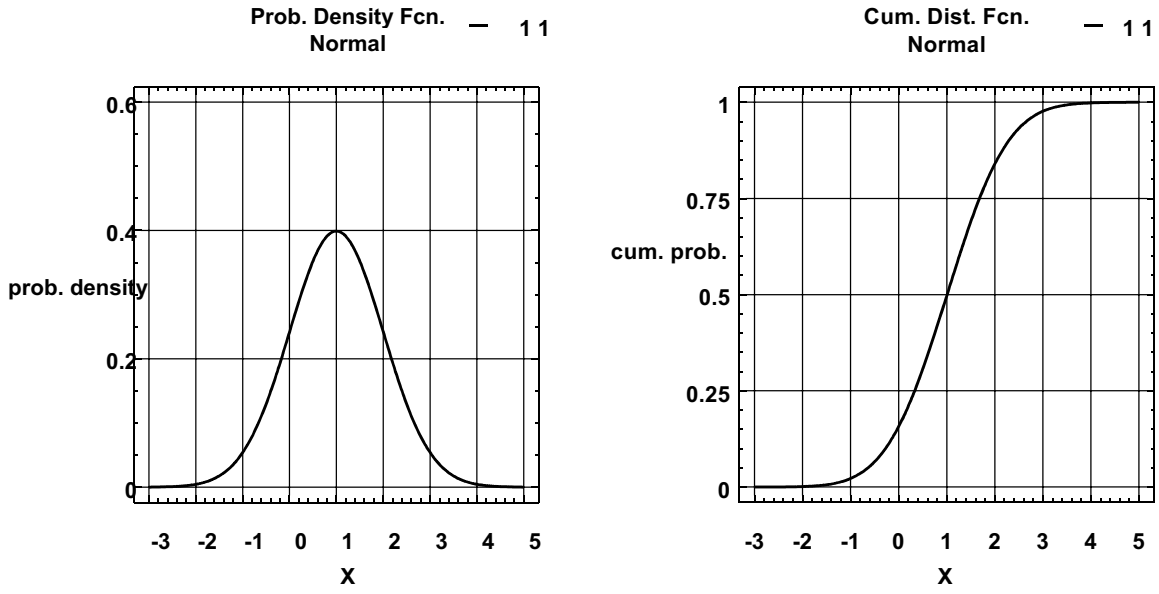
b) $P(X > 20) = 1 - P(X \leq 20) = 1 - F(20) = 1 - \Phi((20 - 20) / 4) = 1 - \Phi(0) =$

$= 1 - 0.5 = 0.5$;

c) $P(12 \leq X \leq 28) = F(28) - F(12) = \Phi((28 - 20) / 4) - \Phi((12 - 20) / 4) = \Phi(2) - \Phi(-2)$

$= \Phi(2) - (1 - \Phi(2)) = 2\Phi(2) - 1 = (2)(0.97725) - 1 = 0.9545$;

d) $P((X < 12) \vee (X > 28)) = 1 - P(12 \leq X \leq 28) = 1 - 0.9545 = 0.0455$.



(a) (b)

Fig. 2.4 The *g*raphs of the density function (a) and the distribution function (b) of a normal probability distribution

Information on further one-dimensional and multi-dimensional continuous probability distributions can be obtained in [1], [2], [3] and [4].

2. DESCRIPTIVE STATISTICS

Basic notions

When performing statistical analysis we deal with events and processes which occur on a mass scale and can be found in a large set of individual objects such as products or persons. We call this set a population. The objects under investigation are called statistical items and we observe them focussing on certain properties - statistical variables such as parameters which have outcomes or values that we observe.

By the type of outcomes variables are either quantitative, with numerical outcomes such as mass, length, strength, price, service life and the like, or qualitative, which are not numeric and can only be expressed by words such as colour, quality class, operation conditions, form, etc. If only one property is observed we speak of a univariate variable, if more properties are ascertained at the same time, we say that we observe a multivariate variable.

Quantitative variables are either discrete if they only take on discrete values (number of defective products, number of faults, number of pieces, etc.) or continuous if they assume all the values of an interval of real numbers (size of a product, time to failure, price index and the like).

Qualitative variables are either ordinal if there is a point in ordering their outcomes expressed in words such as quality classes or classifications or nominal there is no point in ordering them such as colour, form, or suppliers.

Statistical methods are based on the fact that information on the population is not taken from all its elements but rather from a subset of the population defined by taking a sample. This is due to certain limitations such as accessibility of all the statistical items, large size of the population, the way the information is obtained (service life tests, wear tests, etc.), excessive costs of statistical surveys and others. The number of statistical items in a sample is called the size of a sample. If the size

of a sample less than 30 to 50, we say that the sample is small, if the size is several hundreds or thousands we say that the sample is large. This classification is of course arbitrary and may differ depending on circumstances. A sample should be representative, which means that it should provide information without any limitations, and homogeneous (not affected by other factors). This can seldom be achieved with sufficient confidence and that is the reason why we usually select the items of a sample at random, even at the risk of the information about the whole population contained in the sample being biased.

The following is a classification of samples according to the way they have been selected:

- without replacement (each item may only be selected once),
- with replacement (items may be selected several times),
- intentional (typical items are selected),
- regional (the population is first divided into partitions and each partition supplies part of the sample),
- systematic or mechanical (the population is thought to be ordered and every k-th item is selected).

The outcomes of a variable observed or measured in the statistical items of a sample of size n are called sample data of size n. A univariate variable X provides univariate sample data $(x_1,...,x_n)$ where x_i, i = 1,..., n , is the outcome observed in the i-th variable. Similarly, a bivariate variable (X,Y) yields bivariate sample data $((x_1,y_1),...,(x_n, y_n))$ etc.

Processing univariate sample data with a quantitative variable

The primary sample data or raw scores $(x_1,...,x_n)$ of size n are called ungrouped sample data. The outcomes x_i may be listed in order of numerical magnitude, which yields an array of sample data $(x_{(1)},...,x_{(n)})$ where $x_{(i)} \leq x_{(i+1)}$ for

every i, and $x_{(1)} = x_{min}$, $x_{(n)} = x_{max}$. The interval $\langle x_{min}; x_{max} \rangle$ is called the domain of sample data and the number $x_{max} - x_{min}$ is called the range of sample data.

If sample data are very large or if they are to be further processed (some graphical representations or application of mathematical statistical methods) the raw scores are grouped. Sample data are grouped by partitioning the domain into a series of m non-overlapping intervals (usually left-open and right-closed), the so-called classes usually of the same width h. Each class is represented by a pair (x_j^*, f_j) where x_j^* is the midpoint, $x_j^* \le x_{j+1}^*$ and f_j is the frequency (or absolute frequency) of class j, $j = 1,...,m$. The absolute frequency f_j is defined as the number of raw scores that lie in class j. The number f_j/n is called relative frequency and sometimes it is also shown as a percentage. Obviously, we have $\sum_{j=1}^{m} f_j = n$.

The number of classes m is usually selected to be close to $1 + 3.3 \log n$ or $\sqrt{n}$. The length of a class is then $h \doteq (x_{max} - x_{min})/m$ and it should correspond to the accuracy of the outcomes x_i and for the midpoint x_j^* to be a rounded number. For a quantitative variable the midpoint is chosen as one of the outcomes.

The number $F_j = \sum_{k=1}^{j} f_k$ is called cumulative absolute frequency, the number F_j / n is called cumulative relative frequency, $j = 1,...,m$, and can be shown as a percentage as well. Obviously, we have $F_m = n$.

Grouped sample data is shown in a table, which is called frequency distribution for different types of frequency:

x_j^*	x_1^* . . . x_m^*
f_j	f_1 . . . f_m

The properties of sample data are, in a concentrated form, expressed by different measures.

The basic measures of central tendency:

The arithmetic mean (average)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{for ungrouped data,}$$

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} f_j x_j^* \qquad \text{for grouped data.}$$

Mathematical properties:

- $y = ax + b \Rightarrow \bar{y} = a\bar{x} + b$ for arbitrary constants a, b,
- $\overline{x + y} = \bar{x} + \bar{y}$,
- $x_{min} \le \bar{x} \le x_{max}$,
- $\bar{x}$ has the same unit of measurement as variable X.

Sometimes a weighted arithmetic mean

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

is used where $w_i \ge 0$ are the weights of outcomes x_i, which reflect their significance such as accuracy.

The median

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{for odd } n, \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{for even } n. \end{cases}$$

Mathematical properties:

- $y = ax + b \Rightarrow \tilde{y} = a\tilde{x} + b$ for constants a, b,
- $x_{min} \le \tilde{x} \le x_{max}$,
- $\tilde{x}$ has the same unit of measurement as variable X.

The median divides the sample data into "the upper part" and "the lower part" of outcomes x_i. This is a robust measure, which, as compared to the arithmetic mean, is little affected by extreme values. Sometimes a suitable approximation is used to calculate the median (for example if the data is grouped).

The mode $\hat{x}$ is the number whose neighbourhood contains the most outcomes x_i, or the middle x_j^* of the class with the largest absolute frequency f_j. The mode has the same unit of measurement as the variable X and, if it is needed, a suitable approximation is used for calculating it.

The basic measures of variation are the following:

Variance (dispersion)

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 \qquad \text{for ungrouped data,}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{m}f_j(x_j^* - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{m}f_j x_j^{*2}\right) - \bar{x}^2 \qquad \text{for grouped data.}$$

Sometimes we write $s^2(x)$. Mathematical properties:

- $s^2 \ge 0$,
- $y = ax + b \Rightarrow s^2(y) = a^2 s^2(x)$ for constants a, b,
- the unit of measurement for s^2 equals to the square of the unit of measurement for variable X.

The more the outcomes of a variable X are scattered, the more greater is its variance and vice versa. For calculations, sometimes an alternative formula for

variance is used by replacing $\frac{1}{n}$ by $\frac{1}{n-1}$. This variance calculated by this new formula equals to $\frac{n}{n-1}s^2 \rangle s^2$.

Standard deviation

$$s = \sqrt{s^2} .$$

Sometimes written as s(x). Mathematical properties:

- $s \geq 0$,
- $y = ax + b \Rightarrow s(y) = |a|s(x)$ for constants a, b,
- s has the same unit of measurement as variable X.

The more the outcomes of a variable X are scattered, the greater is its standard deviation and vice versa.

Coefficient of variation

$$v = \frac{s}{\overline{x}} .$$

Sometimes written as v(x). This is a relative measure of the variability of variable X and it can also be shown as a percentage. It is only meaningful for a variable X with only positive or only negative outcomes. It holds:

- $v(ax) = v(x)$ for any constant $a \neq 0$,
- v(x) is a dimensionless number.

The basic measure of symmetry for sample data is the coefficient of skewness

$$A = \frac{\frac{1}{n}\sum_{i=1}^{m}(x_i - \overline{x})^3}{s^3} \quad \text{for ungrouped data,}$$

$$A = \frac{\frac{1}{n}\sum_{j=1}^{m} f_j(x_j^* - \overline{x})^3}{s^3} \quad \text{for grouped data.}$$

Mathematical properties:

- $A \rangle 0$ ... the majority of outcomes x_i are less than $\overline{x}$,
- $A = 0$ ... outcomes x_i are symmetric with respect to $\overline{x}$,
- $A \langle 0$ ... the majority of outcomes x_i are greater than $\overline{x}$,
- $y = ax + b \Rightarrow A(y) = \frac{a}{|a|} A(x)$ for any constant $a \neq 0$,
- A is a dimensionless number.

A number of further sample data measures exist. Sometimes the geometric mean

$$\overline{x}_g = \sqrt[n]{x_1 ... x_n}$$

is employed instead of the arithmetic one for some variables which describe ratios such as volume and price indices, interest rates and the like. In special cases we use the harmonic mean

$$\overline{x}_h = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}.$$

Much valuable information on sample data can be obtained from their graphical representation.

For univariate ungrouped or grouped sample data we can use box charts see Fig 2.1 where the box contains the middle part of grouped data (about one half of all the outcomes) while about a quarter of the data is placed on either side of the box. The line on the left (on the right) corresponds to the so-called lower quartile (upper quartile) and the perpendicular line in the middle is in the place of the median. The height of the box is proportional to the size of the data and the line segments on both sides represent acceptable domains for the above quarters of the data. Outcomes beyond these line segments are considered as suspicious or extremely deviated. There are also other modifications of this chart and other graphical tools.

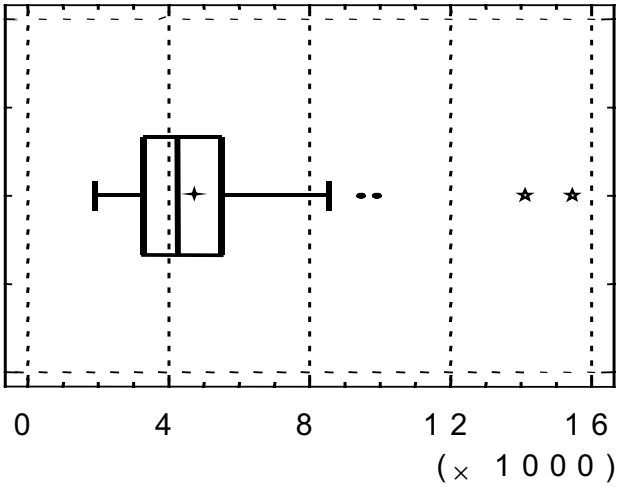


Fig. 2.1

Two other types of charts are frequently used for univariate sorted data: Histograms - see Fig. 2.2 - a histogram is a system of bars in Cartesian co-ordinates where the bases if the bars are the classes and their heights correspond to the absolute (relative, cumulative, etc.) frequencies. Frequency polygons - see Fig. 2.3 - a frequency polygon is a broken line in Cartesian co-ordinates connecting points whose abscissas coincide with the midpoints (or with upper limits) of classes and their ordinates are proportional to the frequency.

Example 2.1

A total of 10 rollers have been measured with the following results: 5.38; 5.36; 5.35; 5.40; 5.41; 5.34; 5.29; 5.43; 5.42; 5.32. Determine the size, domain, and range, arithmetic mean, variance, standard deviation, coefficient of variation, and median of the sample data.

Solution:

The data size is n = 10, the domain is <5.29; 5.43> mm and the range is 5.43 – 5.29 = 0.14 mm.

$\bar{x}$ = (5.38 + ...+ 5.32)/10 = 53.70/10 = 5.37 mm,

$s^2 = (5.38^2 + ...+ 5.32^2)/10 - 5.37^2 = 288.388/10 - 28.8369 = 0.0019$ mm^2,

$s = \sqrt{0.0019} \cong 0.0435889894 \cong 0.044$ mm,

$v = 0.0435889894/5.37 \cong 0.00811713 \cong 0.8117\ \%$,

$\tilde{x} = (5.36 + 5.38)/2 = 5.37$ mm.

E x a m p l e 2.2

When checking the volume of beverage in a bottle for a sample of 50 bottles, the following deviations (in ml) from the values stated on labels have been found:

1.2	2.1	1.7	0.9	0.3	2.0	-1.3	-0.1	3.2	2.8
0.8	4.4	2.9	1.2	0.0	-2.3	1.2	0.9	2.3	- 0.2
0.1	1.9	-1.9	-0.2	-1.3	1.5	0.5	2.0	-1.3	3.7
0.9	1.0	0.4	1.9	1.4	-1.3	1.6	1.4	3.1	-0.1
1.8	0.0	4.1	1.3	3.0	0.4	3.8	-0.8	3.1	0.9

Group the data, set up a frequency distribution, and design a graphical representation. Calculate $\bar{x}$, s^2, s, $\hat{x}$.

S o l u t i o n:

The size of the data is n = 50; x_{min} = - 2.3 ml and x_{max} = 4.4 ml, which means that the domain is <-2.3; 4.4> ml, the range being 4.4 - (-2.3) = 6.7 ml. We choose the number of classes to be m = 7 and the class width h = 1 (approximation of 6.7/7). The selection of classes, their midpoints, the grouping of the data and the calculation of absolute and relative frequencies are shown in the table below (*###* stands for 5 outcomes):

j	class	x_j^*	classification	f_j	F_j
1	-2.5; -1.5	-2	//	2	2
2	-1.5; -0.5	-1	*###*	5	7
3	-0.5; 0.5	0	*###* *###* /	11	18
4	0.5; 1.5	1	*###* *###* ///	13	31
5	1.5; 2.5	2	*###* ////	9	40
6	2.5; 3.5	3	*###* /	6	46
7	3.5; 4.5	4	////	4	50

Histograms and polygons for this sample data are shown in Fig. 2.2 and 2.3. Further calculations are, for the sake of clarity, shown in the following table:

j	x_j^*	f_j	$f_j x_j^*$	$f_j x_j^{*2}$
1	-2	2	-4	8
2	-1	5	-5	5
3	0	11	0	0
4	1	13	13	13
5	2	9	18	36
6	3	6	18	54
7	4	4	16	64
$\sum$	—	50	56	180

Using the table we get:

$\bar{x}$ = 56/50 = 1.12 ml, $s^2 = 180/50 - 1.12^2 = 2.3456$ ml^2,

$s = \sqrt{2.3456} \cong 1.532$ ml, $\hat{x}$ = 1 ml (the midpoint of the class with the greatest frequency).

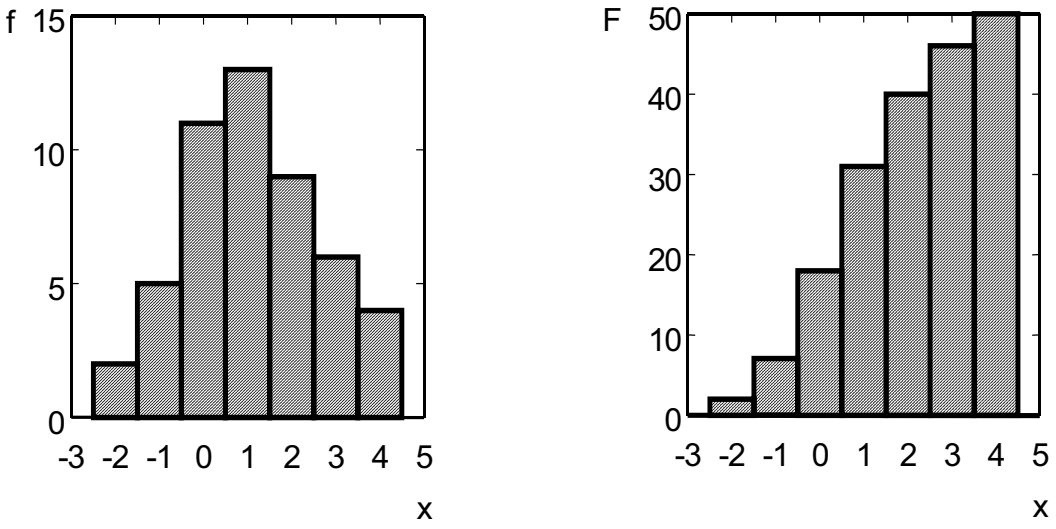


Fig. 2.2

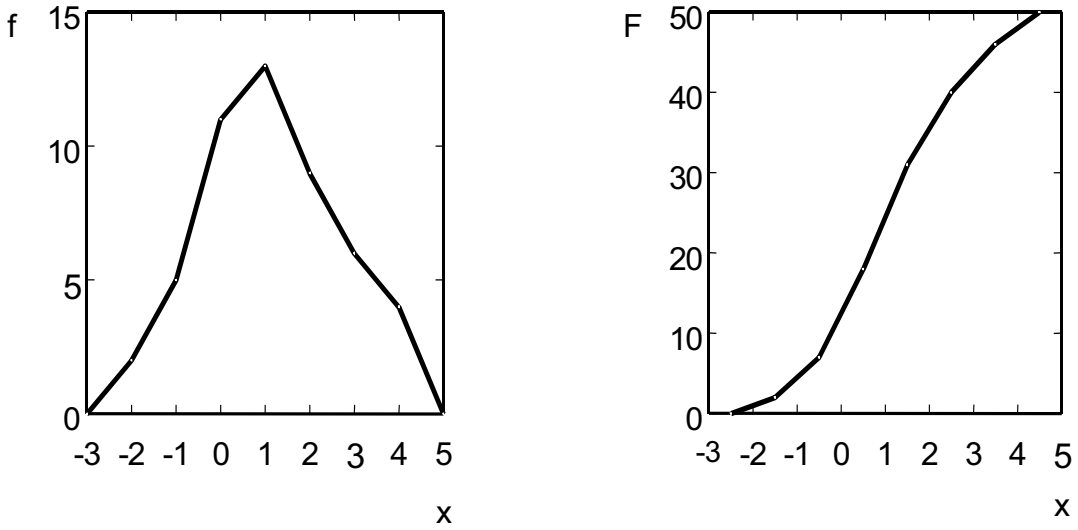


Fig. 2.3

For univariate sorted sample data with a discrete variable, usually the following charts are used. Bar chart - see Fig. 2.4 - is similar to a histogram but there are gaps between the bars and sometimes the bars are positioned horizontally. Pie chart - see Fig. 2.5 - is a circle divided into sections whose perimeter corresponds to the class frequencies. Some of the sections may be shifted in the upward or downward direction. Different colours or types of hatching are used in these charts to make selected pieces of information more prominent and sometimes the charts are further geometrically and artistically modified for better presentation.

