

3 Random variable and random sample

Probability theory is an important part of statistics and statistical methods. The basic concepts include random experiment, random event, and probability of a random event. A *random experiment* is any activity that is repeated several times under the same or approximately the same conditions. A *random event* is any possible outcome of a random experiment, which can be characterized qualitatively (verbally) or quantitatively (numerically). In many cases it is more convenient to characterize the outcome quantitatively (numerically), for which we use a random variable (quantity). The *probability of a random event* A is a number that quantitatively characterizes the possibility of the random event A occurring. We denote it $P(A)$.

3.1 Random variable and its numerical characteristics

A *random variable (quantity)* X is a mapping that assigns a real number to each element of the set of random events. We will denote random variables by capital letters $X, Y, Z, \dots$ and their specific values by lowercase letters $x, y, z, \dots$.

According to the values a random variable takes, we divide it into:

- a *discrete random variable*, which takes a finite or countable number of possible values (e.g. number of customers, age of a respondent),
- a *continuous random variable*, which takes any value from some (bounded or unbounded) interval (e.g. lifetime of a product, width of a product).

Knowing a random variable means knowing its *probability distribution*, i.e. we need to determine what values it takes and with what probabilities. This means that for every real number x we can determine the probability that the value of the random variable X equals the number x (denoted $P(X = x)$). The probability distribution can be given by a table, a function, or a graph. The *distribution function* assigns to every real number the probability that the random variable X takes a value less than or equal to the number x , i.e.

$$F(x) = P(X \leq x). \quad (35)$$

It is a universal form of describing a probability distribution, since it is used for both discrete and continuous random variables. The distribution function can also be used to describe the opposite probability, i.e. the probability that the random variable X takes a value greater than the number x , i.e.

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x).$$

Properties of the distribution function:

1. The distribution function takes only values from the interval $\langle 0; 1 \rangle$, i.e. $0 \leq F(x) \leq 1$ for every $x \in \mathbb{R}$.
2. The distribution function is a non-decreasing function.
3. The distribution function is right-continuous on $\mathbb{R}$.
4. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
5. If $a < b$, then $P(X \in (a, b)) = P(a < X \leq b) = F(b) - F(a)$.

The most important characteristic of location of the random variable X is the *expected value* $E(X)$. Its definition depends on whether we are dealing with a discrete or a continuous random variable, so we will define it separately for these random variables in the following chapters.

The basic characteristic of variability is the *variance (dispersion)* $D(X)$. The variance of the random variable X is the expected value of the squared deviation of the random variable X from its expected value $E(X)$, i.e.

$$D(X) = E([X - E(X)]^2) = E(X^2) - [E(X)]^2. \quad (36)$$

The square root of the variance of the random variable X is called the *standard deviation* $\sigma(X)$, i.e.

$$\sigma(X) = \sqrt{D(X)}. \quad (37)$$

3.1.1 Discrete random variable

A *discrete random variable* is a random variable X that takes values $x_1, x_2, \dots, x_n$ with probabilities $p_1, p_2, \dots, p_n$, i.e.

$$p_i = P(X = x_i) \text{ for } i = 1, 2, \dots, n, \quad (38)$$

where the so-called *normalization condition* is satisfied

$$\sum_{i=1}^n p_i = 1, \quad (39)$$

where $p_i \geq 0$ for $i = 1, 2, \dots, n$.

The probability distribution of a discrete random variable X can be given by a probability table or by the distribution function of the discrete random variable.

A *probability table* (*probability distribution table*) of a discrete random variable (Table 4) is a table that contains the possible values of the random variable x_i and their corresponding probabilities $p_i = P(X = x_i)$ for $i = 1, 2, \dots, n$.

Table 4: Probability table (probability distribution table)

x_i	x_1	x_2	$\cdots$	x_n
p_i	p_1	p_2	$\cdots$	p_n

The *distribution function* of a discrete random variable has the form

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p_i. \quad (40)$$

The graph of the distribution function of a discrete random variable has a so-called step shape, where the size of the jump at the point x_i equals the probability p_i (Fig. 11).