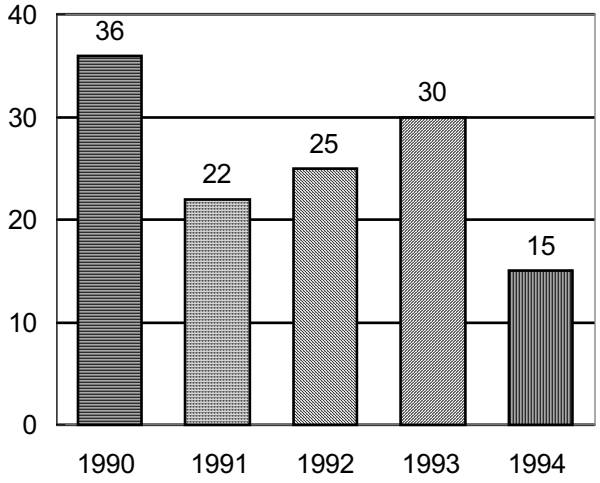


Fig. 2.4

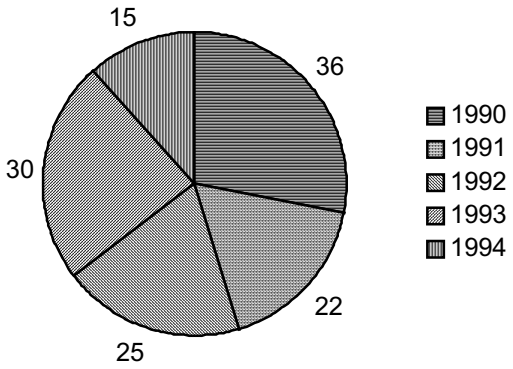


Fig. 2.5

Processing bivariate sample data with quantitative variables

The raw scores $((x_1, y_1),..., (x_n, y_n))$ obtained are called ungrouped data. If we leave out the first and the second value in each pair, we get two sets of univariate sample data $(x_1,..., x_n)$ and $(y_1,..., y_n)$ respectively. Processing these sets we obtain measures like $\overline{x}$, $\overline{y}$, $s^2(x)$, $s^2(y)$ etc.

We can group bivariate sample data by grouping each of the sets of univariate sample data $(x_1,...,x_n)$ and $(y_1,...,y_n)$ where for each data set the number of classes or the widths of classes may be different. In this way we obtain bivariate classes with middles (x_j^*, y_k^*) and absolute frequencies f_{jk}, $j = 1,...,m_1$ and $k = 1,...,m_2$. The relative frequencies f_{jk}/n, cumulative frequencies etc. are calculated similarly.

Grouped bivariate sample data can be summarized in a crosstabulation or contingency table for different types of frequency:

x_j^* \ y_k^*	y_1^* . . . $y_{m_2}^*$	f_{xj}
x_1^*	f_{11} . . . f_{1m_2}	f_{x1}
. . .	. . .	. . .
$x_{m_1}^*$	$f_{m_1 1}$ . . . $f_{m_1 m_2}$	$f_{x m_1}$
f_{yk}	f_{y1} . . . $f_{y m_2}$	n

The numbers f_{xj} and f_{yk} are marginal frequencies and the following formulas hold:

$$f_{xj} = \sum_{k=1}^{m_2} f_{jk} \quad , \quad f_{yk} = \sum_{j=1}^{m_1} f_{jk} \quad , \quad \sum_{j=1}^{m_1} f_{xj} = \sum_{k=1}^{m_2} f_{yk} = \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} f_{jk} = n .$$

For grouped data (x_j^*, f_{xj}), $j = 1,...,m_1$, and (y_k^*, f_{yk}), $k = 1,...,m_2$, we obtain measures like $\bar{x}$, $\bar{y}$, $s^2(x)$, $s^2(y)$ etc.

The correlation coefficient is a measure of the dependence of variables X and Y

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{xy}}{s(x)s(y)}$$ for ungrouped data,

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}f_{jk}(x_j^* - \bar{x})(y_k^* - \bar{y})}{s(x)s(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}f_{jk}x_j^* y_k^* - \overline{xy}}{s(x)s(y)}$$ for grouped data.

The numerators in all fractions define the so-called covariance cov. Sometimes we write r(x,y) and cov(x,y).

Mathematical properties:

- $r(ax+b, cy+d) = \frac{ac}{|ac|}r(x,y)$ for constants a, b, c, d, $a \neq 0$, $c \neq 0$,
- $-1 \leq r \leq 1$,
- y = ax + b (a, b constants), is equivalent to $r = \pm 1$,
- r is a dimensionless number.

The correlation coefficient is a measure of only linear-type dependence between X and Y. The more its value nears 1 or -1, the closer the dependence is and the better points (x_i, y_i) can be fitted with a straight line. Its positive or negative value corresponds to a direct or an indirect linear dependency. A value close to 0 means either that the dependency is not linear or that X and Y are independent.

To graphically represent ungrouped bivariate sample data a scatter diagram may be used - see Fig. 2.6 while grouped bivariate data may be shown in a 3D–histogram - see Fig. 2.7 or in a 3D-bar chart for discrete variables X,Y.

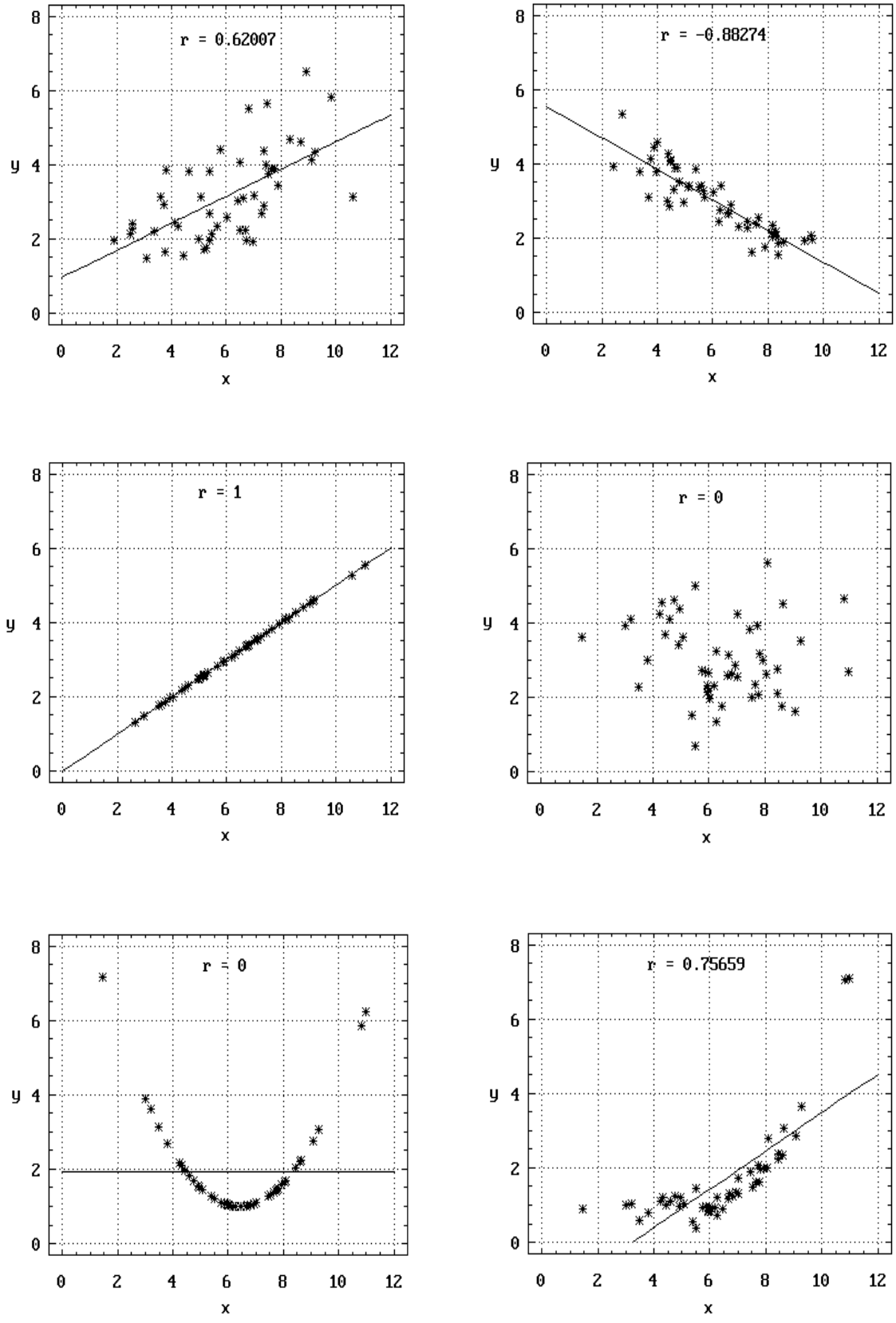


Fig. 2.6

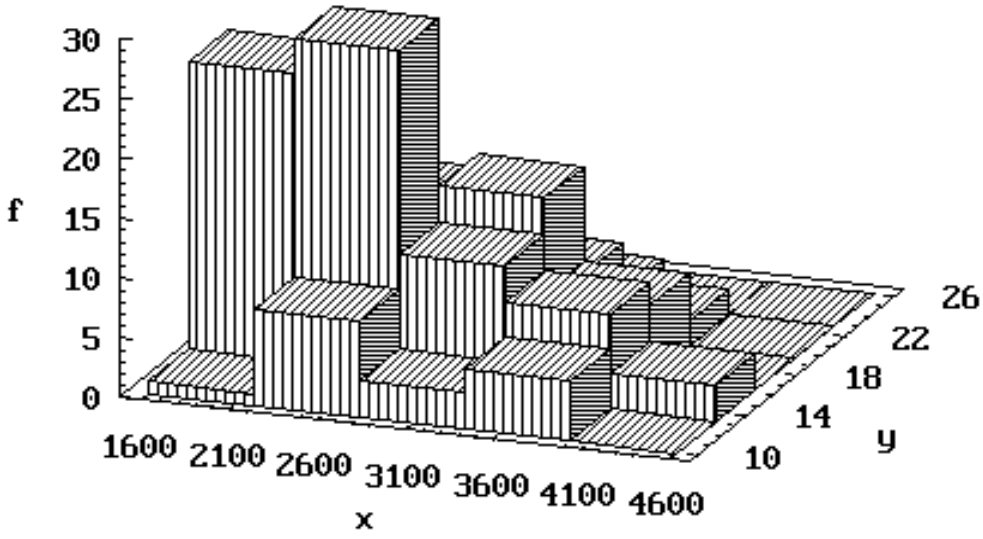


Fig.2.7

E x a m p l e 2.3

A survey of costs x (CZK) and prices y (CZK) of an identical product manufactured by ten different producers has yielded the following bivariate sample data (x_i,y_i):

(30.18; 50.26), (30.19; 50.23), (30.21; 50.27), (30.22; 50.25), (30.25; 50.22), (30.26; 50.32), (30.26; 50.33), (30.28; 50.29), (30.30; 50.37), (30.33; 50.42). Calculate $\bar{x}$, $\bar{y}$, $s^2(x)$, $s^2(y)$, s(x), s(y), cov, r.

S o l u t i o n:

As the data size is only n = 10 the data need not be grouped. Using the above formulas we get:

$\bar{x}$ = (30.18 + ...+ 30.33)/10 = 30.248 CZK,

$\bar{y}$ = (50.26 + ...+ 50.42)/10 = 50.296 CZK,

$s^2(x) = (0.18^2 + ...+ 30.33^2)/10 - 30.248^2 = 0.002096$ CZK^2,

$s^2(y) = (50.26^2 + ...+ 50.42^2)/10 - 50.296^2 = 0.003684$ CZK^2,

$s(x) = \sqrt{0.002096} = 0.0457821$ CZK $\cong 0.0458$ CZK,

$s(y) = \sqrt{0.003684} = 0.0606960$ CZK $\cong 0.0607$ CZK,

cov = [(30.18)(50.26) + ...+(30.33)(50.42)]/10 – (30.248)(50.296) = 0.002292 CZK^2,

r = 0.002292/[(0.0457821)(0.0606960)] = 0.82481996263 ≅ 0.82482.

Judging by the value of the correlation coefficient it may be assumed that there is a dependency between the variables that is fairly close to linear.

Processing sample data with qualitative variables

Sample data $(x_1,...,x_n)$ obtained are represented by a frequency distribution where x_j^* are all possible values of variable X expressed in words and f_j are the frequencies of these values in the original data, $j = 1,...,m$. Measures are used only exceptionally (variability). Bar charts and pie charts are mainly used for graphical representation.

Sample data $((x_1,y_1),...,(x_n,y_n))$ obtained can be grouped and summarized in a crosstabulation or a contingency table as with quantitative variables where (x_j^*, y_k^*) are pairs representing all combinations of the outcomes of variables (X,Y) and f_{jk} are the frequencies of these outcomes for $j = 1,...,m_1$ and $k = 1,...,m_2$. Out of various measures the most frequently used are measures of the dependence of X and Y. 3D - bar charts are used to graphically represent these data.

Exercises

Exercise 2.4

A total of ten metal parts have been machined, for each part the wasted material has been weighed and the corresponding percentage calculated. The following data have been obtained: 40.60; 40,29; 37.51; 38,90; 38.13; 38,15; 34.81; 37,00; 39.95; 40.43. Calculate $\bar{x}$, s^2, s, and v.

Result: $\bar{x}$ = 38.577 %, $s^2 \cong 3.0648$ %^{2}, $s \cong 1.751$ %, $v \cong 0.0454 = 4.54$ %

E x e r c i s e 2.5

Calculate the domain, range, arithmetic mean, variance, and standard deviation for the following data describing the precipitation (the amount of rain and snow fallen in mm) in Brno from 1941 to 1960: 718.5; 492.3; 431.5; 540.5; 514.7; 584.0; 385.0; 532.0; 531.0; 578.3; 551.9; 613.6; 476.0; 661.3; 518.0; 508.5; 488.7; 494.9; 554.6; 673.5.

R e s u l t: <385.0; 718.5> mm; the range is 333.5 mm; $\overline{x}$ = 542.44 mm;

$s^2 \cong 6127.5\ mm^2$; $s \cong 78.28$ mm

E x e r c i s e 2.6

In a survey conducted in 100 households selected at random the following numbers of household members have been ascertained:

2 2 3 5 3 3 2 7 4 7 2 3 5 6 4 4 4 2 4 6 5 3 4 5 5

4 5 7 4 3 4 2 4 4 4 4 4 4 3 2 4 3 3 3 4 2 3 4 2 3

3 3 4 3 5 9 3 3 4 8 5 4 5 3 3 4 3 3 3 4 5 2 3 7 3

5 5 1 4 4 5 3 3 4 3 4 4 4 3 3 4 3 4 2 3 3 5 6 2 4

(a) set up frequency distribution tables for absolute, relative, and accumulative frequencies

(b) calculate the average number of household members, the mode and the median.

R e s u l t: $\overline{x}$ = 3.82; $\hat{x}$ = 3; $\tilde{x}$ = 4

E x e r c i s e 2.7

For a total of 200 parts processed by an automatic machine tool the differences from the required size in micrometres have been measured. The following are the resulting differences:

1.0	1.5	-2.5	0.0	-1.5	1.0	1.0	15.0	-1.0	2.0
2.0	3.0	11.0	-1.0	5.0	4.5	0.5	3.5	8.0	5.0
4.5	3.5	9.5	12.0	7.5	7.5	10.0	8.5	10.0	11.0
14.0	11.0	11.0	13.0	16.0	14.5	19.0	14.0	18.0	19.0
19.0	23.5	22.0	18.5	19.5	17.5	18.0	19.5	17.5	25.5

19.5	22.0	13.5	18.5	21.5	27.5	21.0	13.5	11.5	10.0
7.5	8.5	6.5	8.5	5.5	26.0	12.5	6.5	8.5	7.5
2.5	7.0	4.5	-1.5	4.0	5.5	1.0	4.0	6.5	5.5
4.5	5.0	7.5	5.0	5.5	6.0	6.5	-3.0	5.0	3.5
-3.0	-14.0	17.0	-9.0	-3.0	-12.0	8.5	12.0	6.0	8.5
0.0	7.0	-1.0	-3.0	0.5	0.0	2.0	-4.5	2.0	-10.0
-8.5	-3.5	-11.5	-7.5	-11.5	-6.5	2.0	-11.5	-11.0	-17.5
-15.0	-15.5	1.5	-18.0	-20.0	-15.0	-3.0	-8.0	-1.0	-6.5
-8.0	-13.5	-12.0	-17.0	-10.5	14.5	10.0	9.5	7.0	0.5
21.0	10.5	5.0	0.5	4.0	0.0	0.5	3.5	9.0	2.5
2.0	7.0	7.5	3.5	7.0	4.5	-1.0	11.0	4.0	9.0
4.5	11.5	14.0	10.0	20.0	13.0	7.0	12.0	7.5	2.0
1.0	25.0	0.5	-3.0	4.5	6.0	9.5	12.5	19.0	13.0
1.5	0.5	12.0	4.0	6.5	-9.5	-8.0	-4.5	7.5	-4.0
-9.0	-9.0	2.0	-0.5	3.5	10.5	-5.5	-6.0	-6.5	-8.0

Summarize the data and use the resulting crosstabulation to calculate the arithmetic mean, the standard deviation and the coefficient of skewness.

R e s u l t: x_{min} = -17.5 μm; x_{max} = 27.5 μm; h = 5.0 μm; m = 10; $\bar{x} \cong 4.3$ μm; $s \cong 9.7$ μm; $A \cong -0.102$

E x e r c i s e 2.8

The below frequency distribution shows how a total of 200 workers have met the norm. The numbers in the upper line are the midpoints of percentage classes:

x_j^*	85	95	105	115	125	135	145	155	165	175
F_j	4	21	65	39	24	17	12	9	7	2

Use the frequency distribution to calculate the arithmetic mean, mode, median, standard deviation, and coefficient of skewness.

R e s u l t: $\bar{x}$ = 117.9 %; $\hat{x}$ = 105 %; $\tilde{x}$ = 115 %; $s^2 \cong 380$ %^{2}; $A \cong 0.92$

E x e r c i s e 2.9

In a statistical survey made by an insurance company each person has been asked about the bonus they are paying. The following is the resulting frequency distribution. The upper line shows the bonuses in CZK:

x_j^*	390	410	430	450	470	490	510	530	550	570
F_j	7	10	14	22	25	12	3	3	2	2

Calculate the arithmetic mean, mode, median, variance, standard deviation, coefficient of skewness, and coefficient of variation.

R e s u l t: $\bar{x}$ = 457.4 CZK; $\hat{x}$ = 470 CZK; $\tilde{x}$ = 450 CZK; s^2 = 1493.24 CZK^2;
$s \cong 38.64$ CZK; $A \cong 0.52$; $v \cong 8$ %

E x e r c i s e 2.10

Calculate the measures for the following bivariate sample data:

x_i	18	19	20	21	22	22	25	26	26	26	27	28	29	30	31	33
y_i	26	23	29	27	31	25	22	32	32	33	38	29	36	37	41	42

R e s u l t: $\bar{x} \cong 25.19$; $\bar{y} \cong 31.44$; $s^2(x) \cong 18.777$; $s^2(y) \cong 35.246$;
$s(x) \cong 4.33$; $s(y) \cong 5.94$; cov $\cong 20.980$; $r \cong 0.8155$

E x e r c i s e 2.11

Calculate the measures for the following bivariate sample data:

x_i	2	4	4	5	6	8	10	10	10	10
y_i	1	2	3	4	4	4	5	5	5	6

R e s u l t: $\bar{x}$ = 6.9; $\bar{y}$ = 3.9; $s^2(x)$ = 8.49; $s^2(y)$ = 2.09;
$s(x) \cong 2.91$; $s(y) \cong 1.45$; cov = 3.89; $r \cong 0.9235$

E x e r c i s e 2.12

The following contingency table summarizes last year's (x) and this year's (y) prices of shares in thirty companies selected at random. Find the average prices of shares and the correlation coefficient.

x_j \ y_k	1001 - 2000	2001 - 3000	3001 - 4000
501 - 1000	3	6	0
1001 - 1500	5	8	2
1501 - 2000	0	1	3
2001 - 4000	0	1	1

R e s u l t: $\bar{x}$ = 1283.8 CZK; $\bar{y}$ = 2433.8 CZK; r = 0.4232

Questions

1. Describe the types of variables and show examples.
2. What is a sample, what are its properties and how it is done?
3. Define sample data and show how it is related to the parent population.
4. Describe the way univariate sample data with a quantitative variable are grouped.
5. What are the measures of central tendency for univariate sample data with a quantitative variable, what are their properties and what is their significance?
6. Show the measures of variation for univariate sample data with a quantitative variable, their properties and their significance.
7. Describe types of graphical representation of univariate sample data with a quantitative variable.
8. Show the measures for bivariate sample data with quantitative variables.
9. Describe the way bivariate sample data with quantitative variables are summarized and the types of their graphical presentation.
10. Describe the way sample data with qualitative variables are processed and graphically presented.

3. ANALYSIS OF TIME SERIES

Fundamentals

Two main categories of statistical information exist: cross sections and time series. The economists often estimate the consumption by relating the consumers' costs to the national product or analyse in detail the distribution of consumption at one particular point of time (cross sections). This approach has a broader significance for the practice but is not sufficient if we are interested the dynamic of an event and in particular changes over time.

The basic tool employed to study of the dynamic an event is an analysis of its past development, which helps us grasp the existing laws and to estimate its future development.

A time series is obtained if the data on a particular event over time are arranged in order of increasing time. A well-established time series that can be used for an analysis must meet the following requirements.

- the data must be arranged in order of increasing time,
- the data items must be comparable:
    a) the same period of time over which the data has been acquired,
    b) the same data definition (units of measurement, uniform data collection method).

Failure to comply with any of the above conditions may result in erroneous conclusions.

From the statistical point of view a time series is a sequence $(y_1,...,y_n)$ of the observed values of a statistical variable Y where the index i corresponds to the time t_i or to the i-th interval ending at t_i, $t_i < t_{i+1}$, $i = 1,...,n$. Sometimes we write y_t instead of y_i. Graphically, the time series is mostly represented by a graph in the Cartesian

system of co-ordinates with the indices i or times t_i as abscissas and the values y_i as ordinates. Fig 3.1 shows an example of a time series.

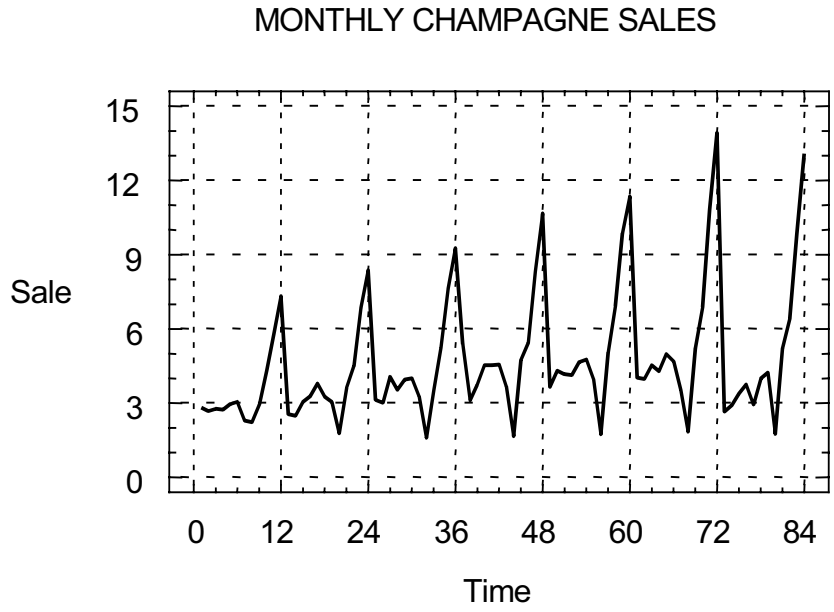


Fig. 3.1

If time series are related to periods of time, they are called interval time series, if they refer to points of time, they are called point time series.

Interval time series are composed of indexes measured for fixed time intervals such as an hour, day, month, or year, etc. They are characterised by the following features:

- the data items express quantities,
- they are dependent on the length of the time interval,
- the sum of the data is meaningful.

Point time series contain data that are related to a fixed point in time. They are characterised by the following features:

- the data express a level or a condition of the variable under investigation,
- there is no point in adding up all the data items.

To analyse a time series correctly certain differences must be taken into consideration that follow from the different character of time series data and from their significance.

The following is a classification of time series:

1) absolute quantities: a) interval
   b) point

2) derived quantities: a) sums: α) cumulative
   β) moving sums
   b) averages: α) cumulative averages
   β) moving averages
   c) quotients

Interval and point time series

Interval series

It is typical of interval time series that they are related a fixed time interval and as such are affected by the length of the interval. The following are the most common interval quantities: production volume, retail sales, revenues, wages and salaries, man-hours, number of children born within a certain period, etc.

Point time series

The quantities used for setting up a time series typically do not refer to an interval but rather to a time point. This may be the first or the last day of a period, an arbitrary but fixed day or moment. The number of inhabitants, workers, the amount of fixed assets, etc may exemplify data of this type. Such data show an instantaneous condition of the event in question. We use the following numerical characteristic called the chronological mean to aggregate the data.

Given the values of a point index (outcomes of the observed variable Y)

$$y_1, y_2,..., y_n$$

for n time points

$$t_1, t_2,..., t_n,$$

we can calculate the averages for the following points

$$t_1 \text{ and } t_2,\ t_2 \text{ and } t_3,..., t_{n-1} \text{ and } t_n\ .$$

Using these partial averages we now calculate the average for all the aggregate point values

$$\bar{y}_1 = \frac{y_1 + y_2}{2},\ \bar{y}_2 = \frac{y_2 + y_3}{2},..., \bar{y}_{n-1} = \frac{y_{n-1} + y_n}{2}.$$

The number dividing the sum of the above partial averages will be one less than n.

If the intervals between the individual values of a point time series are of equal length, the chronological mean is calculated as follows. Denoting the distances between the members of a point time series by

$$d_1, d_2,... , d_{n-1},$$

we have

$$d_1 = d_2 =...= d_{n-1},$$

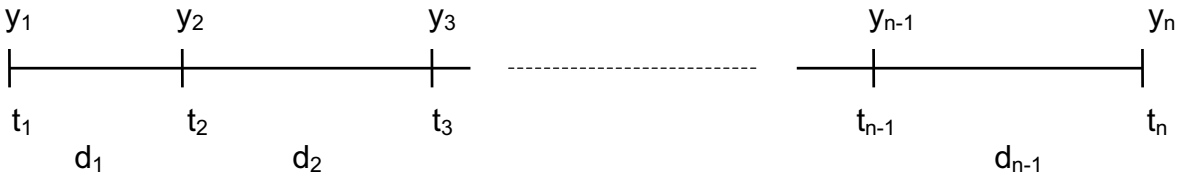
and the chronological mean is

$$\bar{y}_{chr} = \frac{\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \cdots + \frac{y_{n-1} + y_n}{2}}{n-1}\ .$$

After some manipulations we get

$$\bar{y}_{chr} = \frac{0.5y_1 + y_2 + y_3 + \cdots + y_{n-1} + 0.5y_n}{n-1}\ .$$

If the distances between the neighbouring members of a point time series are not equal, the calculation is similar. We reduce the different lengths of time $d_1, d_2,..., d_n$ to one value. We use their weighted average to do this:



The chronological mean is then

$$\bar{y}_{chr} = \frac{\frac{y_1 + y_2}{2} d_1 + \frac{y_2 + y_3}{2} d_2 + \cdots + \frac{y_{n-1} + y_n}{2} d_{n-1}}{d_1 + d_2 + \cdots + d_{n-1}}$$

which yields

$$\bar{y}_{chr} = \frac{y_1 d_1 + y_2 (d_1 + d_2) + \ldots + y_{n-1} (d_{n-2} + d_{n-1}) + y_n d_{n-1}}{2(d_1 + d_2 + \ldots + d_{n-1})} \quad .$$

E x a m p l e 3.1

The following are data on the numbers of employees of a company during the calendar year:

1st Jan	3 500 employees
1st Apr	3 425 employees
1st Jul	3 430 employees
1st Oct	3 390 employees
1st Jan	3 350 employees

Calculate the chronological mean of the time series expressing the numbers of employees.

Solution:

We use the above formulas to calculate the chronological mean, where $y_1 = 3\,500$, $y_2 = 3\,425$, $y_3 = 3\,430$, $y_4 = 3\,390$, $y_5 = 3\,350$ and $d_1 = d_2 = d_3 = d_4$. Substituting these values we get

$$\bar{y}_{chr} = \frac{(0.5)(3\,500) + 3\,425 + 3\,430 + 3390 + (0.5)(3\,350)}{4} = 3\,417.5 \ .$$

The chronological mean of the numbers of employees of the company is 3 417.5 for the given year. For practical use we can round off the value to 3 418.

Example 3.2

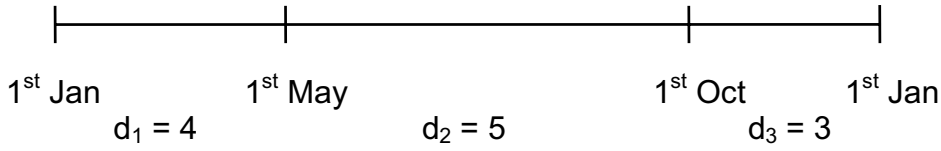
A company keeps an inventory of stock. The total figures are in Czech Korunas. The data for the following dates are available:

1st Jan	20.523 million
1st May	16.100 million
1st Oct	17.230 million
1st Jan	21.432 million

Calculate the average yearly stock in the company.

Solution:

We use the formula for the calculation of the chronological mean, where $y_1 = 20.523$, $y_2 = 16.100$, $y_3 = 17.230$, $y_4 = 21.432$ and



Substituting the data we get

$$\bar{y}_{chr} = \frac{(20.523)(4) + 16.100(4+5) + 17.230(5+3) + (21.432)(3)}{2(4+5+3)} \cong 17.880\,.$$

The average yearly stock in the company was 17.880 million CZK.

Special types of time series

Cumulative time series

Cumulative time series behave like increasing sums. A cumulative time series is formed by gradually adding up the values of a given variable starting from a fixed point. This method is employed for example when monitoring indexes for a certain period of time such as a month or a year. Cumulative values are of considerable help

in matters of strategic decision making. In the following example use of this method is demonstrated.

E x a m p l e 3.3

Using the following data on the production volume for each month of the year analyse the real production for each month and for the whole year.

Production (thousands of tons)						
Month	Monthly values			Cumulative values		
	Plan	Production	(%)	Plan	Production	(%)
January	36.2	36.5	100.8	36.2	36.5	100.8
February	36.5	35.8	98.1	72.7	72.3	99.4
March	35.1	35.8	102.0	107.8	108.1	100.3
April	34.2	35.1	102.6	142.0	143.2	100.8
May	33.0	33.2	100.6	175.0	176.4	100.8
June	33.0	32.1	97.3	208.0	208.5	100.2
July	33.0	31.2	94.5	241.0	239,7	99.5
August	32.5	31.0	95.4	273.5	270.7	99.0
September	32.7	32.3	98.8	306.2	303.0	99.0
October	34.6	33.4	96.5	340.8	336.4	98.7
November	36.8	36.6	99.5	377.6	373.0	98.8
December	38.0	38.1	100.3	415.6	411.1	98.9

From the tabulated values it can be concluded (cf. December) that, as compared with the planned value 415.6 thousand tons, the yearly total was 411.1 thousand tons, which is 98.9 % of the plan.

Time series of cumulative averages

Time series of cumulative averages are derived from interval series. These series show how the cumulative averages approach the total average over the given period of time, which is equal to the last value.

This method is used for example to record the costs in quality control. It is based on a cumulative time series where the values are divided by the number of periods over which it has been accumulated. We will use the data from Example 3.3 to demonstrate this. The quantities are shown in thousands of tons.

E x a m p l e 3.4

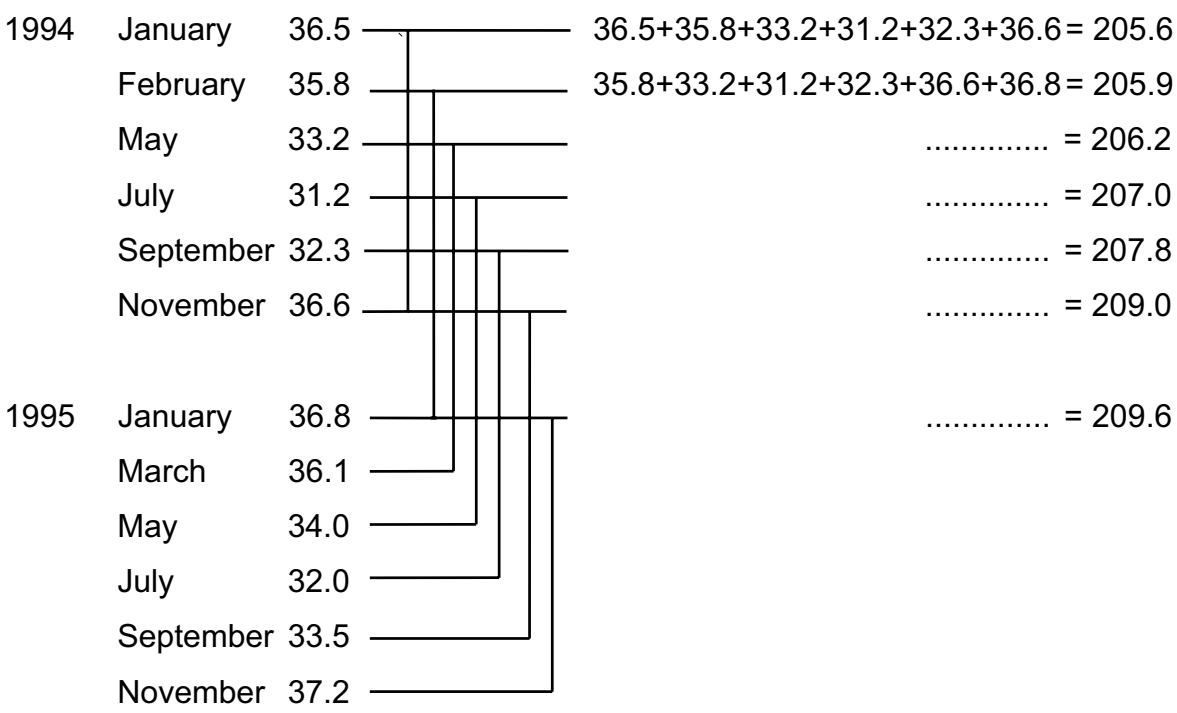
Cumulative production and cumulative averages (thousands of tons):

January	36.5	36.5 / 1 = 36.50
February	72.3	72.3 / 2 = 36.15
March	108.1	108.1 / 3 ≅ 36.03
April	143.2	143.2 / 4 = 35.80
May	176.4	176.4 / 5 = 35.28
June	208.5	208.5 / 6 = 34.75
July	239.7	239.7 / 7 ≅ 34.24
August	270.7	270.7 / 8 ≅ 33.84
September	303.0	303.0 / 9 ≅ 33.67
October	336.4	336.4 /10 = 33.64
November	373.0	373.0 /11 ≅ 33.91
December	411.1	411.1 /12 ≅ 34.26

Cumulative series of moving sums

Series of moving sums are interval-type time series. In practice we usually calculate moving yearly totals using data for individual months. As a rule, we use data collected over a minimum of two years. These series are created by gradually adding up consecutive values. A series of moving sums is suitable for comparing development trends over two different periods.

E x a m p l e 3.5



The values for 1994 and 1995 clearly show that the production trend is increasing.

Time series of moving averages

Moving averages are related to the calculation of moving series. Moving averages are calculated by dividing moving sums by the number of period over which the sum has been made. A series of moving averages smoothes possible seasonal influences of current values. More information on moving averages can be found in [1], [3], [4], and [5].

Development of time series

The simplest numerical characteristics used to analyse time series are the absolute and relative measure of growth and decline. An analysis of absolute and relative measures of growth enables decisions necessary for the selection of a function used for smoothing a time series. For the following methods, we will always assume that the neighbouring boundaries or midpoints of the time intervals are equidistant.

Absolute measures of growth provide absolute comparison between the members of a time series. The following measures are used:

absolute increment (difference)

$$\delta_i = y_i - y_{i-1}, \quad \text{for } i = 2, 3, \ldots, n,$$

mean absolute increment

$$\overline{\delta} = \frac{1}{n-1}\sum_{i=2}^{n}\delta_i = \frac{y_n - y_1}{n-1} ,$$

which is calculated as the simple arithmetic mean of all absolute increments.

If all average increments (also called first differences $\delta_i^{(1)}$) are close to a constant, the time series has a linear trend, which can be expressed in the form of a straight line. Second differences $\delta_i^{(2)}$ are obtained by subtracting two neighbouring first differences. Sometimes also their arithmetic mean $\overline{\delta}^{(2)}$ is calculated. If the second differences of a time series are all close to a constant, the time series may be represented by a parabola. Third differences are calculated as the differences of two neighbouring second differences. Further differences are established in a similar way. If the differences of order k are almost constant, the corresponding time series may be represented by a k-th order polynomial.

The rate of development (growth or decline) of the values of a time series is characterised by relative increments calculated as the ratio of the first difference of the i-th period to the current value of the (i -1)th period. The development rate may be expressed by the following characteristics:

coefficient of growth

$$k_i = \frac{y_i}{y_{i-1}} \quad \text{for} \quad i = 2, 3, \ldots, n,$$

coefficient of increment

$$\kappa_i = \frac{\delta_i}{y_{i-1}} = k_i - 1 \quad \text{for} \quad i = 2, 3, ..., n,$$

average coefficient of growth

$$\bar{k} = \sqrt[n-1]{k_2 k_3 ... k_n} = \sqrt[n-1]{\prod_{i=2}^{n} k_i} ,$$

which is calculated as the geometric mean of coefficients of growth.

The average coefficient of growth may also be calculated as the (n - 1)-th root of the quotient of the first and the last current value in the given time series

$$\bar{k} = \sqrt[n-1]{\frac{y_2}{y_1}\frac{y_3}{y_2}\frac{y_4}{y_3} \dots \frac{y_{n-1}}{y_{n-2}}\frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\frac{y_n}{y_1}} .$$

Coefficients of growth and increment as well as the corresponding average values are sometimes expressed as percentages: $k_i 100\%$, $\kappa_i 100\%$, $\bar{k} 100\%$. If the growth coefficients ki of a time series vary slightly around a constant, the trend of the series is approximately exponential.

E x a m p l e 3.6

The GDP figures (in thousands of millions of CZK) in the Czech republic between 1990 and 1996, recalculated for fixed prices, is given in the table below.

Year	1990	1991	1992	1993	1994	1995	1996
GDP	564	721	859	1015	1192	1338	1579

Determine the average GDP, absolute yearly increments, the average yearly increment, second differences, the average second difference, the coefficient of growth, and the average coefficient of GDP growth.