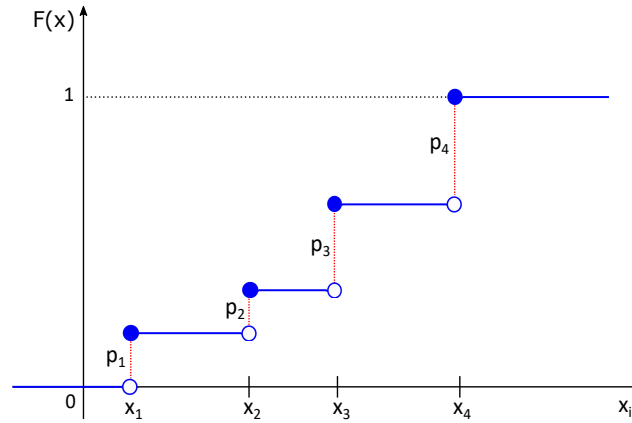


Figure 11: Graph of the distribution function of a discrete random variable

The *expected value* $E(X)$ of a discrete random variable X is defined by the relation

$$E(X) = \sum_{i=1}^n x_i p_i. \quad (41)$$

By substituting the expected value of a discrete random variable into the general formula (36), we obtain for the *variance* $D(X)$ of a discrete random variable X the relation

$$D(X) = \sum_{i=1}^n (x_i - E(X))^2 p_i = \sum_{i=1}^n x_i^2 p_i - \left(\sum_{i=1}^n x_i p_i \right)^2. \quad (42)$$

3.1.2 Continuous random variable

A *continuous random variable* is a random variable X for which there exists a real, non-negative, integrable function $f(x)$ such that for all $x \in \mathbb{R}$ the distribution function can be expressed in the form

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt, \text{ for all } x \in \mathbb{R}. \quad (43)$$

The function $f(x)$ is called the *probability density function of the continuous random variable* and has the following properties:

1. If the value $f(x)$ exists, then $f(x) \geq 0$.
2. If $F'(x)$ exists, then $F'(x) = f(x)$ for all $x \in \mathbb{R}$.
3. *Normalization condition:*

$$\int_{-\infty}^{\infty} f(x) dx = 1, \text{ for all } x \in \mathbb{R}. \quad (44)$$

4. If $a < b$, then

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a \leq X < b) = \\ &= P(a < X < b) = \int_a^b f(x) dx = F(b) - F(a). \end{aligned}$$

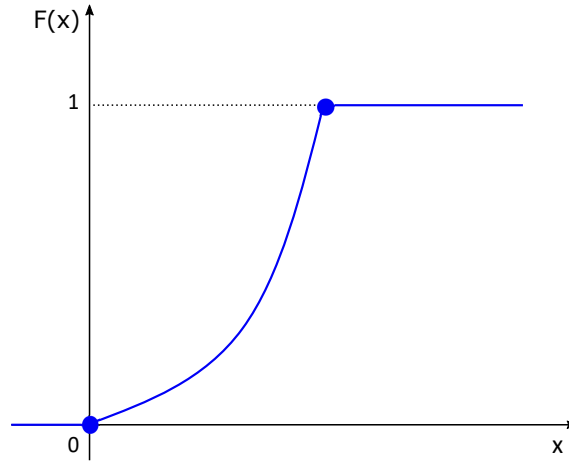


Figure 12: Graph of the distribution function of a continuous random variable

The probability distribution of a continuous random variable X can be given by the probability density $f(x)$ of the continuous random variable X or by the distribution function of the continuous random variable X (Fig. 12).

The relationship between the probability density and the distribution function of a continuous random variable X , which follows from relation (43), is illustrated in Fig. 13.

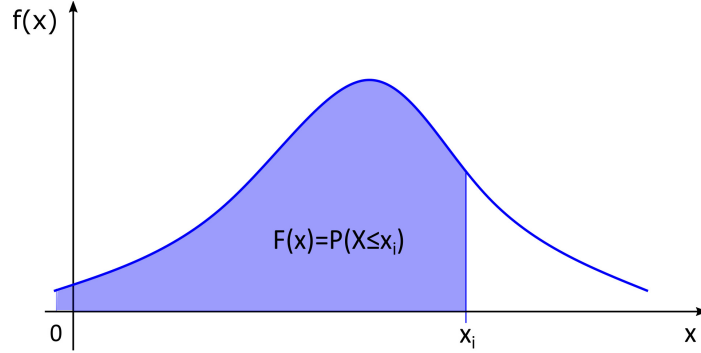


Figure 13: Relationship between the probability density and the distribution function

The *expected value* $E(X)$ of a continuous random variable X is defined by the relation

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx. \quad (45)$$

By substituting the expected value of a continuous random variable into the general formula (36), we obtain for the *variance* $D(X)$ of a continuous random variable X the relation

$$D(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - [E(X)]^2. \quad (46)$$

3.2 Selected distributions of random variables

In this chapter we mention the probability distributions of random variables that we will encounter in this textbook. For each distribution we give its basic properties and characteristics.

3.2.1 Binomial distribution

The binomial distribution belongs among the probability distributions of discrete random variables. We have n independent trials, each of which ends in only one of two possible outcomes: the event occurs, or the event does not occur. The probability that the event occurs is p , and the probability that the event does not occur is $1 - p$, where $0 < p < 1$ and p is the same in every trial.

A discrete random variable X has a *binomial probability distribution* determined by the parameters n and p if it takes the values $x = 0, 1, \dots, n$ with probability

$$p(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad p \in (0, 1), \quad n = 1, 2, \dots, \quad (47)$$

where p is the probability of the event occurring and n is the number of independent trials. We write $X \sim \text{Bi}(n, p)$.

For the numerical characteristics we have:

- Expected value:

$$E(X) = np, \quad (48)$$

- Variance:

$$D(X) = np(1 - p). \quad (49)$$

3.2.2 Poisson distribution

The Poisson distribution belongs among the probability distributions of discrete random variables. It is used to model processes in which we observe the number of occurrences of a certain event in a given time interval or spatial region (e.g. number of hours of work without a breakdown, number of traffic accidents per month, number of customers served per week).

A discrete random variable X has a *Poisson probability distribution* determined by the parameter λ , $\lambda > 0$, if it takes the values $x = 0, 1, \dots, n$ with probability

$$p(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}. \quad (50)$$

We write $X \sim \text{Po}(\lambda)$.

For the numerical characteristics we have:

- Expected value:

$$E(X) = \lambda, \quad (51)$$

- Variance:

$$D(X) = \lambda. \quad (52)$$

3.2.3 Uniform distribution

The uniform probability distribution belongs among the probability distributions of continuous random variables. The continuous random variable X takes values only on an interval of a specific length, with all values of the random variable in the given interval being equally probable (e.g. waiting time for a bus, arrival of trains).

A continuous random variable X has a *uniform probability distribution* determined by the parameters a, b , $a < b$, if its probability density $f(x)$ has the form

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in \langle a, b \rangle, \\ 0 & \text{for } x \notin \langle a, b \rangle. \end{cases} \quad (53)$$

We write $X \sim R(a, b)$.

For the numerical characteristics and distribution function we have:

- Expected value:

$$E(X) = \frac{a+b}{2}, \quad (54)$$

- Variance:

$$D(X) = \frac{(b-a)^2}{12}, \quad (55)$$

- Distribution function:

$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } x \in \langle a, b \rangle, \\ 1 & \text{for } x > b. \end{cases} \quad (56)$$

The graphs of the probability density and distribution function of the uniform distribution are shown in Fig. 14.

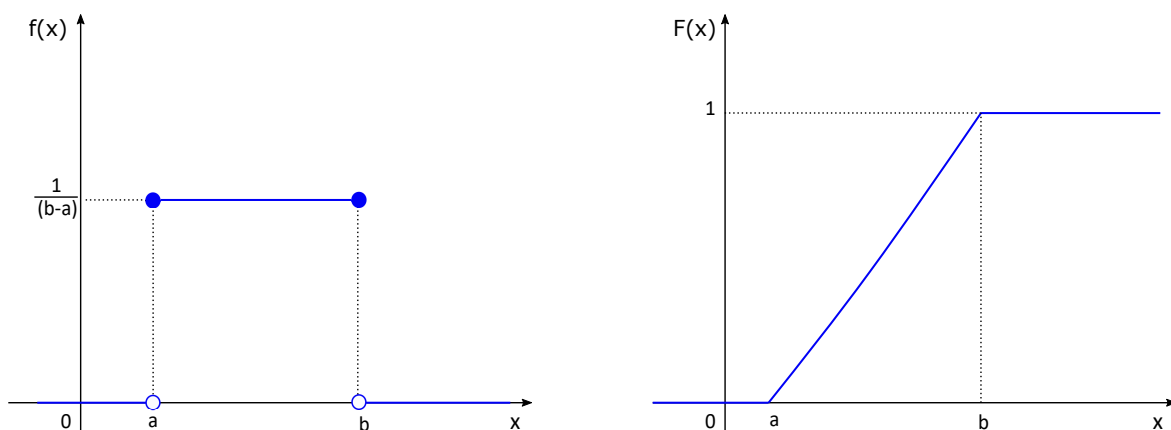


Figure 14: Probability density and distribution function of the uniform distribution