Solution:

Part of the results can be found in the table where t as the time variable is used instead of i:

t	y_t	δ_t	$\delta_t^{(2)}$	k_t	$k_t 100\%$	κ_t	$\kappa_t 100\%$
1990	564	---	---	---	---	---	---
1991	721	157	---	1.2784	127.84	0.2784	27.84
1992	859	138	-19	1.1914	119.14	0.1914	19.14
1993	1015	156	18	1.1816	118.16	0.1816	18.16
1994	1192	177	21	1.1744	117.44	0.1744	17.44
1995	1338	146	-31	1.1225	112.25	0.1225	12.25
1996	1579	241	95	1.1801	118.01	0.1801	18.01
Σ	7268	1015	84	---	---	---	---

The average yearly GDP is

$$\bar{y} = \frac{7268}{7} \cong 1038.2857 \cong 1038.3 \text{ thousands of million CZK.}$$

Absolute increments δ_t, second differences $\delta_t^{(2)}$, coefficients of growth k_t and coefficients of increment κ_t are shown in the table. Hence we see that 241 thousands of million CZK was the greatest value of the GDP reached in 1996, and in 1992 it dropped to a minimum of 138 thousands of million CZK. However the largest relative growth of the GDP was attained in 1991 (the coefficient of growth being 1.2784 or 127.84 %) and the least relative growth of the GDP was recorded in 1995 (the coefficient of growth being 1.1225 or 112.25 %). Further the table shows that the greatest absolute acceleration of the GDP (the greatest positive second difference) was reached in 1996 and the greatest slow-down of the GDP (the least negative value of the second difference) rate was recorded in 1995. The average yearly absolute increment of the GDP is

$$\bar{\delta} = \frac{1075}{7-1} = 169.1666\ldots \cong 169 \text{ thousands of million CZK.}$$

The average yearly growth coefficient for the GDP is

$$\bar{k} = \sqrt[7-1]{\frac{1579}{564}} \cong \sqrt[6]{2.7996454} \cong 1.1872 \text{ or } 118.72\ \%.$$

Hence the average yearly growth coefficient for the GDP is 0.1872 or 18.72 %. The calculation of the average coefficient of growth that uses the arithmetic mean may be misleading but unfortunately this often occurs in economic applications. The average yearly second difference of the GDP (in thousands of millions of CZK) is

$$\bar{\delta}^{(2)} = \frac{84}{7-2} = 16.8 > 0,$$

which means that the overall growth of the GDP is increasing.

Time series analysis

Time series is the measurement of a variable over time. The following major components, or movements, of time series may be identified:

- trend (long-term influence),
- periodic influences (recurring regularly) affecting the values of a time series,
- irregular influences (occurring at random, forecasting is difficult).

The trend of a time series

The trend is an important component of time series. It may be linear or any other non-linear function of time. The trend is the general movement over a long period of time. When using time series analysis in economy, we are interested in the trend both in terms of the present situation and a forecast of its future development. A number of methods have been devised and computer programs have been written to express the trend of time series [1], [3], and [4], [5] a [6]. The basic methodology is described below in the paragraph on time series smoothing.

Periodic influences

Periodic influences account for periodic variations of time series over time. The length of the periods varies as well and can be used for further subdivision of periodic

influences as follows:

- cycles (wavelike repetitive movements fluctuating about the trend of the series),

- seasonal variation (repetitive fluctuating movements occurring within a time period of one year or less),

- short-term influences (fluctuations with regular periods such as a day in the week, a week in the month and the like).

A number of methods have been devised and computer programs have been written to express the periodic changes of time series [1], [3], [4], [5] and [6].

Irregular influences

The irregular component of a time series refers to random movements around its trend or seasonal component. We consider these movements as interfering components. They can be represented by random variables and may be described by statistical methods [1] and [2]. Some of the methods are treated in Chapter 5.

Time series decomposition

A time series may be thought of as the result of its trend component T_t, periodic component P_t and random component E_t. The periodic component may further be decomposed into a cyclic component C_t and a seasonal component S_t. The decomposition of a time series is mostly based on an additive model

$$y_t = T_t + P_t + E_t \,, \quad \text{or} \quad y_t = T_t + C_t + S_t + E_t$$

or a multiplicative model

$$y_t = T_t \, P_t \, E_t \,, \quad \text{or} \quad y_t = T_t \, C_t \, S_t \, E_t \,.$$

Time series smoothing

When analysing the trend component of a time series, we try to identify the influence of those factors that are stable and determine the trend. Graphically, this corresponds to drawing a curve that best fits the time series trend when plotted. Such

a curve may be obtained by graphically, mechanically, or analytically smoothing the time series.

For graphical smoothing the time series is plotted in a graph (Fig. 3.1) and its trend is estimated (smoothed) graphically. This method can only serve as a guideline and sometimes it may be misleading.

Mechanical smoothing of a time series is based on moving sums. When moving sums are divided by the number of periods, moving averages are obtained with values mostly close to the original values but devoid of their seasonal variations. Thus the moving average curve tends to be smoother than the original one. Moreover, it becomes more monotonous (and shorter) as more periods are taken as the basis to calculate the moving sums. A major advantage of this method is its simplicity and the fact that it informs us well on the development trend of the time series without interfering seasonal and cyclic fluctuations.

Analytical smoothing of a time series is based on the assumption that a function of time exists that approximates the values of the time series. We try to select a function that best suits its development. The selection can be guided by a graphical representation of the time series empirical values or by calculating its first (second or other) differences. The following is the general form of a smoothing function:

$$y'_t = f(t) + e_t ,$$

where: y'_t - denotes the smoothed values of the dependent variable,

t - denotes the independent time variable,

e_t - denotes the so-called residual component.

The function f(t) ought to render the trend of the time series correctly, that is, to smooth the time series as well as possible. Linear, parabolic, and exponential functions are the ones most frequently used. Generally, any function may be used - more details can be found in Chapter 5.

The most frequently used method is linear smoothing if the trend of the time

series appears to be linear. The function f(t) has the form

$$y'_t = b_0 + b_1 t \ ,$$

and the parameters b_0 and b_1 are determined from the so-called system of normal equations

$$b_0 T + b_1 \sum_{t=1}^{T} t = \sum_{t=1}^{T} y_t \ ,$$

$$b_0 \sum_{t=1}^{T} t + b_1 \sum_{t=1}^{T} t^2 = \sum_{t=1}^{T} y_t t \ .$$

The first coefficient b_0 determines the point at which the smoothing straight line intersects the y-axis. It is interpreted as the smoothed value of the time series in period zero. The second coefficient b_1 is the slope of the straight line and expresses the actual trend. It determines the change of the smoothed values y'_t for a unitary change of t or the average change of the original values y_t when t is increase by one. We can test the suitability of the smoothing function by looking at the plotted diagram or by calculating the correlation coefficient of the pairs (t, y_t) or the sum of the squared differences $\sum(y_t - y'_t)^2$.

This method may be simplified if we shift the time variable for the sum of its shifted values to be equal to zero. This can be achieved by shifting the origin (0) to the central period, that is, by decreasing t by its mean value

$$\bar{t} = \frac{T+1}{2} \ .$$

Thus, instead of t, we consider the variable $t' = t - \bar{t}$. By this transformation the terms in the system of normal equations with Σt turn to zero, which yields the following explicit formulas

$$b'_0 = \frac{\sum_{t=1}^{T} y_t}{T} \ , \quad b'_1 = \frac{\sum_{t=1}^{T} y_t t'}{\sum_{t=1}^{T} t'^2} \ .$$

For the original coefficients we have

$$b_0 = b'_0 - b'_1 \bar{t} \ , \quad b_1 = b'_1 \ .$$

Example 3.7

Determine the trend component of the time series representing the development of the gross domestic product in the Czech Republic from 1990 to 1996 as shown in Example 3.6.

Solution:

As the original time variable t takes on the values 1990, 1991,…, 1996, we will use the transformation $t' = t - 1993$ since then we have

$$\bar{t} = (1990 + 1991 + \ldots + 1996)/7.$$

The number of periods T = 7 so that we get

$$b_0' = \frac{564 + 721 + \cdots + 1579}{7} = \frac{7268}{7} = 1038.2857 \cong 1038.3 \ ,$$

$$b_1' = \frac{(-3)(564) + (-2)(721) + \cdots + (3)(1579)}{(-3)^2 + (-2)^2 + \cdots + 3^2} = \frac{4612}{28} = 164.71428 \cong 164.7 \ .$$

The original coefficients being

$$b_0 = 1038.2857 - (164.71428)(1993) = -327\ 237.28 \cong -327\ 237.3 \ ,$$

$$b_1 = 164.71428 \cong 164.7 \ .$$

Hence we get the following straight line, which smoothes the time series

$$y_t' = -327\ 237.3 + 164.7\ t \ .$$

For t = 1993, say, we obtain $y_{1993}' = -327\ 237.3 + (164.7)(1993) = 1009.8$, which is in good correspondence with the real GDP y_{1993} = 1015. Also the value b_1 = = 164.7 is close to the average increment $\bar{\delta}$ = 169 from Example 3.6.

Exercises

Exercise 3.8

The below data describe the amount of fixed assets in a company over a calendar year (the accounting value):

1st Jan	101.230 million CZK	1st Aug	100.250 million CZK
1st Mar	105.100 million CZK	1st Dec	99.800 million CZK
1st Apr	105.500 million CZK	1st Jan	103.150 million CZK

Calculate the average fixed assets for the calendar year.

R e s u l t: $\bar{y}_{chr}$ = 102.05875 million CZK

E x e r c i s e 3.9

The figures displayed in the table show electricity consumption in Czechoslovak industry between 1967 and 1972 in thousands of millions of kilowatt-hours.

Year	1967	1968	1969	1970	1971	1972
Consumption	27.6	29.1	30.3	31.8	33.6	35.4

Determine the average yearly electricity consumption, absolute yearly increments, the average yearly increment, second differences, the average second difference, the coefficient of growth, the coefficient of increment, and the average coefficient of growth of electricity consumption.

R e s u l t: $\bar{y} = 31.3$; $\bar{\delta} = 1.56$; $\bar{\delta}^{(2)} = 0.075$; $\bar{k} \cong 1.051$

E x e r c i s e 3.10

Smooth the average fixed assets figures (in millions of CZK) shown in the table below for the years 1978 to 1985 using the linear smoothing method and calculate a forecast of the average fixed assets for 1987 assuming that the trend of the time series remains does not change.

Year	1978	1979	1980	1981	1982	1983	1984	1985
Amount	675.0	681.4	684.0	689.6	690.8	698.2	706.0	712.0

R e s u l t: $y'_t = -9352.193 + 5.069t$, $y'_{1987} = 719.91$ mill. CZK

E x e r c i s e 3.11

A transport company recorded the following average numbers vehicles in its fleet:

1993…54 , 1994…63 , 1995…69 , 1996… 72.

The company's real figures as of particular dates of 1997 are shown in the below table:

Date	1st Jan	20th Mar	15th Apr	30th Jul	25th Sep	31st Dec
Number	70	66	71	80	82	90

Determine the average number of the company's vehicles for 1997, the average yearly coefficient of growth and characterise the trend of the development from 1993 to 1997 using a linear function.

R e s u l t: $\bar{y}_{chr} \cong 77$ (weighted yearly average); $\bar{k} \cong 1.093$ (average yearly increase by 9.3 %); $y'_t = -10\,905.5 + 5.5t$

Questions

1. Define a time series and show examples.
2. What is the classification of time series. Show examples of each type.
3. How is the average of a time series calculated?
4. What characteristics describe movements of a time series?
5. Describe the components of a time series and its decomposition.
6. What methods are used to smooth a time series?

4. INDEX NUMBERS

Basic notions

Index numbers are relative statistical measures that express change in magnitude of a quantitative variable or a group of variables describing the behaviour of one or several items over a period of time or as influenced by a factor. They usually take the form of a fraction where the numerator is the value of the variable for the current period and the denominator is its value for the base period. Depending on the character, the construction complexity, and the influence of the quantities observed, indexes are either simple or composite.

Simple indexes measure the relative change from the base period for a single item or for a group of homogeneous items. In the former case they are called single indexes (describing such quantities as the price or the amount of a single product) and in the latter case they are called group indexes (when used to measure the change in one variable (such as the price or quantity) for a group of homogeneous items. As opposed to that, composite indexes measure the relative change from the base period in a group of inhomogeneous items or an aggregate (such as a bundle of commodities of different types).

Indexes that measure the number of items, the production volume and the like are called quantity indexes and are denoted by q while those that measure such quantities as price or intensity are called value indexes and are denoted by p. Among the indexes of the first group the most frequently used are volume indexes while those in the second group are usually price indexes.

Simple index numbers

Single indexes

The single index i_q for a quantity or i_p for a value is given by

$$i_q = \frac{q_1}{q_0} \quad \text{or} \quad i_p = \frac{p_1}{p_0} \ ,$$

respectively where the numerator corresponds to the current period and the denominator corresponds to the base period. The correct selection of the base period is important. Those values that best represent the outcomes of the variable should be chosen as the base. Sometimes it is best to use the average of several values. When calculating and interpreting index numbers, we must ensure comparability of the periods and a factual agreement of the aspects concerned otherwise the expressive power of the index could be significantly impaired.

Example 4.1

The production volume of a steel works reached 2780 tons in 1994 when the price of steel was 8750 CZK and in 1995 the production volume rose to 2950 tons with a price of 9690 CZK. Calculate indexes for the production and price of steel.

Solution: $i_q = \frac{2950}{2780} \cong 1.061 = 106.1\%$, $i_p = \frac{9690}{8750} \cong 1.107 = 110.7\%$

Group indexes

While single indexes are used to analyse single items such as the quantity of cement produced by one plant, group indexes are related to a group of similar items such as the cement production volumes for a group of plants. The distinction between single and group indexes is of great importance. For group indexes, the comparability of periods, facts and composition plays an important role. For a quantity variable, the group index is given by

$$i_q = \frac{\sum_i q_1^{(i)}}{\sum_i q_0^{(i)}} \ .$$

Example 4.2

The production figures for four cement production plants are given in the following table:

Plant	January (0)	February (1)	March (2)
A	2300 t	2450 t	2390 t
B	5210 t	4800 t	5100 t
C	8100 t	8600 t	9000 t
D	6250 t	6250 t	5900 t
Total production	21860 t	22400 t	22090 t

If we consider all the plants as one unit, the group indexes for a quantity variable are as follows (in fact they are what will be later called chain production indexes):

$$i_{q1} = \frac{22\,400}{21860} \cong 1.025 \quad , \qquad i_{q2} = \frac{22\,090}{22\,400} \cong 0.986 \ .$$

Group indexes may be further specified as variable composition indexes, fixed composition indexes, and structure indexes. A group index describing changes in values (such as average prices) is called a variable composition index and is given by

$$i_{var.comp.} = \frac{\overline{p}_1}{\overline{p}_0} = \frac{\dfrac{\sum_i p_1^{(i)} q_1^{(i)}}{\sum_i q_1^{(i)}}}{\dfrac{\sum_i p_0^{(i)} q_0^{(i)}}{\sum_i q_0^{(i)}}} \ .$$

Example 4.3

The following table gives a summary of prices in a company:

Supply	No. of items in period		Price per item in period		Supply value in period (in thousands of CZK)			
	Base	Curr.	Base	Curr.				
	q_0	q_1	p_0	p_1	p_0q_0	p_1q_1	p_1q_0	p_0q_1
Contractual	8000	8000	12.-	13.2	96	105.6	105.6	96
Surplus	500	4600	30.-	30.-	15	138	15	138
Total	8500	12600	-----	-----	111	243.6	120.6	234

The value under investigation is the price. Price changes in individual types of supply may be measured by single indexes for contractual (1) and surplus (2) supplies:

$$\frac{p_1^{(1)}}{p_0^{(1)}} = \frac{13.2}{12} = 1.100 \ , \qquad \frac{p_1^{(2)}}{p_0^{(2)}} = \frac{30}{30} = 1.000 \ .$$

For contractual supplies the price has risen by 10% while for surplus supplies it has remained the same. The average price for the base period is

$$\overline{p}_0 = \frac{\sum p_0^{(i)} q_0^{(i)}}{\sum q_0^{(i)}} = \frac{111000}{8\,500} \cong 13.06 \ \text{CZK}$$

and the average price for the current period is

$$\overline{p}_1 = \frac{\sum p_1^{(i)} q_1^{(i)}}{\sum q_1^{(i)}} = \frac{243\,600}{12\,600} \cong 19.33 \ \text{CZK},$$

which yields a variable composition index of

$$i_{\text{var.comp.}} = \frac{19.33}{13.06} \cong 1.480 = 148\% \, .$$

Composite indexes

The basic property of composite (aggregate) indexes is that they can be used to measure changes in quantities and values of inhomogeneous variables. If, for example, the prices of consumer goods have changed, we would like to know the percentage of the drop in prices of products as a whole. Since we have to deal with a range of different types of goods, the average price of a unified item cannot be used.

The change may be determined by considering it to be the sum of values multiplied by the corresponding quantities. Thus in the above example the total value will be the result of two variables:

a) value variable (such as price, total costs, and labour intensity),

b) quantity variable (a carrier of the above values).

Then the so-called composite value index (retail trade turnover) is given by

$$I_h = \frac{\sum_i q_1^{(i)} p_1^{(i)}}{\sum_i q_0^{(i)} p_0^{(i)}} \quad .$$

To determine the influence of only one of the variables, the influence of the other must be eliminated, which is done by fixing it at a given constant level in each of the aggregates that are being compared. A constant level for a value variable (p) or a quantity variable (q) to calculate the index may be achieved in two ways: by fixing it at the level of either the base period or the current period. Accordingly, we get the following indexes:

the Laspeyres composite index for quantity $$I_q^L = \frac{\sum_i q_1^{(i)} p_0^{(i)}}{\sum_i q_0^{(i)} p_0^{(i)}} \quad ,$$

the Laspeyres composite index for value $$I_p^L = \frac{\sum_i q_0^{(i)} p_1^{(i)}}{\sum_i q_0^{(i)} p_0^{(i)}} \, ,$$

the Paasche composite index for quantity $$I_q^P = \frac{\sum_i q_1^{(i)} p_1^{(i)}}{\sum_i q_0^{(i)} p_1^{(i)}} \, ,$$

and the Paasche composite index for value $$I_p^P = \frac{\sum_i q_1^{(i)} p_1^{(i)}}{\sum_i q_1^{(i)} p_0^{(i)}} \, .$$

The following relationships are easily established between the composite indexes for quantity and value and the composite value index

$$I_h = I_p^L I_q^P = I_q^L I_p^P \, ,$$

since we have

$$\frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum q_0 p_1}{\sum q_0 p_0} \frac{\sum q_1 p_1}{\sum q_0 p_1} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \frac{\sum q_1 p_1}{\sum q_1 p_0} \, .$$

Neither the Laspeyres nor the Paasche index expresses a change in a satisfactory manner since a change in p or q between the base period and the current one may cause a change in q or p respectively. For example a change in prices may influence the consumption and vice versa. To eliminate this drawback, we use sometimes the Fisher ideal index

$$I_q^F = \sqrt{I_q^L \, I_q^P} \quad \text{or} \quad I_p^F = \sqrt{I_p^L \, I_p^P} \ .$$

However, not even the Fisher index does reflect the changes sufficiently (it is only a compromise between the Laspeyres and the Paasche index) and therefore further indexes are used as well [1].

Example 4.4

The following table contains sales figures for products A, B, C in a trading company:

Product	Items sold		Item retail price	
	Base period	Current period	Base period	Current period
	q_0	q_1	p_0	p_1
A	1 000	1 200	60	69
B	6 000	4 500	10	11
C	8 000	9 000	8	7

Determine: a) the growth of the total retail trade turnover in the company for the given period,
b) the growth of the sales volume,
c) the change in the price of the commodities sold.

Solution:

The following table contains auxiliary calculations:

Product	i_q	i_p	q_0p_0	q_1p_1	q_0p_1	q_1p_0
A	1.200	1.150	60 000	82 800	69 000	72 000
B	0.750	1.100	60 000	49 500	66 000	45 000
C	1.125	0.875	64 000	63 000	56 000	72 000
Total	---	---	184 000	195 300	191 000	189 000

Hence the retail trade turnover index

$$I_h = \frac{195\,300}{184\,000} \cong 1.0614 = 106.14\%.$$

The above index tells us that the retail trade turnover in the given company rose by 6.14 %. The Laspeyres index for the sales volume is

$$I_q^L = \frac{189\,000}{184\,000} \cong 1.0272 = 102.72\%$$

and the Paasche index for the prices is

$$I_p^P = \frac{195\,300}{189\,000} \cong 1.0333 = 103.33\%.$$

The sales volume rose by 2.72% and the price of the commodities sold rose by 3.33%. The following relationship may be established for the two indexes

$$I_h = I_q^L \, I_p^P \cong (1.0272)(1.0333) \cong 1.0614\,.$$

Further indexes are calculated in a similar way:

$$I_q^P = \frac{195\,300}{191000} \cong 1.0225 = 102.25\%$$

and

$$I_p^L = \frac{191000}{184\,000} \cong 1.0380 = 103.80\%\,.$$

Index numbers and absolute quantities

Indexes express only relative changes in observed variables. They are not sufficient to analyse the development to the full extent. For this reason we must know

also the absolute value of the change expressed by an index. To do this we proceed in different ways for quantity and value variables. For quantities, apart from the index

$$i_q = \frac{q_1}{q_0},$$

we are also interested in the absolute value of the quantity under investigation, which is given by

$$q_1 - q_0 .$$

In value variables, the difference between the numerator and denominator of the index fraction only indicates the increase (decrease) of the level.