3.2.4 Exponential distribution

The exponential distribution belongs among the probability distributions of continuous random variables and is typical for problems related to the lifetime of devices or products, where the mean lifetime is given by the parameter λ .

A continuous random variable X has an *exponential probability distribution* determined by the parameter λ , $\lambda > 0$, if its probability density $f(x)$ has the form

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases} \quad (57)$$

We write $X \sim \text{Ex}(\lambda)$.

For the numerical characteristics and distribution function we have:

- Expected value:

$$E(X) = \lambda, \quad (58)$$

- Variance:

$$D(X) = \lambda^2, \quad (59)$$

- Distribution function:

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{\lambda}} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases} \quad (60)$$

The graphs of the probability density and distribution function of the exponential distribution are shown in Fig. 15.

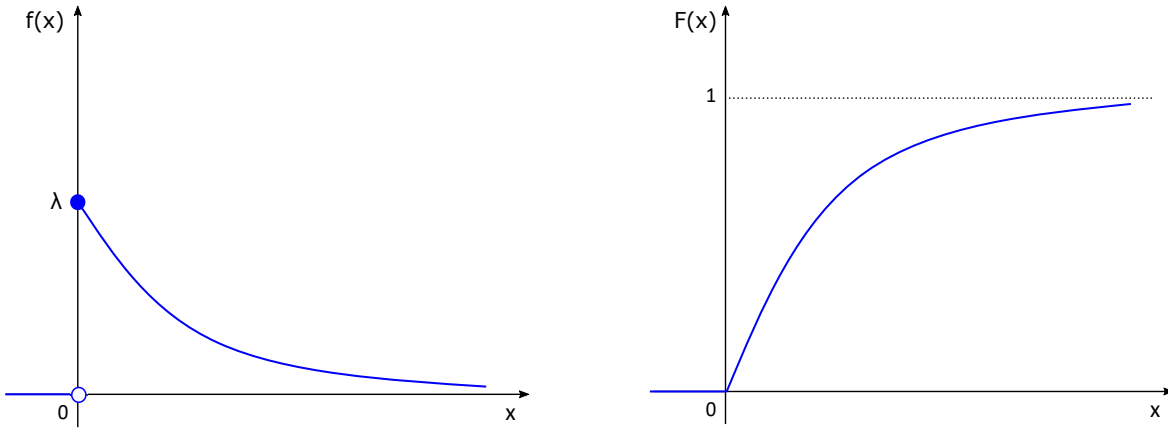


Figure 15: Probability density and distribution function of the exponential distribution

Remark: In some literature, the exponential distribution is given with a parameter α , where $\alpha = \frac{1}{\lambda}$.

3.2.5 Normal distribution

The normal distribution is used as a model in problems where the random variable X fluctuates around the expected value μ with variance σ^2 . In practice it is one of the most widely used probability distributions (e.g. measurement errors).

A continuous random variable X has a *normal probability distribution* determined by the parameters μ , σ^2 , $\sigma > 0$, if its probability density $f(x)$ has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for every } x \in \mathbb{R}. \quad (61)$$

We write $X \sim N(\mu, \sigma^2)$.

For the numerical characteristics and distribution function we have:

- Expected value:

$$E(X) = \mu, \quad (62)$$

- Variance:

$$D(X) = \sigma^2, \quad (63)$$

- Distribution function:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt. \quad (64)$$

The graph of the probability density $f(x)$ is called the *Gaussian curve* and has a so-called bell shape, symmetric about the expected value μ . Together with the graph of the distribution function of the normal distribution, it is shown in Fig. 16.