Basic and chain indexes

When analysing real events we must sometimes set up a whole series (sequence) of indexes for several subsequent periods. The following two indexes are mainly used.

a) One period is taken for a base with the value y_0 of the variable to be analysed and the ratios of other outcomes y_n in current (other) periods to this base period are calculated. In this way we obtain what is called basic indexes or constant base indexes,

$$i_{n/0} = \frac{y_n}{y_0}, \quad n = 0, 1, 2, \ldots$$

b) A given period is always compared with the previous one, y_n being divided by y_{n-1}. In this way we obtain chain indexes or moving base indexes,

$$i_{n/n-1} = \frac{y_n}{y_{n-1}}, \quad n = 1, 2, \ldots$$

Chain indexes must pass the so-called intercalation test

$$i_{m/k}\, i_{n/m} = i_{n/k} .$$

We use both basic and chain indexes to set up time series to be processed by methods described in Chapter 3. For example, the geometric mean of chain indexes is the average index that expresses the same relative change in the given variable

between individual periods of equal length. Note that, in this case, we cannot use the arithmetic mean.

E x a m p l e 4.5

Calculate the basic and chain indexes for the data in the following table.

Period	Production	Denotation
January	75 000 t	q_0
February	75 250 t	q_1
March	81 000 t	q_2
April	82 100 t	q_3

Solution:

Basic indexes: $i_{0/0} = 1.0000$; $i_{1/0} \cong 1.0033$; $i_{2/0} = 1.0800$; $i_{3/0} \cong 1.0947$

Chain indexes: $i_{1/0} \cong 1.0033$; $i_{2/1} \cong 1.0764$; $i_{3/2} \cong 1.0136$

Exercises

E x e r c i s e 4.6

The beer production of a brewery for individual types of beer with the average yearly prices in 1996 and 1997 is given in the table where q is the quantity (hl) and p is the price (CZK/hl):

Type / Period	10^0		11^0		12^0	
	q	p	q	p	q	p
1996	8350	1000.-	2560	1200.-	3870	1450.-
1997	9460	1100.-	3440	1350.-	2800	1600.-

Since the products are homogeneous, calculate both single and group indexes

related to the base year 1996 and interpret them in terms of absolute changes in price, quantity, and value of the beer production.

Exercise 4.7

Using single indexes, composite indexes (of the Laspeyres, Paasche, and Fisher type) for both value and quantity, and the cost index, calculate changes for a "small" market basket of a typical four-member family. The average retail prices p (CZK/kg or CZK/l) and the quantities of purchased food q (kg or l) are shown in the following table. Interpret the resulting indexes in terms of absolute changes in the price of the basket.

Food / Period	meat		bread		pastry		soft drinks	
	q	p	q	p	q	p	q	p
May 94	12	90.-	20	11.-	8	21.-	30	5.-
May 95	11	110.-	18	12.-	10	26.-	30	7.-

Exercise 4.8

The table contains the figures of monthly loans (in millions of CZK) given by a bank in 1997. Calculate the basic and chain indexes of the amounts loaned. For the base period take (a) January, (b) July. What was the average index of monthly loans?

I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
53.2	56.2	49.5	48.0	47.6	52.8	54.5	42.9	49.2	56.0	57.1	55.0

Exercise 4.9

Use the basic index (with 1978 as the base period) and chain index to analyse the data from Exercise 3.10 on the average fixed assets (in millions of CZK) for the years 1978 to 1985. On the assumption that the trend of the time series does not change

use the average chain index to make a forecast of the average fixed assets in 1987.

Questions

1. Define an index and describe its types.
2. How are single indexes determined and what are their properties?
3. How are group indexes determined and what are their properties?
4. How are composite indexes determined and what are their properties?
5. Define basic and chain indexes and show their use.
6. What are the drawbacks of the Laspeyres and Paasche index? Exemplify by a particular market basket.

5. MATHEMATICAL STATISTICS

Random sample and its characteristics

Using the methods of mathematical statistics, we can describe quantities of random character from observed values. Most frequently, we try to establish the properties of the probability distribution of a random variable - estimating parameters or quantitative characteristics, testing the hypotheses that they have certain properties, analysing the relationships between them etc.

If we conduct n experiments whose results are the values of a random variable X with a distribution function $F(x,\vartheta)$ where ϑ is the parameter (or a vector parameter or its function) of a given probability distribution, we, in fact, observe a random vector $\mathbf{X} = (X_1, ..., X_n)$ whose components are independent random variables X_i which all have same distribution function equal to that of X. This random vector is called a random sample or a simple random sample from random variable X. Its size is n and its joint distribution function is

$$F(\mathbf{x}, \vartheta) = F(x_1, \vartheta) \dots F(x_n, \vartheta) = \prod_{i=1}^{n} F(x_i, \vartheta).$$

We define a random sample from a random vector in a similar way.

The numbers $x_1, ..., x_a$ where x_i is an observed value of X_i, i = 1,..., n are said to be sample data of size n. Sample data processing is described in Chapter 2.

A function of a random selection $T(X_1, ..., X_n)$ is called a sample characteristic or a statistic. The value $t = T(x_1, ..., x_n)$ it takes on at sample data $x_1, ..., x_n$ is said to be an empirical characteristic or an observed value of statistic T. The following statistics are frequently used:

a) sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

b) sample variance $S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$,

c) sample standard deviation $S = \sqrt{S^2}$,

d) sample correlation coefficient $R = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{S(X)S(Y)}$.

for a random sample from random vector (X, Y) where S(X) and S(Y) are the sample standard deviations of X and Y.

We have:

a) $E(\overline{X}) = E(X), \quad D(\overline{X}) = \dfrac{D(X)}{n}, \quad E(S^2) = \dfrac{n-1}{n}D(X)$.

b) If X has a normal distribution $N(\mu,\sigma^2)$, then $\dfrac{\overline{X}-\mu}{S}\sqrt{n-1}$ has the so-called t-distribution S(n - 1) or Student's distribution and $\dfrac{nS^2}{\sigma^2}$ has the so-called chi-square distribution $\chi^2(n-1)$.

You can find more information on statistics and their distributions in [4].

Parameter estimation

Point and interval estimations

We usually do not know the real value of a parameter ϑ of the probability distribution of an observed random variable X (or a random vector) and we try to estimate it using sample data. These parameters often include number characteristics of a random variable such as its expected value, variance, etc. We can either make a point estimate or an interval estimate.

An estimator T of a parameter ϑ is a statistic $T(X_1, ..., X_n)$ that assumes values close to the parameter ϑ whatever its value is. An estimator T is unbiased if its expected value $E(T) = \vartheta$. If the variance of such an estimator is the least possible of all unbiased estimators, we call T the minimum variance unbiased estimator. An estimator T is called consistent, if $\lim_{n\to\infty} P(|T - \vartheta| < \varepsilon) = 1$ for an arbitrary real $\varepsilon > 0$.

Further types of estimators (such as maximum likelihood estimators) are described in [1] and [4]. It holds:

a) $\overline{X}$ is an unbiased consistent estimator of the mean value E(X).

b) $\frac{n}{n-1}S^2$ is an unbiased consistent estimator of the variance D(X).

c) If X is normally distributed, the estimators under a) and b) are also the minimum variance estimators .

A point estimate of a parameter ϑ is the value $t = T(x_1,..,x_n)$ an estimator T assumes for sample data $(x_1,...x_n)$. Point estimates of the basic number characteristics are calculated as follows:

$$E(X) = \bar{x}, \quad D(X) = \frac{n}{n-1}s^2, \quad \sigma(X) = \sqrt{\frac{n}{n-1}}s, \quad \rho(X,Y) = r \ .$$

An interval estimator for a parameter ϑ with confidence 1 - α is a pair of statistics $(T_1;T_2)$ such that

$$P(T_1 \le \vartheta \le T_2) = 1-\alpha$$

for any value of ϑ. A 1 - α confidence interval or a 1 - α interval estimate for a parameter ϑ is an interval $\langle t_1; t_2 \rangle$ where t_1, t_2 are the values the statistics T_1, T_2 assume for a sample data.

The confidence 1 - α is mostly selected to be close to one, usually 0.95 or 0.99. Sometimes the confidence is shown as a percentage. Confidence intervals are either one-sided or both-sided the former being limited on one side and the latter on both sides. A confidence interval usually becomes shorter as n grows and longer as the confidence increases. Given a confidence interval's length the size of sample data needed can be established [1], [2] and [3].

Estimations of normal distribution parameters

We assume that the observed random variable X, or the random vector (X,Y) is normally distributed with parameters μ, σ^2, or ρ respectively. In the sequel we will

focus on both-sided interval estimations. One-sided estimations and interval estimations for distributions other than normal can be found in [1].

Point estimators are calculated as follows

$$\mu = \bar{x}, \quad \sigma^2 = \frac{n}{n-1} s^2, \quad \sigma = \sqrt{\frac{n}{n-1}}\, s \ , \quad \rho = r \ .$$

A (1- α) confidence interval for the mean value μ where the variance σ^2 is unknown is given by the formula

$$\left\langle \bar{x} - \frac{s}{\sqrt{n-1}}\, t_{1-\alpha/2} \ , \ \bar{x} + \frac{s}{\sqrt{n-1}}\, t_{1-\alpha/2} \right\rangle \ ,$$

where $t_{1-\alpha/2}$ is the $\left(1 - \alpha/2\right)$- quantile of a t-distribution with n - 1 degrees of freedom shown in Table T2.

A (1-α) confidence interval for the variance σ^2 is given by

$$\left\langle \frac{n s^2}{\chi^2_{1-\alpha/2}} \ , \ \frac{n\, s^2}{\chi^2_{\alpha/2}} \right\rangle \ ,$$

where χ^2_P is the P-quantile of a chi-square distribution with n - 1 degrees of freedom listed in Table T3. By taking the square root of this confidence interval we get an confidence interval for the standard deviation σ.

E x a m p l e 5.1

By measuring the lengths of 10 rollers sample data have been obtained with sample characteristics $\bar{x}$ = 5.37 mm, s^2 = 0.0019 mm^2 and s = 0.044 mm (see Example 2.1). Calculate the minimum variance unbiased point estimates for the mean value, variance and standard deviation. Assuming that the observed length is normally distributed, calculate 0.95 confidence intervals for these characteristics.

S o l u t i o n:

The point estimates are calculated as follows:

mean value μ = 5.37 mm ,

variance $\sigma^2 = \frac{10}{9} 0.0019 = 0.00211 \text{ mm}^2$,

standard deviation $\sigma = \sqrt{0.00211} \cong 0.046$ mm.

A 0.95 confidence interval for the mean value μ is calculated for $t_{0.975} = 2.262$ for 9 degrees of freedom using Table T2,

$$\mu \in \langle 5.37 - \frac{\sqrt{0.0019}}{\sqrt{10-1}} 2.262;\ 5.37 + \frac{\sqrt{0.0019}}{\sqrt{10-1}} 2.262 \rangle \cong \langle 5.337;\ 5.403 \rangle \text{ mm.}$$

A 0.95 confidence interval for variance σ^2 is calculated for $\chi^2_{0,025} = 2.70$ and $\chi^2_{0.975} =$ $= 19.0$ for 9 degrees of freedom using table T3,

$$\sigma^2 \in \langle \frac{10(0.0019)}{19.0};\ \frac{10(0{,}0019)}{2.70} \rangle \cong \langle 0.00100;\ 0.00704 \rangle \text{ mm}^2,$$

which means that a 0.95 confidence interval for standard deviation σ is

$$\sigma \in \langle \sqrt{0.00100};\ \sqrt{0.00704} \rangle \cong \langle 0.0316;\ 0.0839 \rangle \text{ mm.}$$

A (1-α) confidence interval for the correlation coefficient ρ for $n \geq 10$ is given by

$$\langle \text{tgh } z_1, \text{tgh } z_2 \rangle,$$

where

$$z_1 = w - \frac{u_{1-\alpha/2}}{\sqrt{n-3}}, \quad z_2 = w + \frac{u_{1-\alpha/2}}{\sqrt{n-3}}, \quad w = \frac{1}{2}\left(\ln \frac{1+r}{1-r} + \frac{r}{n-1} \right), \quad \text{tgh} z = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{e^{2z} - 1}{e^{2z} + 1}$$

and $u_{1-\alpha/2}$ is the $\left(1 - \alpha/2\right)$- quantile of the standard normal distribution tabulated in Table T1. For 1 - α = 0.95 and 1 - α = 0.99, we have $u_{0.975} = 1.960$ and $u_{0.995} = 2.576$ respectively.

Example 5.2

A survey launched to examine the costs and price of a product manufactured by 10 different companies two-dimensional provided sample data to calculate a correlation

coefficient of r = 0.82482 (see Example 2.3). Calculate the minimum variance unbiased point estimator and find a 0.99 confidence interval for the correlation coefficient ρ of the parent population.

S o l u t i o n:

The point estimate calculated for the correlation coefficient is $\rho = 0.82482$.

Substituting it, we get

$$w = \frac{1}{2}\left(\ln\frac{1+0.82482}{1-0.82482} + \frac{0.82482}{10-1}\right) \cong 1.21753.$$

In Table T1 we can find $u_{0.995} = 2.576$ so that

$$z_1 = 1.21753 - \frac{2.576}{\sqrt{10-3}} \cong 0.24397,\ z_2 = 1.21753 + \frac{2.576}{\sqrt{10-3}} \cong 2.19110$$

and a 0.99 confidence interval for the correlation coefficient ρ is shown below

$$\rho \in \langle \text{tgh}0.24397;\ \text{tgh}2.19110\rangle \cong \langle 0.239242;\ 0.975313\rangle .$$

Estimations of parameter of a binomial distribution

Let a random variable X have an alternative distribution with the parameter equal to p, that is, a binomial distribution Bi(1; p). When estimating p, we actually estimate the ratio p of the parent population with a desired property where X_i assumes values $x_i = 1$ or 0 depending on the i-th element chosen at random having or having not the property. Let x be the number of elements that have the property in a random sample of size n. Thus we have

$$x = \sum_{i=1}^{n} x_i .$$

The minimum variance unbiased point estimator is

$$p = \frac{x}{n} .$$

The interval estimator for n > 30 is

$$\left\langle \frac{x}{n} - u_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}; \frac{x}{n} + u_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} \right\rangle$$

where $u_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the standard normal distribution N(0;1), which can be found in Table T1.

Example 5.3

When asked about a new product by a marketing research agency, 80 out of the 400 customers of the STAMET supermarket answered that they would buy it. Calculate the minimum variance unbiased point estimator and find an α confidence interval for the ratio p of such customers to all the STAMET customers.

Solution:

Since x = 80 and n = 400, the point estimator assumes the value p = $\frac{80}{400}$ = 0.2, or 20% of the customers.

For confidence 0.95 we have $u_{0.975} = 1.960$, which yields the following 0.95 confidence interval for p

$$p \in <\frac{80}{400} - 1.960\sqrt{\frac{\frac{80}{400}(1-\frac{80}{400})}{400}}; \frac{80}{400} + 1.960\sqrt{\frac{\frac{80}{400}(1-\frac{80}{400})}{400}} > =...=$$

$$= < 0.1608; 0.2392 >.$$

Similarly, a 0.99 confidence interval is calculated to be

$$p \in < 0.1485; 0.2515 >.$$

We can say with a 0.95 or 0.99 confidence that about 16% to 24% or 15% to 25% of the STAMET customers will buy the new product. If there are about 10 000 STAMET customers it may be expected that 2 000 products will be sold. A 0.95 confidence interval tells us that STAMET will sell approximately 10 000(0.16) = 1600 to 10 000(0.24) = 2400 new products.

Testing statistical hypotheses

Statistical hypothesis and its testing

When observing random variables we must often test certain conditions or assumptions on their properties using experiment data. For example we might want to make a decision whether a parameter ϑ has the value ϑ_o, writing $H: \vartheta = \vartheta_o$. If $\bar{H}: \vartheta \neq \vartheta_o$, we call it a <u>two-tailed</u> alternative hypothesis and if $\bar{H}: \vartheta \rangle \vartheta_o$ or $\vartheta \langle \vartheta_o$, we have a <u>one-tailed</u> alternative hypothesis.

To test a hypothesis $H: \vartheta = \vartheta_o$ a suitable statistic $T(X_1, ..., X_n)$ is constructed the so-called <u>test criterion or test statistic</u>. For $\vartheta = \vartheta_o$, the range of values of the test criterion is divided into two disjunct subsets - the critical range W_α and its complement $\overline{W}_\alpha$ in such a way that, for $\vartheta = \vartheta_o$, the probability of $T(X_1, ..., X_n)$ taking on a value from $\overline{W}_\alpha$ is 1 - α . The number $\alpha > 0$ is called a level of significance or level of significance and is chosen close to zero, usually 0.05 or 0.01.

If, for sample data $(x_1, ..., x_n)$, the test criterion assumes a value t = $= T(x_1, ..., x_n)$ from the critical range that is $t \in W_\alpha$, we reject the hypothesis $\bar{H}$ and do not reject the hypothesis H. If, on the other hand, t lies outside the critical range or $t \in \overline{W}_\alpha$, we <u>reject</u> the hypothesis $\bar{H}$ and <u>do not reject</u> the hypothesis H at the level of significance α. If a hypothesis H or $\bar{H}$ is rejected, it does not necessarily mean that its validity has been proved since via the random sample we have only acquired information which does not suffice to reject it. Whenever possible, the size of the sample data should be increased before accepting a given hypothesis H and it should be tested again.

When testing a hypothesis H, the following errors may be made

(1) the so called <u>first type error</u>, when H is true but $t \in W_\alpha$, so that we reject it (the probability of such an error is $P(T \in W_\alpha / H) = \alpha$),

(2) the so called second type error, when H is not true but $t \notin W_\alpha$, so that we do not reject it (the probability of this error is $P\left(T \notin W_\alpha / \overline{H}\right)$).

Since a test criterion T is a random variable, the range $\overline{W}_\alpha$ is often in the form of an interval $\left\langle t_1 ; t_2 \right\rangle$ where t_1, t_2 are quantiles of the test statistic T (the so-called critical values) as in confidence intervals. You can find more about statistical hypotheses and their testing in [1], [2], [3] and [4].

Testing hypotheses concerning parameters of normal distribution

We assume that random variables and vectors are normally distributed. The following testing criteria are for two-tailed alternative hypotheses such as $\overline{H} : \mu \neq \mu_0$ except for the variance equality test.

Assumptions on distribution types may be tested using the so-called goodness of fit tests (such as Pearson's test or Kolmogorov's test) or we can use a rough graphical test using the so-called probability paper or its graphical version on a computer. These tests and further details on testing hypotheses with one-tailed alternative hypotheses tests about parameters of other distributions and other tests can be found in [1], [2] and [4].

Testing a hypothesis $H : \mu = \mu_0$ with unknown variance σ^2. The test criterion is calculated using the formula

$$t = \frac{\bar{x} - \mu_0}{s} \sqrt{n-1},$$

and $\overline{W}_\alpha = \left\langle -t_{1-\alpha/2} ; t_{1-\alpha/2} \right\rangle$ where $t_{1-\alpha/2}$ is the $\left(1 - \alpha/2\right)$-quantile of the t-distribution with n - 1 degrees of freedom. These values can be found in Table T2. This is the so-called one sample t - test.

E x a m p l e 5.4

By measuring the lengths of 10 rollers, empirical characteristics $\bar{x}$ = 5.37 mm and s^2 = 0.0019 mm^2 have been established (see Problem 2.1). At the 0.05 level of significance test the hypothesis that the mean value of the roller length measured is 5.40 mm, and so H : μ = 5.40.

S o l u t i o n:

The test criterion assumes the value

$$t = \frac{5.37 - 5.40}{\sqrt{0.0019}}\sqrt{10-1} \cong -2.0647.$$

For 10 - 1 = 9 degrees of freedom, we find $t_{0.975}$ = 2.262 in Table T2 so that $\overline{W}_{0.05}$ = = <-2.262; 2.262>. Since t $\in$ $\overline{W}_{0,05}$, we do not reject the hypothesis. To test this hypothesis we could also use the 0.95 confidence interval from Example 5.1. Since this interval includes the hypothetical value of 5.40, we do not reject the hypothesis at the level of significance 1 - 0.95 = 0.05.

Testing hypothesis H : $\sigma^2 = \sigma_0^2$. The test criterion is given by

$$t = \frac{ns^2}{\sigma_0^2}$$

and $\overline{W}_\alpha = \left\langle \chi^2_{\alpha/2}; \chi^2_{1-\alpha/2} \right\rangle$ where χ^2_P is a P-quantile of the chi-squared distribution with n - 1 degrees of freedom listed in Table T3.

E x a m p l e 5.5

At the 0.05 level of significance test a hypothesis that the variance of the value of the roller length measured in Example 5.2 is 0.0025 mm^2 so that H : σ^2 = 0.0025.

S o l u t i o n:

The test criterion assumes the value

$$t = \frac{10(0.0019)}{0.0025} = 7.6 .$$

For 10 - 1 = 9 degrees of freedom we determine $\chi^2_{0.025}$ = 2.70 and $\chi^2_{0.975}$ = 19.0 from Table T3 so that $\overline{W}_{0.05}$ = <2.70; 19.0>. Since t $\in$ $\overline{W}_{0.05}$, we do not reject the hypothesis.

Testing hypothesis H : $\rho = \rho_0$. The value of the test criterion is calculated for n $\geq$ 10, $|r| \neq 1$ and $|\rho_0| \neq 1$ using the formula

$$t = \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{n-1} \right) \frac{\sqrt{n-3}}{2}$$

and $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2} \, ; u_{1-\alpha/2} \right\rangle$ where $u_{1-\alpha/2}$ is the $\left(1 - \alpha/2\right)$ - quantile of the normal distribution N(0, 1), which can be found in Table T1.

E x a m p l e 5.6

By monitoring the costs X and prices Y of an identical product with ten manufacturers, a two-dimensional statistical data have been collected and a sample correlation coefficient r = 0.82482 calculated (see Example 2.3). At the 0.01 level of significance, test the hypothesis that the random variables X and Y are not correlated (independent with respect to normal distribution) so that H : ρ = 0.

S o l u t i o n:

The test criterion assumes the value

$$t = \left(\ln \frac{1+0.82482}{1-0.82482} - \ln \frac{1+0}{1-0} - \frac{0}{10-1} \right) \frac{\sqrt{10-3}}{2} \cong 3.1001.$$

For the given level of significance we find $u_{0.995}$ = 2.576 in Table T1 so that $\overline{W}_{0.01}$ = = <–2.576; 2.576 >. Since t $\notin \overline{W}_{0.01}$, we reject the hypothesis and consider X, Y as dependent.

Testing hypothesis H : $\mu(X) = \mu(Y)$ for pairs. Denote by (x_i, y_i), where i = 1,…,n, the values of pairs observed in the random vector (X, Y), by $d_i = x_i - y_i$ their differences and by $\bar{d}$ and $s^2(d)$ their empirical characteristics . The test criterion is given by

$$t = \frac{\bar{d}}{s(d)}\sqrt{n-1}$$

and $\overline{W}_\alpha = \left\langle -t_{1-\alpha/2} ; t_{1-\alpha/2} \right\rangle$, where $t_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the chi-squared distribution with n -1 degrees of freedom, which can be found in Table T2. This is the so-called t - test for paired values.

E x a m p l e 5.7

Using two thermometers the following pairs of temperature values have been measured over eight days: $(x_i; y_i)$ = (51.8; 49.5), (54.9; 53.3), (52.2; 50.6), (53.3; 52.0), (51.6; 46.8), (54.1; 50.5), (54.2; 52.1), (53.3; 53.0) (^oC). At the level of significance 1%, test the hypothesis that the difference of the mean values is insignificant so that H : $\mu(X) = \mu(Y)$.

S o l u t i o n:

For $d_i = x_i - y_i$, i = 1,...,8, we get $\bar{d}$ = 2.2 ^oC and s(d) = 1.3172 ^oC. The test criterion assumes the value

$$t = \frac{2.2}{1.3172}\sqrt{8-1} \cong 4.4190.$$

For 8 - 1 = 7 degrees of freedom we have $t_{0.995}$ = 3.499 from Table T2 so that $\overline{W}_{0.01}$ = = <-3.499; 3.499>. Since $t \notin \overline{W}_{0.01}$, we reject the hypothesis at the level of significance 1%. The difference between the two measurements is statistically significant.

For the following tests we assume that by observing two independent random variables X and Y normally distributed with parameters $\mu(X)$, $\sigma^2(X)$ and $\mu(Y)$, $\sigma^2(Y)$ sample data of sizes n_1 and n_2 have been obtained.

Testing hypothesis H : $\mu(X) - \mu(Y) = \mu_0$ with unknown variances. The test criterion is calculated as follows

$$t = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{n_1 s^2(x) + n_2 s^2(y)}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

and $\overline{W}_\alpha = \langle -t_{1-\alpha/2} ; t_{1-\alpha/2} \rangle$ where $t_{1-\alpha/2}$ is the $\left(1 - \alpha/2\right)$ - quantile of the chi-squared distribution with $n_1 + n_2 - 2$ degrees of freedom listed in Table T2. This is the so-called two-sample t - test.

Example 5.8

By testing the strength of wire manufactured by two different technologies two sample data have been obtained with the following sample characteristics n_1 = 33, $\bar{x}$ = 5.4637 kN, $s^2(x)$ = 0.3302 kN^2, n_2 = 28, $\bar{y}$ = 6.1179 kN, $s^2(y)$ = 0.4522 kN^2. At the level of significance 0.05 test the hypothesis that the different technologies do not affect the expected value of the wire strength (assuming that the variances σ^2 (X) and σ^2 (Y) are the same) so that H : μ(X) - μ(Y) = 0.

Solution:

The test criterion assumes the value

$$t = \frac{5.4637 - 6.1179 - 0}{\sqrt{33(0.3302) + 28(0.4522)}} \sqrt{\frac{33(28)(33 + 28 - 2)}{33 + 28}} \cong 4.030.$$

For 33 + 28 - 2 = 59 degrees of freedom we get $t_{0.975}$ = 2.001 by interpolating in Table T2 so that $\overline{W}_{0.05}$ = <-2.001; 2.001>. Since $t \notin \overline{W}_{0.05}$, we reject the hypothesis. The different technologies do affect the expected value of the wire strength.

Testing hypothesis H : $\mu(X) - \mu(Y) = \mu_0$ for unknown variances $\sigma^2(X) \neq \sigma^2(Y)$.

The test criterion is calculated as

$$t = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{\frac{s^2(x)}{n_1 - 1} + \frac{s^2(y)}{n_2 - 1}}}$$

and $\overline{W}_\alpha = \langle -\bar{t}_{1-\alpha/2}; \bar{t}_{1-\alpha/2} \rangle$ where

$$\bar{t}_{1-\alpha/2} = \frac{\frac{s^2(x)}{n_1 - 1} t(x) + \frac{s^2(y)}{n_2} t(y)}{\frac{s^2(x)}{n_1 - 1} + \frac{s^2(y)}{n_2 - 1}}$$

and t(x) or t(y), is the $\left(1 - \alpha/2\right)$ - quantile of the chi-squared distribution with $n_1 - 1$ or $n_2 - 1$ degrees of freedom respectively shown in Table T2. This is the so-called <u>two-sample t - test</u>.

<u>E x a m p l e 5.9</u>

Surveys made to determine the mean service life of products in two different systems of extreme conditions yielded two sets of sample data with the following sample characteristics $n_1 = 21$, $\bar{x} = 3.581$, $s^2(x) = 0.114$, $n_2 = 23$, $\bar{y} = 3.974$, $s^{2(}y) = 0.041$ (the length of the life is in hours). At the level of significance 0.05, test the hypothesis that the second system of extreme conditions increases the mean service life by 0.5 h as compared with the first one (assuming different variances $\sigma^2(X)$ and $\sigma^2(Y)$) so that H : $\mu(X) - \mu(Y) = - 0.5$.

S o l u t i o n:

The test criterion assumes the value

$$t = \frac{3.581 - 3.974 - (-0.5)}{\sqrt{\frac{0.114}{21 - 1} + \frac{0.041}{23_2 - 1}}} \cong 1.2303.$$

Entering Table T2 at $1 - \alpha/2 = 0.975$ we see that t(x) = 2.086 for 21 - 1 = 20 degrees of freedom and t(y) = 2.074 for 23 - 1 = 22 degrees of freedom. Then we can calculate

$$\bar{t}_{0,975} = \frac{\frac{0.114}{21 - 1} 2.086 + \frac{0.041}{23 - 1} 2.074}{\frac{0.114}{21 - 1} + \frac{0.041}{23 - 1}} \cong 2.083.$$

and $\overline{W}_{0.05}$ = <-2.083; 2.083>. Since $t \in \overline{W}_{0.05}$, we do not reject the hypothesis that the second system increases the mean service life by 0.5 h.

Testing hypothesis $H : \sigma^2(X) = \sigma^2(Y)$ against alternative hypothesis $\overline{H} : \sigma^2(X) > \sigma^2(Y)$. The test criterion is calculated as follows

$$t = \frac{\frac{n_1 s^2(x)}{n_1 - 1}}{\frac{n_2 s^2(y)}{n_2 - 1}} \geq 1$$

and $\overline{W}_\alpha = \langle 1; F_{1-\alpha} \rangle$ where $F_{1-\alpha}$ is the $(1-\alpha)$ - quantile of the Fisher - Snedecor distribution with $k_1 = n_1 - 1$ and $k_2 = n_2 - 1$ degrees of freedom listed in table T4. If $t < 1$ we swap X and Y. This is the so-called F - test or Fisher's test. It is used to test assumptions on variances made in the previous tests.

E x a m p l e 5.10

At the level of significance 0.05, test the hypothesis that the variances σ^2 (X) > σ^2 (Y) are different in Example 5.9 where $s^2(x) = 0.114$, $n_1 = 21$, $s^{2(}y) = 0.041$, $n_2 = 23$.

S o l u t i o n:

Let us, on the contrary, test the hypothesis H : σ^2 (X) = σ^2 (Y).

The test criterion assumes the value

$$t = \frac{\frac{21(0.114)}{21-1}}{\frac{23(0.041)}{23-1}} \cong 2.7926.$$

For k_1 = 21 - 1 = 20 and k_2 = 23 - 1 = 22 degrees of freedom we obtain $F_{0.95}$ = 2.071 by interpolating in Table T4 so that $\overline{W}_{0.05}$ = <1; 2.071>. Since $t \notin \overline{W}_{0.05}$, we reject the hypothesis. We consider the assumption that the variances are different as correct.

Testing hypotheses on the parameter of a binomial distribution

Let us observe a random variable X which has an alternative probability distribution with a parameter p, thus it has the binomial distribution Bi(1; p). When testing a hypothesis H : p = p_0 we actually test a hypothesis that the ratio of those elements of the parent population that have a desired property is p_0 on the basis of

the fact that x elements out of the n elements in a random sample have the property (see estimations of parameters).

Testing hypothesis H : $p = p_0$. The testing criterion is calculated for $n > 30$ as follows

$$t = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2} ; u_{1-\alpha/2} \right\rangle$ where $u_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the normal distribution N(0, 1) which can be found in Table T1.

E x a m p l e 5.11

According to an expert's opinion 20% of the customers will be interested in a new product. Out of the 400 customers asked, 62 showed interest. At the 0.05 significance level test the hypothesis that the expert's forecast is real, that is H : p = = 0.2.

S o l u t i o n:

For x = 62 and n = 400 the testing criterion assumes the value

$$t = \frac{\frac{62}{400} - 0.2}{\sqrt{\frac{(0.2)(1-0.2)}{400}}} = \frac{-0.045}{0.02} = -2.25 .$$

From Table T1 we get $u_{0.975} = 1.960$. Since $t = -2.25 \notin \overline{W}_{0,05} = \langle -1.960; 1.960 \rangle$, we reject the hypothesis that 20% of customers will be interested at the 0.05 significance level. The real interest will probably be lower. Note that we would not reject the hypothesis at the 0.01 significance level since $u_{0.995} = 2.576$.

For the following test we assume that we observe two independent random variables X, Y which have s alternative distributions with parameters p_1, p_2 and that two independent sample data have been obtained of sizes n_1, n_2 respectively and

the respective numbers x, y of elements with the desired property (see estimations of parameters).

Testing hypothesis H : $p_1 = p_2$. The testing criterion can be calculated as shown below on the assumption that $n_1 > 50$ a $n_2 > 50$

$$t = \frac{\frac{x}{n_1} - \frac{y}{n_2}}{\sqrt{\bar{f}(1-\bar{f})}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

for $\bar{f} = \frac{x+y}{n_1+n_2}$ and $\overline{W}_\alpha = \left\langle -u_{1-\alpha/2} ; u_{1-\alpha/2} \right\rangle$ where $u_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the normal distribution N(0, 1) with values shown in Table T1.

E x a m p l e 5.12

Shop inspectors bought 250 items of food and 200 items of hard goods to test their quality. Subsequently, they found 108 items of food and 73 items of hard goods to be defective. At the 0.05 significance level test, if the quality of food and hard goods is equal, that is the hypothesis H: $p_1 = p_2$ where p_1, p_2 are the theoretical ratios (probabilities) of buying defective items for the given kinds of goods.

S o l u t i o n:

For x = 108, n_1 = 250, y = 73, n_2 = 200 we get

$$\bar{f} = \frac{108+73}{250+200} \cong 0.40222$$

and the testing criterion assumes the value

$$t = \frac{\frac{108}{250} - \frac{73}{200}}{\sqrt{0.40222(1-0.40222)}} \sqrt{\frac{250(200)}{250+200}} \cong \frac{0.067(10.5409)}{0.49035} \cong 1.4403.$$

We establish $u_{0.975}$ = 1.960 from Table T1. Since t = 1.4403 $\in$ $\overline{W}_{0.05}$ = <–1.960; 1.960>, at the 0.05 significance level we do not reject the hypothesis that the probabilities of buying a defective item are the same for both kinds of goods and consider both kinds of goods to be of equally bad quality.

Regression analysis

Regression function

An important role of statistics is to find and analyse dependencies of variables whose values we observe when conducting experiments. Because of their random character, the independent variables are represented by a random vector $\mathbf{X} = (X_1,...,X_k)$ and the dependent variable is represented by a random variable Y. We employ regression analysis to determine the dependence of Y on **X** and this dependence is expressed by a regression function

$$y = \varphi(\mathbf{x}, \boldsymbol{\beta}) = E(Y / \mathbf{X} = \mathbf{x}),$$

where $\mathbf{x} = (x_1,...,x_k)$ is a vector of independent variables (the value of **X**), y is a dependent variable (the value of Y) and $\boldsymbol{\beta} = (\beta_1,...,\beta_m)$ is a vector of parameters, the so-called regression coefficients. **X** can also be a non-random vector as it is often the case in applications, for example if the variances of the vector constituents X_j, $j = 1, \ldots, m$, can be neglected as compared to the variance of Y.

When determining the dependence of Y on **X** we conduct experiments to obtain a sequence of points $[\mathbf{x}_1,y_1],...,[\mathbf{x}_n,y_n]$ where y_i is the result of measuring Y_i and $\mathbf{x}_i$ is the resulting value of measuring the vector of independent variables, i = = 1,..., n. To obtain estimates of the regression coefficients β_j, j = 1,..., m, we minimize the so-called residual sum of squares

$$S^* = \sum_{i=1}^{n} \left[y_i - \varphi(\mathbf{x}_i, \boldsymbol{\beta}) \right]^2$$

using the least-square method.

Linear regression function

A linear regression function (linear with respect to the regression coefficients) can be written in the form

$$y = \sum_{j=1}^{m} \beta_j f_j(\mathbf{x}),$$

where $f_j(\mathbf{x})$ are known functions not containing $\beta_1, \ldots, \beta_m$.

Let us consider a model based on the following assumptions:

1. The vector **x** is non-random, so that the functions $f_j(\mathbf{x})$ assume non-random values $f_{ji} = f_j(\mathbf{x}_i)$ for j = 1,..., m and i = 1,..., n.
2. The matrix $\mathbf{F} = (f_{ji})$ is an (m, n)-matrix with rank m < n.
3. For i = 1,..., n, random variable Y_i has the expected value $E(Y_i / \mathbf{X_i} = \mathbf{x_i}) = \sum_{j=1}^{m} \beta_j f_{ji}$ and a constant variance $D(Y_i / \mathbf{X}_i = \mathbf{x}_i) = \sigma^2 > 0$.
4. Random variables Y_i are not correlated and are normally distributed for i = 1,..., n.

Using the following denotations

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}, \quad \mathbf{H} = \mathbf{F}\mathbf{F}^T, \quad \mathbf{g} = \mathbf{F}\mathbf{y},$$

we get:

a) For j = 1,..., m, β_j is the minimum variance unbiased <u>point estimate for the regression coefficient</u> b_j where **b** is the solution of the system of linear algebraic equations (the so-called <u>system of normal equations</u>)

$$\mathbf{H}\mathbf{b} = \mathbf{g}.$$

b) <u>The minimum variance unbiased point estimate for the linear regression function</u> is

$$y = \sum_{j=1}^{m} b_j f_j(\mathbf{x}).$$

c) <u>The minimum variance unbiased point estimate for variance</u> σ^2 is

$$s^2 = \frac{S^*_{min}}{n - m},$$

where

$$S^*_{min} = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} b_j f_{ji} \right)^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=1}^{m} b_j g_j$$

and g_j is an element of matrix **g**.

d) The (1 - α) - confidence interval for the regression coefficient β_j is

$$\left\langle b_j - s\sqrt{h^{jj}}\ t_{1-\alpha/2}\ ;\ b_j + s\sqrt{h^{jj}}\ t_{1-\alpha/2} \right\rangle,$$

j = 1, ..., m, where h^{jj} is the j–th diagonal element of the matrix $\mathbf{H}^{-1}$ and $t_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the t–distribution with n - m degrees of freedom.

e) The (1 - α) - confidence interval for the expected value of y is

$$\left\langle \sum_{j=1}^{m} b_j f_j(\mathbf{x}) - t_{1-\alpha/2} s\sqrt{h^*}\ ;\ \sum_{j=1}^{m} b_j f_j(\mathbf{x}) + t_{1-\alpha/2} s\sqrt{h^*} \right\rangle,$$

where $h^* = \mathbf{f}(\mathbf{x})^T \mathbf{H}^{-1} \mathbf{f}(\mathbf{x})$ and $t_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the t-distribution with n - m degrees of freedom. A (1 - α) - confidence interval for an individual function value of y is calculated in a similar way taking $1 + h^*$ instead of h^*.

f) To test the hypothesis $H: \beta_j = \beta_{j0}$ with an α level of significance where j is an arbitrary index, j = 1,..., m, the testing criterion must be calculated

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{h^{jj}}}$$

and $\overline{W}_\alpha = \left\langle -t_{1-\alpha/2}\ ;\ t_{1-\alpha/2} \right\rangle$, where $t_{1-\alpha/2}$ is the $\left(1-\alpha/2\right)$ - quantile of the t – distribution with n - m degrees of freedom.

The simplest and the most used linear regression function is the so-called regression line

$$y = \beta_1 + \beta_2 x\,.$$

For this function, k = 1, $\mathbf{x} = x_1 = x$ (we write x instead of x_1), m = 2, $f_1(x) = 1$, $f_2(x) = x$, so that

$$\mathbf{F} = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

When calculating with a pocket calculator, we can use the following explicit formulas ($\sum$ stands for $\sum_{i=1}^{n}$):

1) $\mathbf{H} = \begin{pmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$, $\mathbf{g} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$, $\sum 1 = n$,

2) $\det \mathbf{H} = n\sum x_i^2 - \left(\sum x_i\right)^2$, $b_2 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{\det \mathbf{H}}$, $b_1 = \bar{y} - b_2 \bar{x}$,

3) $S^*_{min} = \sum (y_i - b_1 - b_2 x_i)^2 = \sum y_i^2 - b_1 \sum y_i - b_2 \sum x_i y_i$, $s^2 = \dfrac{S^*_{min}}{n-2}$,

4) $h^{11} = \dfrac{\sum x_i^2}{\det \mathbf{H}}$, $h^{22} = \dfrac{n}{\det \mathbf{H}}$,

5) $h^* = \dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{\sum x_i^2 - n(\bar{x})^2} = \dfrac{1}{n} + \dfrac{n(x - \bar{x})^2}{\det \mathbf{H}}$,

6) $r = \sqrt{1 - \dfrac{S^*_{min}}{\sum y_i^2 - n(\bar{y})^2}} = |r(x, y)|$, where r(x,y) is the correlation coefficient – see Chapter 2.