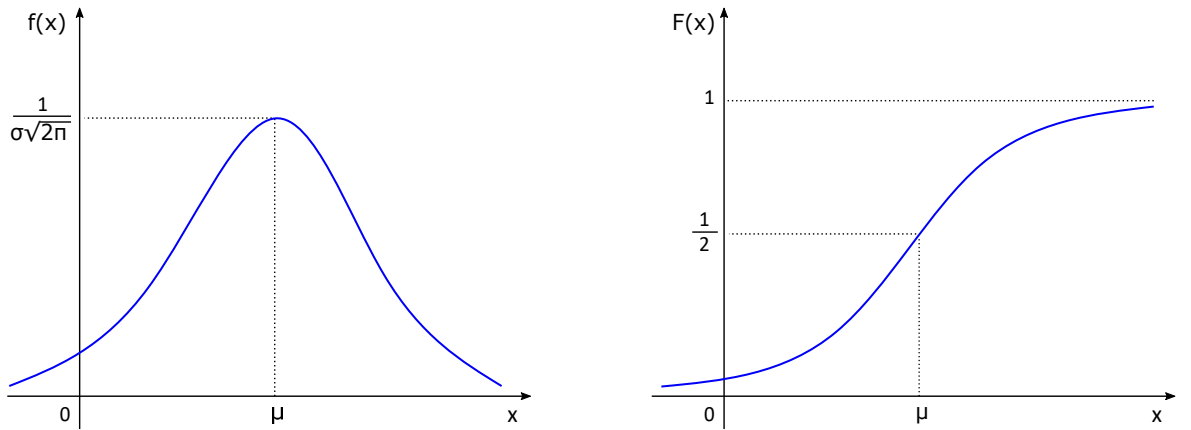


Figure 16: Probability density and distribution function of the normal distribution

Standard normal distribution

A continuous random variable Y has a *standard normal probability distribution* determined by the parameters $\mu = 0$, $\sigma^2 = 1$, if its probability density $\varphi(y)$ has the form

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \quad \text{for every } y \in \mathbb{R}. \quad (65)$$

We write $Y \sim N(0, 1)$.

For the numerical characteristics and distribution function we have:

- Expected value:

$$E(Y) = 0, \quad (66)$$

- Variance:

$$D(Y) = 1, \quad (67)$$

- Distribution function:

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{t^2}{2}} dt. \quad (68)$$

Since the function $\varphi(y)$ is even and the normalization condition (44) holds for it, for the distribution function $\Phi(y)$ we obtain the symmetry property, i. e.

$$\Phi(-y) = 1 - \Phi(y). \quad (69)$$

Values of the distribution function $\Phi(y)$ are tabulated only for $y \geq 0$ (Table 27). Based on condition (69) we can also determine the values $\Phi(y)$ for $y < 0$.

The graphs of the probability density and distribution function of the standard normal distribution are shown in Fig. 17.

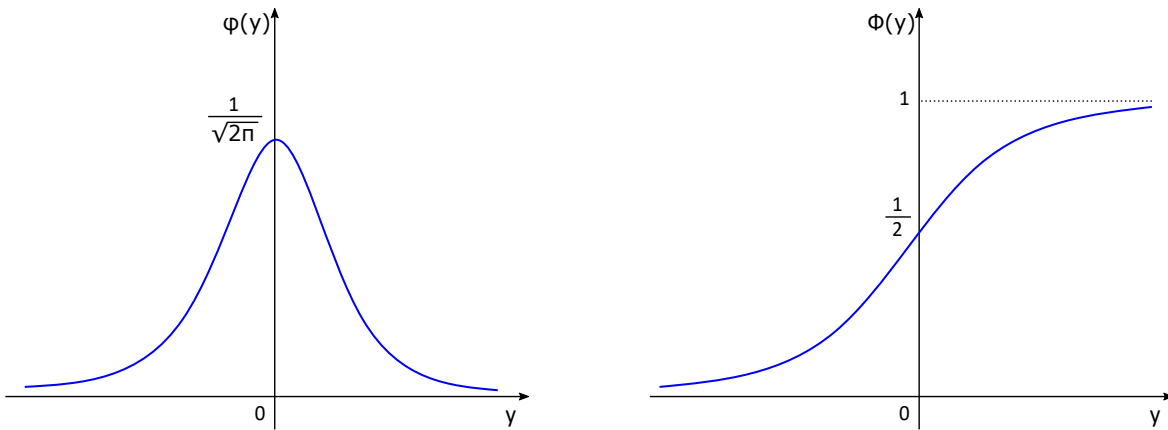


Figure 17: Probability density and distribution function of the standard normal distribution