Example 5.13

A survey at eight tax consultancy companies chosen at random produced the following numbers of employees x and annual sales y (in millions of CZK):

x_i	3	5	5	8	9	11	12	15
y_i	0.8	1.2	1.5	1.9	1.8	2.4	2.5	3.1

Find out how the annual sales of a company depend on the number of its employees. Using the formula $y = \beta_1 + \beta_2 x$, calculate a 0.95 confidence interval for β_2, with a 0.05 level of significance test the hypothesis H : $\beta_1 = 0.2$, find the minimum variance unbiased point estimate and a 0.95 confidence interval for y(10). Using a graph and the correlation coefficient analyse the suitability of the regression function.

S o l u t i o n:

The table below contains auxiliary calculations:

i	x_i	y_i	x_i^2	x_iy_i	y_i^2
1	3	0.8	9	2.4	0.64
2	5	1.2	25	6.0	1.44
3	5	1.5	25	7.5	2.25
4	8	1.9	64	15.2	3.61
5	9	1.8	81	16.2	3.24
6	11	2.4	121	26.4	5.76
7	12	2.5	144	30.0	6.25
8	15	3.1	225	46.5	9.61
Σ	68	15.2	694	150.2	32.80

This is a regression line and so using the above formulas and table, for n = 8, we can set up the matrix $\mathbf{H} = \begin{pmatrix} 8 & 68 \\ 68 & 694 \end{pmatrix}$ with det **H** = (8)(694) - 68^2 = 928. This yields the point estimate for β_2:

$$b_2 = \frac{(8)(150.2) - (68)(15.2)}{928} = 0.1810344 \cong 0.181.$$

Next we have $\bar{x}$ = 68/8 = 8.5 and $\bar{y}$ = 15.2/8 = 1.9, so that the point estimate for β_1 is

$$b_1 = 1.9 - (0.1810344)(8.5) = 0.3612068 \cong 0.361.$$

Thus we get the point estimate for the regression function y = 0.361 + 0.181x.

The minimum value of the residual sum of squares is

$$S^*_{min} = 32.80 - (0.3612068)(15.2) - (0.1810344)(150.2) \cong 0.1182758$$

and the point estimates for σ^2 and σ, are the following

$$s^2 = 0.1182758/(8-2) = 0.0197126, \text{ resp. } s = \sqrt{0.0197126} \cong 0.1404017.$$

The diagonal elements of $\mathbf{H^{-1}}$ are $h^{11} = 694/928 \cong 0.7478448$ and $h^{22} = 8/928 \cong$ $\cong 0.00862069$. In Table T2, for 8 - 2 = 6 degrees of freedom, we find $t_{0.975} = 2.447$. The 0.95 confidence interval for the regression coefficient β_2 is

$$\beta_2 \in \langle 0.1810344 - (2.447)(0.1404017)\sqrt{0.00862069};$$

$$0.1810344 + (2.447)(0.1404017)\sqrt{0.00862069} \rangle =$$

$$= \langle 0.1491353; 0.2129334 \rangle \cong \langle 0.149; 0.213 \rangle.$$

We have found a point estimate of 181 000 CZK for the increase in annual sales corresponding to an increase in the number of employees by one. A 0.95 confidence interval for this value is 149 000 CZK to 213 000 CZK.

For $H: \beta_1 = 0.2$ the testing criterion assumes the value

$$t = \frac{0.3612068 - 0.2}{0.1404017\sqrt{0.7478448}} \cong 1.3277.$$

For the alternative hypothesis $\overline{H}: \beta_1 \neq 0.2$ we get $\overline{W}_{0,05} = \langle -2.447; 2.447 \rangle$. Since $t \in \overline{W}_{0.05}$, we do not reject the hypothesis $\beta_1 = 0.2$ at a 0.05 level of significance. As it is, at this level of significance we would not actually reject the hypothesis that a company without employees (the owners themselves work), since $y(0) = \beta_1$, will have annual sales of about 200 000 CZK.

The point estimate for the average and individual sales at a company that has ten employees is the following

$$y(10) = 0.3612068 + (0.1810344)(10) = 2.1715508 \cong 2.172.$$

For the given company annual sales of about 2 172 000 CZK may be expected Since

$$h^* = \frac{1}{8} + \frac{8(10 - 8.5)^2}{928} = 0.1443965,$$

is a 0.95 confidence interval for the annual sales of a company with ten employees,

$$y(10) \in \langle 2.1715508 - (2.447)(0.1404017)\sqrt{0.1443965};$$

$$2.1715508 + (2.447)(0.1404017)\sqrt{0.1443965} \rangle =$$

$$= \langle 2.0409985; 2.3021031 \rangle \cong \langle 2.040; 2.302 \rangle.$$

With probability 0.95 it may be expected that the average annual sales of such a company will range between 2 040 000 CZK and 2 302 000 CZK.

If we employ $1 + h^*$ instead of h^* for the calculation, we get a 0.95 confidence interval for the value of the annual sales of a firm with ten employees

$$y(10) \in \langle 2.1715508 - (2.447)(0.1404017)\sqrt{1.1443965}\,;$$

$$2.1715508 + (2.447)(0.1404017)\sqrt{1.1443965}\rangle \cong \langle 1.804;\ 2.539\rangle.$$

With a 0.95 probability it may be expected that the annual sales (its individual value) of such a company will range from 1 804 000 CZK to 2 539 000 CZK.

The correlation coefficient (calculated by the formula in Chapter 2) is $r = 0.984798$. It follows from the graph in Figure 5.1 and from the value of the correlation coefficient that the form of the regression function chosen fits the given dependence quite well.

Dependence of annual sale on number of employees

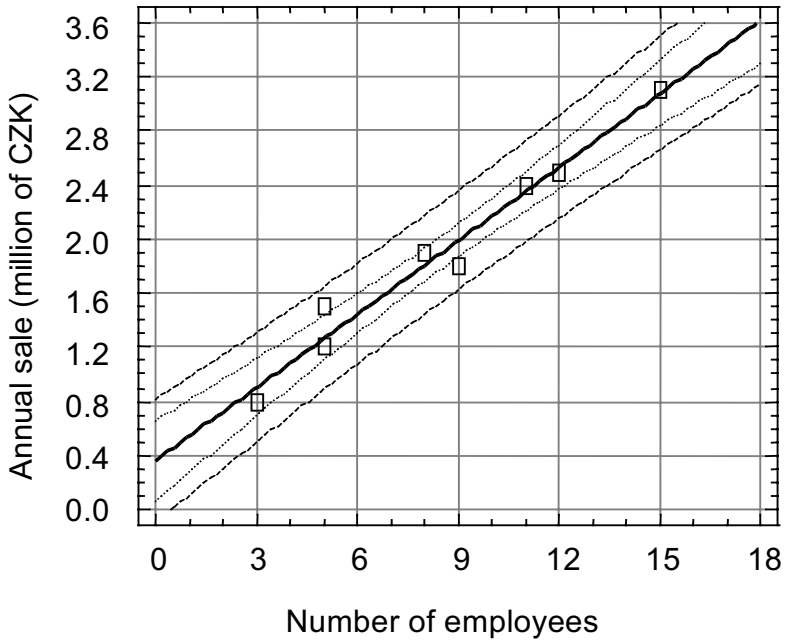


Fig. 5.1

The regression functions are either linear or nonlinear (with respect to their regression coefficients). Some of the nonlinear regression functions can be

transformed into linear ones using a suitable linearization (such as taking a logarithm of a power or an exponential function). You will find more details about linearization, tests of the suitability of a linear regression function, regression diagnostics and other topics in [1], [2], [3] and [4].

Exercises

Exercise 5.14

Find the minimum variance unbiased point estimate and a 0.99 confidence interval for the parameters μ and σ^2 of a normal distribution, using sample data from a random sample of size n = 18 with the sample mean $\bar{x}$ = 50.1 and variance s^2 = 17.64.

Result: $\mu = 50.1$; $\sigma^2 \cong 18.678$; $\mu \in \langle 47.09; 53.10 \rangle$; $\sigma^2 \in \langle 8.894; 55.705 \rangle$

Exercise 5.15

Sample data of size n = 12 have sample mean $\bar{x}$ = 77.55 and variance s^2 = 1045.65. Calculate the point estimator and find a 0.99 confidence interval for μ and σ of the parent population.

Result: $\mu = 77.55$; $\sigma \cong 33.78$; $\mu \in \langle 47.267; 107.833 \rangle$; $\sigma \in \langle 21.638; 69.47 \rangle$

Exercise 5.16

A total of 100 workers of the same category selected at random have been asked about their wages per hour and the empirical characteristics $\bar{x}$ = 28.64 CZK and s^2 = = 1.1979 CZK have been calculated. Find the minimum variance unbiased point estimate and a 99% confidence interval for the expected value of wages per hour μ and standard deviation σ provided that the parent population is normally distributed.

Result: μ = 28.64 CZK; $\sigma \cong$ 1.10 CZK;

$\mu \in \langle 28.35; 28.93 \rangle$ CZK; $\sigma \in \langle 0.93; 1.34 \rangle$ CZK

E x e r c i s e 5.17

From fifteen independent measurements of the maximum velocity of an aeroplane the minimum variance unbiased point estimates 424.7 ms^{-1} for the mean value and 8.7 ms^{-1} for the standard deviation have been calculated. Find a 95% confidence interval for the expected value and standard deviation of the maximum speed provided that it is normally distributed.

R e s u l t: $\mu \in \langle 419.88; 429.52 \rangle$ ms^{-1} ; $\sigma \in \langle 6.37; 13.72 \rangle$ ms^{-1}

E x e r c i s e 5.18

A total of fifty values of a random variable with a normal distribution $N(\mu,\sigma^2)$ has been used to calculate a sample mean $\bar{x} = 610$ and a variance $s^2 = 2770.4$. Find interval estimates for μ with confidences of 0.9; 0.95; and 0.99.

R e s u l t: <597.45; 622.55>; <594.97; 625.07>; <592.77; 627.33>

E x e r c i s e 5.19

Five independent and equally accurate measurements have been carried out to determine the volume of a vessel: 4.781; 4.792; 4.795; 4.779; 4.769 (in litres). Find a 0.99 level of significance confidence interval for the expected value of the volume of vessel provided that it is normally distributed.

R e s u l t: <4.761; 4.805> l

E x e r c i s e 5.20

A sample of size n = 128 has been used to determine the correlation coefficient and the result was r = 0.560. Calculate the minimum variance unbiased point estimator and find a 95% confidence interval for ρ.

R e s u l t: $\rho = 0.560$; $\rho \in \langle 0.441; 0.679 \rangle$

E x e r c i s e 5.21

In a total of 46 households chosen at random a survey has been made to find the relationship between income X and expenses Y. A sample correlation coefficient of

r = 0.638 was calculated. Calculate the minimum variance unbiased point estimator and determine a 0.95 confidence interval for the correlation coefficient provided that income and expenses have a two-dimensional normal distribution.

R e s u l t: $\rho = 0.638$; $\rho \in \langle 0.423 ; 0.780 \rangle$

E x e r c i s e 5.22

Questionnaires have been sent to 190 customers of a pension fund chosen at random with questions about their annual income X, premium amount Y and employment time Z. Calculations have been performed to determine the correlation coefficients with the following results: $\rho(X,Y) = 0.55$; $\rho(X,Z) = 0.30$; $\rho(Y,Z) = 0.37$. Find a 99% level of significance confidence intervals of the above coefficients assuming a normal distribution.

R e s u l t: $\rho(X,Y) \in \langle 0.42; 0.68 \rangle$; $\rho(X,Z) \in \langle 0.13; 0.47 \rangle$; $\rho(Y,Z) \in \langle 0.21; 0.53 \rangle$

E x e r c i s e 5.23

During a check of expiry periods of a certain type of tinned meat in a food industry warehouse, a total of 320 tins have been chosen at random. It has been found that in 59 of them the guarantee periods have expired. Calculate the minimum variance unbiased point estimate and find a 95% level of significance confidence interval for the percentage of tins in the warehouse with expired guarantee periods. Do the same for a warehouse with 20 000 tins.

R e s u l t: $p = 59/320 \cong 0.184 = 18.4\ \%$; $p \in \langle 0.142; 0.226 \rangle = \langle 14.2; 22.6 \rangle\ \%$;
$N = 3680$; $N \in \langle 2840; 4520 \rangle$

E x e r c i s e 5.24

Within a random sample from tyres produced by a large European multi-national company, 10% of the tyres do not comply with the rules of a new standard. Find a 95% confidence interval for the percentage p (of the whole parent population) of tyres that do not comply with the new standard if the size of the random sample is a) n = = 100, b) n = 400, c) n = 1600.

R e s u l t: <0.041; 0.159>; <0.071; 0.129>; <0.085; 0.115>

E x e r c i s e 5.25

Sample data of size n = 10 provide a sample mean of $\bar{x} = 32$ and a variance of $s^2 = 15$. At a 0.05 level of significance test the hypothesis that the expected value of the parent population is $\mu = 30$.

R e s u l t: $t \cong 1.549$; $t_{0.975} = 2.262$; we do not reject the hypothesis

E x e r c i s e 5.26

The following sample data have been provided by a random sample from normal distribution.

x_j^*	-2	-1	0	1	2	3
f_j	1	4	7	3	3	2

At a 0.05 level of significance test the hypothesis $\mu = 0.1$.

R e s u l t: $\bar{x} = 0.45$; $s = 1.3592$; $t \cong 1.1224$; $t_{0.975} = 2.093$; we do not reject the hypothesis

E x e r c i s e 5.27

The expected value of humidity in roasted coffee is set at 4.2% and the standard deviation at 0.4%. The actual percentages of humidity determined by analyzing 20 samples are the following: 4.5; 4.3; 4.1; 4.9; 4.6; 3.2; 4.4; 5.1; 4.8; 4.0; 3.7; 4.4; 3.9; 4.1; 4.2; 4.1; 4.7; 4.3; 4.2; 4.4. At a 5% level of significance test the hypotheses that

a) the expected value of humidity for the parent population complies with the standard and

b) the standard deviation of humidity for the parent population complies with the standard.

R e s u l t: a) $t \cong 1.033$; $t_{0.975} = 2.093$; we do not reject the hypothesis

b) $t = 22.25$; $\chi^2_{0.025} = 8.91$; $\chi^2_{0.975} = 32.9$; we do not reject the hypothesis

E x e r c i s e 5.28

Using sample data of size n = 10 and a variance of $s^2 = 2.0$ at a 0.01 level of significance, test the hypothesis that the variance of the parent population is $\sigma^2 = 0.2$.

R e s u l t: t = 100; $\chi^2_{0.005} = 1.73$; $\chi^2_{0.995} = 23.6$; we reject the hypothesis

E x e r c i s e 5.29

Using a random sample from a two-dimensional normal distribution sample data have been obtained of size n = 44 and the correlation coefficient has been found to be r = 0.7417. At a 1% level of significance test the hypothesis that the random variables for the parent population are independent.

R e s u l t: H : $\rho = 0$; $t \cong 6.110$; $u_{0.995} = 2.576$; we reject the hypothesis

E x e r c i s e 5.30

With two-dimensional sample data of size n = 20 a sample correlation coefficient of r = 0.3998 has been calculated. At a 5% level of significance test the hypothesis that the random variables for the parent population are independent.

R e s u l t: H : $\rho = 0$; $t \cong 1.747$; $u_{0.975} = 1.960$; we do not reject the hypothesis

E x e r c i s e 5.31

To compare the accuracy of two methods of measurement a total of 8 measurements have been carried out and the differences in the pairs of the corresponding measurements have been calculated. In this way the average deviation $\bar{d} = 0.244$ and standard deviation s(d) = 0.192 have been found. At a 0.05 level of significance test the hypothesis that both methods may be taken for equally accurate.

R e s u l t: $t \cong 3.362$; $t_{0.975} = 2.365$; we reject the hypothesis

E x e r c i s e 5.32

Using two analytical scales a total of 10 samples have been weighed with the following results: $(x_i; y_i) = (25; 28), (30; 31), (28; 26), (50; 52), (20; 24), (40; 36), (32; 33), (36; 35), (42; 45), (38; 40)$ (mg). With a 0.01 level of significance find out

whether the different results are statistically insignificant provided that they are normally distributed.

R e s u l t: $t \cong -1.13$; $t_{0.995} = 3.250$; we do not reject the hypothesis, the different results are statistically insignificant

E x e r c i s e 5.33

The following are the results of measurements made before and after a scale of a packing machine has been calibrated: $n_1 = 12$, $\bar{x} = 31.2$ g, $s^2(x) = 0.770$ g^2 and $n_2 = 18$, $\bar{y} = 29.2$ g, $s^2(y) = 0.378$ g^2. Suppose the variances are equal and the distribution is normal. At a 0.05 level of significance test the hypothesis that the expected value has not been changed by the calibration.

R e s u l t: $t \cong 7.1$; $t_{0.975} = 2.048$; we reject the hypothesis

E x e r c i s e 5.34

The average grading of a total of twenty study groups of students in a particular year are shown in the table below:

x_j^*	1.70	1.86	2.01	2.23	2.27	2.411
f_j	2	3	5	7	2	1

The overall average grading in the previous year for 20 study groups was $\bar{y} = 2.201$ and the variance was $s^2(y) = 0.012$. Test the hypothesis that the average gradings of the two years do not differ if we assume that the distribution is normal and variances are equal.

R e s u l t: $\bar{x} = 2.0795$; $s^2(x) = 0.0399$; $t \cong -2.325$;

$t_{0.975} = 2.023$ - we reject the hypothesis

$t_{0.995} = 2.712$ - we do not reject the hypothesis

E x e r c i s e 5.35

Two types of rope have been tested for tensile strength. Two samples of an equal size of n = 18 have been taken and the following values have been calculated:

$\bar{x}$ = 3389.3 N, $s^2(x)$ = 1144.4 N^2, $\bar{y}$ = 3339.2 N, $s^2(y)$ = 3453.5 N^2. Assuming the variances to be different test the hypothesis that the expected tensile strengths of the ropes are the same. Use a 0.05 level of significance.

R e s u l t: $t \cong 3.046$; $\bar{t}_{0.975} = 2.110$; we reject the hypothesis

E x e r c i s e 5.36

Two sample data of sizes $n_1 = 20$ and $n_2 = 10$ with the following characteristics $\bar{x}$ = 10.24; $\bar{y}$ = 11.09; $s^2(x)$ = 4.231 and $s^2(y)$ = 18.457 have been calculated using two normally distributed random samples from random variables X and Y with different variances. At a 1% level of significance test the hypothesis that the expected values of X and Y are the same.

R e s u l t: $t \cong -0.5637$; $\bar{t}_{0.975} = 3.212$; we do not reject the hypothesis

E x e r c i s e 5.37

Two different methods have been used to determine the fat content of milk. Using the first method for a sample of 12 analyses a variance of $s^2(x)$ = 0.0224 has been calculated and with the second method a sample of 8 analyses has been used to produce a variance of $s^2(y)$ = 0.0263. At a 0.01 level of significance test the hypothesis that both methods are equally accurate in terms of variance.

R e s u l t: $t \cong 1.23$; $F_{0.95} = 4.907$; we do not reject the hypothesis

E x e r c i s e 5.38

Test the hypothesis that the variances in Exercise 5.19 are equal. Use a 0.05 level of significance.

R e s u l t: $t = 2.1$; $F_{0.95} = 2.41$; we do not reject the hypothesis

E x e r c i s e 5.39

The Board of Directors of a large company consider selling shares of the company's stock to their own employees. The estimate that about 20% of the employees will buy the shares. A total of 400 employees chosen at random have been asked whether

they will buy the shares. The answer has been yes in 66 cases. Using a 0.05 level of significance test the hypothesis that the directors' estimate is realistic.

R e s u l t: t = -1.75; $u_{0.975}$ = 1.960; we do not reject the hypothesis, the estimate is realistic

E x e r c i s e 5.40

A sample of size n = 200 has been taken from products manufactured using a new technology. Out of those 200 products 31 have been found to be defective. Ascertain that the new technology has changed the wastage rate of the products as compared to a previous rate of 10% determined by long experience. Use a 1% level of significance.

R e s u l t: $t \cong 2.593$; $u_{0.995}$ = 2.576; we do not reject the hypothesis (the new technology has changed the wastage rate)

E x e r c i s e 5.40

In two plants the same type of product is manufactured. The wastage rates in the two plants should be the same as both use the same technology. In the first plant, 10 products out of a total of 200 chosen at random and checked are defective while in the second plant it is 23 defective products out of a total of 250. Using a 0.01 level of significance test whether there is a statistically significant difference in quality between the two plants.

R e s u l t: $t \cong -1.699$; $u_{0.995}$ = 2.576; we do not reject the hypothesis, (no statistically significant difference exists between the two plants)

E x e r c i s e 5.41

A survey in a certain region has ascertained that out of a sample of 58 farmers 23 were ill while in a sample of 43 workers 28 were ill. At a 5% significance level test the hypothesis that the sickness rate in workers is higher than in farmers.

R e s u l t: $t \cong -2.534$; $u_{0.975}$ = 1.960; we do not reject the hypothesis (the sickness rate is higher in workers than it is in farmers)

E x e r c i s e 5.42

When investigating the dependence of quantity y on quantity x the following data has been measured:

x_i	3.4	4.3	5.4	6.7	8.7	10.6
y_i	4.5	5.8	6.8	8.1	10.5	12.7

Determine the regression function $y = \beta_1 + \beta_2 x$, calculate the point estimator for y(5.4) and find 0.95 interval estimates for β_1, β_2, y(5.4).

R e s u l t: $y = 0.77 + 1.12x$; $y(5.4) \cong 6.82$; $\beta_1 \in \langle 0.31;1.24\rangle$; $\beta_2 \in \langle 1.05;1.19\rangle$;
$y(5.4) \in \langle 6.41;7.23\rangle$

E x e r c i s e 5.43

To determine the dependence of this year's demand y on last year's demand x for a certain type of goods the following data have been collected from 6 businessmen (pieces):

x_i	20	60	70	100	150	260
y_i	50	60	60	120	230	320

Calculate the minimum variance unbiased point estimate and find a 95% interval estimate for the coefficients of the regression line and for the value of this year's demand at 110 pieces of last year's demand. With a 5% level of significance test the hypothesis that $\beta_1 = 0$ and determine the correlation coefficient.

R e s u l t: $y = 0.687 + 1.266x$; $\beta_1 \in \langle -57.194; 58.568\rangle$; $\beta_2 \in \langle 0.836; 1.696\rangle$;
$y(110) \cong 140$; $y(110) \in \langle 106.55; 173.45\rangle$, resp. $\langle 51.50; 228.50\rangle$;
we do not reject the hypothesis; r = 0.97198

E x e r c i s e 5.44

The values y^* (in thousands of items) of the demand for a certain type of goods at prices x^* (in thousands of CZK) are shown in the following table:

x_i^*	100	110	140	160	200
y_i^*	120	89	56	41	22

Fit the data using a power regression function $y^* = \gamma x^{*\delta}$ and find point and interval estimates (using a 0.95 confidence) for the regression coefficients and for the demand at price 120 CZK.

(Hint: take a logarithm of the power function.)

R e s u l t: $\ln y^* = 15.64395 - 2.36035 \ln x^*$; $y^* = 6.224.10^6 x^{*-2.36}$;

$\ln \gamma = \beta_1 \in$ <13.95342; 17.33448>; $\delta = \beta_2 \in$ <-2.37817; -2.34253>;

$y^*(120) \cong 77$; $y^*(120) \in$ <69.8; 84.9> or <62.0; 95.6>

E x e r c i s e 5.45

The values y^* (in thousands of CZK) of net sales at a company over the first six years x^* of operation are shown below:

x_i^*	1	2	3	4	5	6
y_i^*	112	149	238	354	580	867

Approximate the data using an exponential regression function $y^* = \gamma \exp(\delta x^*)$ and calculate point and interval estimates (using a 95 % confidence) of the regression coefficients and make a forecast of the company's net sales in the seventh year of operation.

(Hint: take a logarithm of the exponential function.)

R e s u l t: $\ln y^* = 4.22798 + 0.4202x^*$; $y^* = 68.579\exp(0.4202x^*)$;

$\gamma = \exp(\beta_1) \in$ <59.41163; 79.15991>; $\delta = \beta_2 \in$ <0.38336; 0.45704>;

$y^*(7) \cong 1299.04$; $y^*(7) \in$ <1052.39; 1603.49>

Questions

1. Define a random sample and sample data.
2. What is meant by a sample characteristic and what are its properties?

3. Define the notion of a parameter and its types.
4. Define a point estimator and show point estimators for the basic number characteristics.
5. Describe an interval estimator and a confidence interval for parameters.
6. Hoe does a change in the confidence level and the sample size affect the length of a confidence interval?
7. Define a statistical hypothesis and describe its types.
8. What is a testing criterion and a critical range?
9. Describe type 1 and type 2 errors made at testing statistical hypotheses.
10. What is understood by regression analysis and a linear regression model?
11. What estimations and tests of statistical hypotheses are used in regression analysis?
12. How is the suitability of a regression function appraised?

BIBLIOGRAPHY

1. BOWERMAN, B. L., and O´CONNELL, R. T.: Applied Statistics; Improving Business Processes. IRWIN, Chicago, 1997.
2. WONNACOT, T.H., and WONNACOT, R. J.: Introductory Statistic for Business and Economics. John Wiley & Sons, Inc., New York, 1993.
3. SPRINTHALL, R. C.: Basic Statistical Analysis (5^{th} ed.). Allyn and Bacon, Boston, 1997.
4. ACZEL, A. D.: Complete Business Statistics. IRWIN, Chicago, 1989.
5. TRIOLA, M. F.: Elementary Statistics. B/C Publishing Comp., Redwood City, 1989.
6. AMEMIYA, T.: Introduction to Statistics & Econometrics. Harvard University Press, 1994.
7. BERENSON, M.: Business Statistics; A First Course. New York, Prentice Hall 1997.
8. WEIERS, R. M.: Introduction to Business Statistics. 3. e. Brooks/Cole Publishing 1997.
9. SEGER, J., and HINDLS, R.: Statistické metody v tržním hospodářství. Victoria Publishing, Praha, 1995.
10. ŠIKULOVÁ, M., and KARPÍŠEK, Z.: Matematika IV (Pravděpodobnost a matematická statistika). ES VUT, Brno, 1996.
11. LIKEŠ, J., CYHELSKÝ, L., and HINDLS, R.: Úvod do statistiky a pravděpodobnosti (Statistika A). VŠE, Praha, 1995.
12. JAROŠOVÁ, E.: Statistika B. Řešené příklady. VŠE, Praha, 1994.
13. MELOUN, M., and MILITKÝ, J.: Statistické zpracování experimentálních dat. PLUS, Praha, 1994.
14. SEBEROVÁ, H.: Statistika I, II. VVŠ PV, Vyškov, 1995.
15. CIPRA, T.: Analýza časových řad s aplikacemi v ekonomii. SNTL/Alfa, Praha, 1986.

STATISTICAL TABLES

T1 Distribution function $\Phi(u)$ of the standard normal distribution N(0;1)

u	0	1	2	3	4	5	6	7	8	9
0.0	0.50000	50399	50798	51197	51596	51994	52392	52791	53188	53586
0.1	53983	54380	54776	55172	55567	55962	56356	56750	57143	57535
0.2	57926	58317	58707	59096	59484	59871	60257	60642	61026	61409
0.3	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0.4	65542	65910	66276	66640	67003	67365	67724	68082	68439	68793
0.5	69146	69498	69847	70195	70540	70884	71226	71566	71904	72241
0.6	72575	72907	73237	73565	73892	74216	74537	74857	75175	75490
0.7	75804	76115	76424	76731	77035	77337	77637	77935	78231	78524
0.8	78815	79103	79389	79673	79955	80234	80511	80785	81057	81327
0.9	81594	81859	82121	82382	82639	82894	83147	83398	83646	83891
1.0	84135	84375	84614	84850	85083	85314	85543	85769	85993	86214
1.1	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1.2	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1.3	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1.4	91924	92073	92220	92364	92507	92647	92786	92922	93056	93189
1.5	93319	93448	93574	93699	93822	93943	94062	94179	94295	94408
1.6	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1.7	95543	95637	95728	95819	95907	95994	96080	96164	96246	96327
1.8	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1.9	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2.0	97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2.1	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2.2	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2.3	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2.4	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2.5	99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2.6	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2.7	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2.8	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2.9	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3.0	99865	99869	99874	99878	99882	99886	99889	99893	99896	99900
3.1	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3.2	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3.3	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3.4	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3.5	99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3.6	99984	99985	99985	99986	99986	99987	99987	99988	99988	99989
3.7	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3.8	99993	99993	99993	99994	99994	99994	99994	99995	99995	99995
3.9	99995	99995	99996	99996	99996	99996	99996	99996	99997	99997

4.00	99997	4.10	99998	4.20	99999	4.30	99999	4.40	99999	4.50	99999

T2 Quantiles t_P of the Student distribution S(k)

k \ P	0.95	0.975	0.99	0.995	0.999	0.9995
1	6.314	12.706	31.821	63.656	318.289	636.578
2	2.920	4.303	6.965	9.925	22.328	31.600
3	2.353	3.182	4.541	5.841	10.214	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.894	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.768
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.689
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.660
30	1.697	2.042	2.457	2.750	3.385	3.646
35	1.690	2.030	2.438	2.724	3.340	3.591
40	1.684	2.021	2.423	2.704	3.307	3.551
45	1.679	2.014	2.412	2.690	3.281	3.520
50	1.676	2.009	2.403	2.678	3.261	3.496
60	1.671	2.000	2.390	2.660	3.232	3.460
70	1.667	1.994	2.381	2.648	3.211	3.435
80	1.664	1.990	2.374	2.639	3.195	3.416
90	1.662	1.987	2.368	2.632	3.183	3.402
100	1.660	1.984	2.364	2.626	3.174	3.390
120	1.658	1.980	2.358	2.617	3.160	3.373
140	1.656	1.977	2.353	2.611	3.149	3.361
160	1.654	1.975	2.350	2.607	3.142	3.352
180	1.653	1.973	2.347	2.603	3.136	3.345
200	1.653	1.972	2.345	2.601	3.131	3.340
300	1.650	1.968	2.339	2.592	3.118	3.323
400	1.649	1.966	2.336	2.588	3.111	3.315
500	1.648	1.965	2.334	2.586	3.107	3.310
1000	1.646	1.962	2.330	2.581	3.098	3.300
∞	1.645	1.960	2.326	2.576	3.090	3.290

T3 Quantiles χ^2_P of the Pearson distribution χ^2 (k)

k \ P	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.878	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.994
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.335
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	44.985	48.232	52.191	55.002
32	15.134	16.362	18.291	20.072	46.194	49.480	53.486	56.328
33	15.815	17.073	19.047	20.867	47.400	50.725	54.775	57.648
34	16.501	17.789	19.806	21.664	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	53.384	56.895	61.162	64.181
39	19.996	21.426	23.654	25.695	54.572	58.120	62.428	65.475
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	58.124	61.777	66.206	69.336
43	22.860	24.398	26.785	28.965	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	60.481	64.201	68.710	71.892
45	24.311	25.901	28.366	30.612	61.656	65.410	69.957	73.166

T3 Quantiles χ^2_P of the Pearson distribution χ^2(k) (continued)

k \ P	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
46	25.041	26.657	29.160	31.439	62.830	66.616	71.201	74.437
47	25.775	27.416	29.956	32.268	64.001	67.821	72.443	75.704
48	26.511	28.177	30.754	33.098	65.171	69.023	73.683	76.969
49	27.249	28.941	31.555	33.930	66.339	70.222	74.919	78.231
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
51	28.735	30.475	33.162	35.600	68.669	72.616	77.386	80.746
52	29.481	31.246	33.968	36.437	69.832	73.810	78.616	82.001
53	30.230	32.019	34.776	37.276	70.993	75.002	79.843	83.253
54	30.981	32.793	35.586	38.116	72.153	76.192	81.069	84.502
55	31.735	33.571	36.398	38.958	73.311	77.380	82.292	85.749
56	32.491	34.350	37.212	39.801	74.468	78.567	83.514	86.994
57	33.248	35.131	38.027	40.646	75.624	79.752	84.733	88.237
58	34.008	35.914	38.844	41.492	76.778	80.936	85.950	89.477
59	34.770	36.698	39.662	42.339	77.930	82.117	87.166	90.715
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
61	36.300	38.273	41.303	44.038	80.232	84.476	89.591	93.186
62	37.068	39.063	42.126	44.889	81.381	85.654	90.802	94.419
63	37.838	39.855	42.950	45.741	82.529	86.830	92.010	95.649
64	38.610	40.649	43.776	46.595	83.675	88.004	93.217	96.878
65	39.383	41.444	44.603	47.450	84.821	89.177	94.422	98.105
66	40.158	42.240	45.431	48.305	85.965	90.349	95.626	99.330
67	40.935	43.038	46.261	49.162	87.108	91.519	96.828	100.554
68	41.714	43.838	47.092	50.020	88.250	92.688	98.028	101.776
69	42.493	44.639	47.924	50.879	89.391	93.856	99.227	102.996
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
71	44.058	46.246	49.592	52.600	91.670	96.189	101.621	105.432
72	44.843	47.051	50.428	53.462	92.808	97.353	102.816	106.647
73	45.629	47.858	51.265	54.325	93.945	98.516	104.010	107.862
74	46.417	48.666	52.103	55.189	95.081	99.678	105.202	109.074
75	47.206	49.475	52.942	56.054	96.217	100.839	106.393	110.285
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	107.522	112.393	118.236	122.324
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.170
110	75.550	78.458	82.867	86.792	135.480	140.916	147.414	151.948
120	83.852	86.923	91.573	95.705	146.567	152.211	158.950	163.648
130	92.223	95.451	100.331	104.662	157.610	163.453	170.423	175.278
150	109.142	112.668	117.985	122.692	179.581	185.800	193.207	198.360
200	152.241	156.432	162.728	168.279	233.994	241.058	249.445	255.264
500	422.303	429.387	439.936	449.147	553.127	563.851	576.493	585.206
1000	888.563	898.912	914.257	927.594	1074.68	1089.53	1106.97	1118.95

T4 Quantiles F_P of the Fisher – Snedecor distribution $F(k_1,k_2)$ for P = 0.95

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161.446	199.499	215.707	224.583	230.160	233.988	238.884	243.905	249.052	254.313
2	18.513	19.000	19.164	19.247	19.296	19.329	19.371	19.412	19.454	19.496
3	10.128	9.552	9.277	9.117	9.013	8.941	8.845	8.745	8.638	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.041	5.912	5.774	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.818	4.678	4.527	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.147	4.000	3.841	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.726	3.575	3.410	3.230
8	5.318	4.459	4.066	3.838	3.688	3.581	3.438	3.284	3.115	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.230	3.073	2.900	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.072	2.913	2.737	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	2.948	2.788	2.609	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.849	2.687	2.505	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.767	2.604	2.420	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.699	2.534	2.349	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.641	2.475	2.288	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.591	2.425	2.235	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.548	2.381	2.190	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.510	2.342	2.150	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.477	2.308	2.114	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.447	2.278	2.082	1.843
21	4.325	3.467	3.072	2.840	2.685	2.573	2.420	2.250	2.054	1.812
22	4.301	3.443	3.049	2.817	2.661	2.549	2.397	2.226	2.028	1.783
23	4.279	3.422	3.028	2.796	2.640	2.528	2.375	2.204	2.005	1.757
24	4.260	3.403	3.009	2.776	2.621	2.508	2.355	2.183	1.984	1.733
25	4.242	3.385	2.991	2.759	2.603	2.490	2.337	2.165	1.964	1.711
26	4.225	3.369	2.975	2.743	2.587	2.474	2.321	2.148	1.946	1.691
27	4.210	3.354	2.960	2.728	2.572	2.459	2.305	2.132	1.930	1.672
28	4.196	3.340	2.947	2.714	2.558	2.445	2.291	2.118	1.915	1.654
29	4.183	3.328	2.934	2.701	2.545	2.432	2.278	2.104	1.901	1.638
30	4.171	3.316	2.922	2.690	2.534	2.421	2.266	2.092	1.887	1.622
35	4.121	3.267	2.874	2.641	2.485	2.372	2.217	2.041	1.833	1.558
40	4.085	3.232	2.839	2.606	2.449	2.336	2.180	2.003	1.793	1.509
45	4.057	3.204	2.812	2.579	2.422	2.308	2.152	1.974	1.762	1.470
50	4.034	3.183	2.790	2.557	2.400	2.286	2.130	1.952	1.737	1.438
55	4.016	3.165	2.773	2.540	2.383	2.269	2.112	1.933	1.717	1.412
60	4.001	3.150	2.758	2.525	2.368	2.254	2.097	1.917	1.700	1.389
70	3.978	3.128	2.736	2.503	2.346	2.231	2.074	1.893	1.674	1.353
80	3.960	3.111	2.719	2.486	2.329	2.214	2.056	1.875	1.654	1.325
90	3.947	3.098	2.706	2.473	2.316	2.201	2.043	1.861	1.639	1.302
100	3.936	3.087	2.696	2.463	2.305	2.191	2.032	1.850	1.627	1.283
120	3.920	3.072	2.680	2.447	2.290	2.175	2.016	1.834	1.608	1.254
150	3.904	3.056	2.665	2.432	2.274	2.160	2.001	1.817	1.590	1.223
250	3.879	3.032	2.641	2.408	2.250	2.135	1.976	1.791	1.561	1.166
500	3.860	3.014	2.623	2.390	2.232	2.117	1.957	1.772	1.539	1.113
∞	3.841	2.996	2.605	2.372	2.214	2.099	1.938	1.752	1.517	1.000

T4 Quantiles F_P of the Fisher – Snedecor distribution $F(k_1,k_2)$ for P = 0.99

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5980.95	6106.68	6234.27	6365.59
2	98.502	99.000	99.164	99.251	99.302	99.331	99.375	99.419	99.455	99.499
3	34.116	30.816	29.457	28.710	28.237	27.911	27.489	27.052	26.597	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.799	14.374	13.929	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.289	9.888	9.466	9.020
6	13.745	10.925	9.780	9.148	8.746	8.466	8.102	7.718	7.313	6.880
7	12.246	9.547	8.451	7.847	7.460	7.191	6.840	6.469	6.074	5.650
8	11.259	8.649	7.591	7.006	6.632	6.371	6.029	5.667	5.279	4.859
9	10.562	8.022	6.992	6.422	6.057	5.802	5.467	5.111	4.729	4.311
10	10.044	7.559	6.552	5.994	5.636	5.386	5.057	4.706	4.327	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.744	4.397	4.021	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.499	4.155	3.780	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.302	3.960	3.587	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.140	3.800	3.427	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.004	3.666	3.294	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	3.890	3.553	3.181	2.753
17	8.400	6.112	5.185	4.669	4.336	4.101	3.791	3.455	3.083	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.705	3.371	2.999	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.631	3.297	2.925	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.564	3.231	2.859	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.506	3.173	2.801	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.453	3.121	2.749	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.406	3.074	2.702	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.363	3.032	2.659	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.324	2.993	2.620	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.288	2.958	2.585	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.256	2.926	2.552	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.226	2.896	2.522	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.198	2.868	2.495	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.173	2.843	2.469	2.006
35	7.419	5.268	4.396	3.908	3.592	3.368	3.069	2.740	2.364	1.891
40	7.314	5.178	4.313	3.828	3.514	3.291	2.993	2.665	2.288	1.805
45	7.234	5.110	4.249	3.767	3.454	3.232	2.935	2.608	2.230	1.737
50	7.171	5.057	4.199	3.720	3.408	3.186	2.890	2.563	2.183	1.683
55	7.119	5.013	4.159	3.681	3.370	3.149	2.853	2.526	2.146	1.638
60	7.077	4.977	4.126	3.649	3.339	3.119	2.823	2.496	2.115	1.601
70	7.011	4.922	4.074	3.600	3.291	3.071	2.777	2.450	2.067	1.540
80	6.963	4.881	4.036	3.563	3.255	3.036	2.742	2.415	2.032	1.494
90	6.925	4.849	4.007	3.535	3.228	3.009	2.715	2.389	2.004	1.457
100	6.895	4.824	3.984	3.513	3.206	2.988	2.694	2.368	1.983	1.427
120	6.851	4.787	3.949	3.480	3.174	2.956	2.663	2.336	1.950	1.381
150	6.807	4.749	3.915	3.447	3.142	2.924	2.632	2.305	1.918	1.331
250	6.737	4.691	3.861	3.395	3.091	2.875	2.583	2.256	1.867	1.244
500	6.686	4.648	3.821	3.357	3.054	2.838	2.547	2.220	1.829	1.164
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.511	2.185	1.791	1.000