Relationship between the standard normal distribution and the normal distribution:

1. If a random variable X has the normal distribution $N(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma}$ is the corresponding standardized random variable and has the standard normal distribution $N(0, 1)$.
2. If a random variable X has the normal distribution $N(\mu, \sigma^2)$ with distribution function $F(x)$, then for every $x \in \mathbb{R}$ we have

$$P(X \leq x) = F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right). \quad (70)$$

3. If a random variable X has the normal distribution $N(\mu, \sigma^2)$ with distribution function $F(x)$, then for all $a, b \in \mathbb{R}$, $a < b$ we have

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a \leq X < b) = \\ &= P(a < X < b) = F(b) - F(a) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right). \end{aligned}$$

3.2.6 Other distributions

In this chapter we give an overview of the probability distributions that we have not covered in the previous chapters but that we will use in estimation theory and in statistical hypothesis testing.

First we need to introduce the concept of the quantile function. Let $F(x)$ be the distribution function of a random variable X and $\alpha \in (0; 1)$. Then the *quantile function* $F^{-1}(\alpha)$ is defined as the smallest number q , $q \in \mathbb{R}$, such that $F(q) \geq \alpha$. The value $q = F^{-1}(\alpha)$ is called the α -*quantile* of the probability distribution given by the distribution function $F(x)$.

Student's t-distribution

- The shape of the density of the distribution depends on the parameter k , which determines the number of degrees of freedom.
- It is symmetric about zero, and as the number of degrees of freedom increases it increasingly resembles the normal distribution. For $k > 30$ it can be well approximated by the normal distribution $N(0, \frac{k}{k-2})$.
- The quantiles of Student's t -distribution $t_\alpha(k)$ are tabulated for various degrees of freedom (Table 29).

χ^2 -distribution

- The shape of the density of the distribution depends on the parameter k , which determines the number of degrees of freedom.
- It is not symmetric for small k , but for $k > 100$ it can be well approximated by the normal distribution $N(k, 2k)$.
- The quantiles of the χ^2 -distribution $\chi_\alpha^2(k)$ are tabulated for various degrees of freedom (Tables 30 – 32).

Fisher's F-distribution

- The shape of the density of the distribution depends on two parameters k_1 and k_2 .
- As the parameters increase, the probability density becomes more symmetric and increasingly resembles the normal distribution.
- It is not a symmetric distribution.
- The quantiles of Fisher's F -distribution $F_\alpha(k_1, k_2)$ are tabulated for various degrees of freedom (Tables 33–40) and it holds that

$$F_\alpha(k_1, k_2) = \frac{1}{F_{1-\alpha}(k_2, k_1)}.$$

Remark: We will denote the quantiles of the standard normal distribution $N(0, 1)$ by u_α , and they are tabulated for various values of α (Table 28).

3.3 Random sample and sample numerical characteristics

In practice we do not always have information available about all statistical units of the population under study. Therefore we often have to limit our analysis to a smaller part (subset, sample) of the statistical population. The set that is the subject of the research is called the *population*. The set of statistical units that have been selected from it in a certain way is called the *sample* (Fig. 18). The knowledge obtained by analyzing the sample can then be generalized to the population. The sample should be representative and should have a structure similar to the structure of the population. The most common way of achieving these properties is *random sampling*. This means that inclusion in the sample is decided by chance alone. Examples of such random sampling include drawing lots or using a random number generator. Random sampling is carried out from an unsorted population, and every statistical unit has the same probability of being selected.

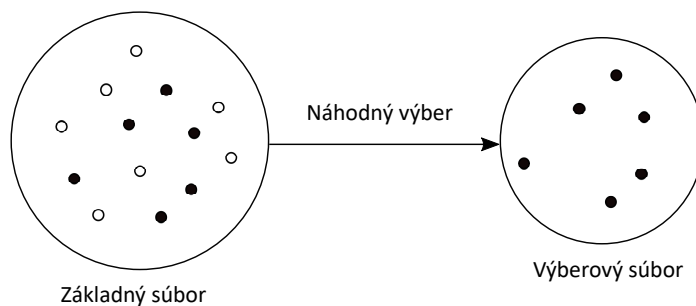


Figure 18: Population and sample

Let V_n denote a random sample from the population with measured values $x_1, x_2, \dots, x_n$ and size n . As we showed in the previous chapter, a statistical population can be described using various numerical characteristics. The numerical characteristics of the population are estimated using the corresponding sample characteristics. The basic sample numerical characteristics include the sample arithmetic mean, the sample variance, and the sample standard deviation.

The *sample arithmetic mean* $\bar{x}$ of the random sample V_n is defined as the sum of the values $x_1, x_2, \dots, x_n$ divided by their count n , i. e.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (71)$$

Formula (71) is used for an ungrouped statistical population. If the values of the statistical characteristic are grouped into k classes, then the sample arithmetic mean can be calculated using the formula

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k x_j n_j, \quad (72)$$

where n_j is the absolute frequency of the value x_j of the statistical characteristic.

The *sample variance (dispersion)* s^2 of the random sample V_n is defined as follows

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (73)$$

where $\bar{x}$ is the sample arithmetic mean of the random sample V_n . Formula (73) is used for an ungrouped statistical population. If the values of the statistical characteristic are grouped into k classes, then the sample variance can be calculated using the formula

$$s^2 = \frac{1}{(n-1)} \sum_{j=1}^k (x_j - \bar{x})^2 n_j, \quad (74)$$

where n_j is the absolute frequency of the value x_j of the statistical characteristic.

Remark: If the expected value of the random sample V_n is known, the sample variance is defined as the arithmetic mean of the squared deviations of the individual values $x_1, x_2, \dots, x_n$ from their expected value μ , i. e.

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2. \quad (75)$$

Formula (75) is used for an ungrouped statistical population. If the values of the statistical characteristic are grouped into k classes, then the sample variance s_0^2 can be calculated using

the formula

$$s_0^2 = \frac{1}{n} \sum_{j=1}^k (x_j - \mu)^2 n_j, \quad (76)$$

where n_j is the absolute frequency of the value x_j of the statistical characteristic.

The *sample standard deviation* s of the random sample V_n is the positive square root of the sample variance s^2 , i. e.

$$s = \sqrt{s^2}. \quad (77)$$

Example 3.1 When weighing twenty one-kilogram bags of sugar, we obtained the following values [in kg]: 1.00; 1.01; 1.05; 0.99; 0.95; 1.00; 0.98; 0.99; 1.04; 1.06; 0.93; 1.00; 1.03; 0.97; 1.00; 0.99; 1.05; 1.01; 0.94; 1.00. For the given sample, calculate

- a) the sample arithmetic mean,
- b) the sample variance,
- c) the sample standard deviation.

Solution:

- a) We calculate the sample arithmetic mean $\bar{x}$ using relation (71)

$$\begin{aligned} \bar{x} = \frac{1}{20} (1,00 + 1,01 + 1,05 + 0,99 + 0,95 + 1,00 + 0,98 + 0,99 + 1,04 + 1,06 + 0,93 + \\ + 1,00 + 1,03 + 0,97 + 1,00 + 0,99 + 1,05 + 1,01 + 0,94 + 1,00) = 0,9995. \end{aligned}$$

This value expresses the average amount of sugar in one bag, in kilograms.

- b) We calculate the sample variance s^2 using relation (73)

$$\begin{aligned} s^2 = \frac{1}{19} [(1,00 - 0,9995)^2 + (1,01 - 0,9995)^2 + (1,05 - 0,9995)^2 + (0,99 - 0,9995)^2 + \\ + (0,95 - 0,9995)^2 + (1,00 - 0,9995)^2 + (0,98 - 0,9995)^2 + (0,99 - 0,9995)^2 + (1,04 - 0,9995)^2 + \\ + (1,06 - 0,9995)^2 + (0,93 - 0,9995)^2 + (1,00 - 0,9995)^2 + (1,03 - 0,9995)^2 + (0,97 - 0,9995)^2 + \\ + (1,00 - 0,9995)^2 + (0,99 - 0,9995)^2 + (1,05 - 0,9995)^2 + (1,01 - 0,9995)^2 + (0,94 - 0,9995)^2 + \\ + (1,00 - 0,9995)^2] = 0,00126. \end{aligned}$$

- c) We calculate the sample standard deviation s using relation (77)

$$s = \sqrt{s^2} = 0,0355.$$

This value can be regarded, roughly, as the average deviation of the measured values from the mean value.