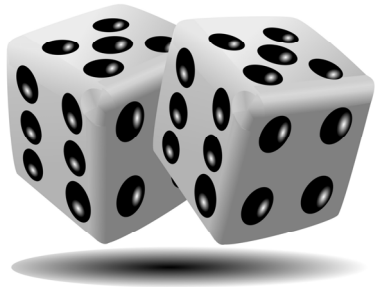
Probabilistic Expression of Random Variables

- Random Variable
- Probability and Probability Distribution
- Histogram
- Statistical Moments, Quantiles

Random Variable



A **random variable** (also called **random quantity**, **aleatory variable**, or **stochastic variable**) is a mathematical formalization of a quantity or object which depends on **random phenomenon**.

Random phenomenon means a repeatable activity performed under the same (or approximately the same) conditions, the result of which is **uncertain** and **depends on chance**.

Random variable is quantity whose values under constant conditions depend on **randomness**, and each of these values appears with a certain **probability**. It is any quantity that can be measured repeatedly in time and its values subjected to processing by methods of **probability theory** or **mathematical statistics**.

Probability

Probability of a random phenomenon is a number that is a measure of the expected occurrence of a random event (with what certainty a random event can be expected).

Probability of an event is a number between 0 and 1, where 0 indicates impossibility of the event and 1 indicates its certainty. The probability can also be given as a percentage (0 to 100%).

Example: Tossing of coin has two outcomes ("heads" and "tails") and both are equally probable (the probability of "heads" equals the probability of "tails"). Since no other outcomes are possible, the probability of either "heads" or "tails" is 0.5 or 50%.

Random Phenomenon

Examples of random events: dice roll, target shoot or lottery draw.

In the theory of structural reliability: **failure** (failure state) and the state when the structure is **reliable**.

Applies to:

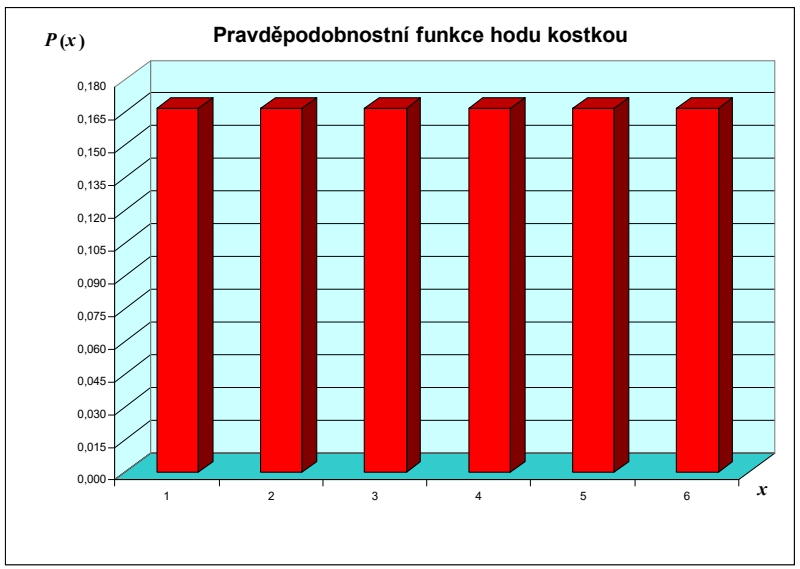
$$P_f + P_r = 1$$

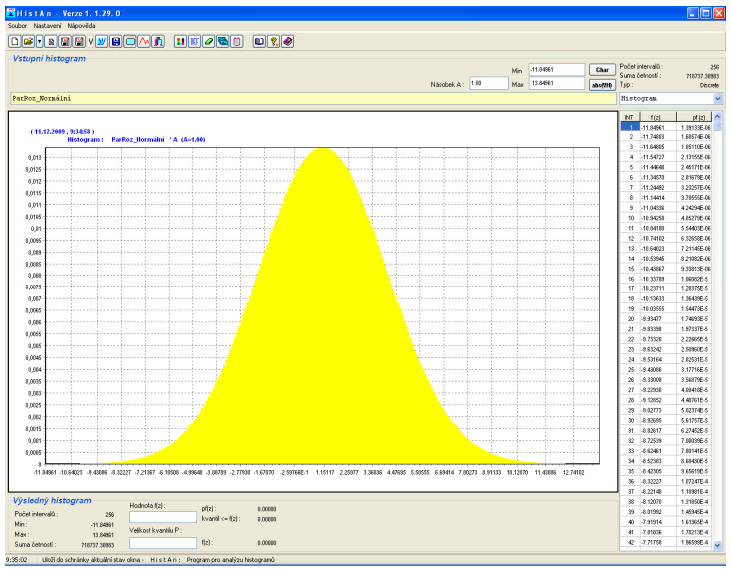
where P_f is **probability of failure** and P_r is probability that the structure will be preserved (is **reliable**).

Random Variable

Random variable can be:

- **discrete** - random variable is valued in a discrete set (such as a finite set),
- **continuous** - random variable is valued in an interval of real numbers.





Probability distribution: function that assigns probabilities to random events.

Discrete Random Variable

The random variable X is **discrete** if the elements of the selection space Ω appear on the axis of real numbers as isolated points $x_1, x_2, \cdots x_k$, each of these points having a nonzero probability.

Probability mass function (**PMF**) is a function f in probability theory and statistics that indicates the probability that a discrete random variable X equals the corresponding value x:

$$f_X(x) = P(X = x)$$

It also applies:

$$\sum_{x \in \Omega} f_X(x) = 1$$

Cumulative Distribution Function (**CDF**) is a function F that indicates the probability that a value of random variable X is less or equal than corresponding value x:

$$F_X(x) = P(X \leq x)$$

Continuous Random Variable

The random variable X is **continuous** if there is a non-negative function f, for which assuming $a < b$ applies:

$$P(a < X < b) = \int_a^b f_X(x)\mathrm{d}x$$

Function f is **probability density function** (**PDF**) for which it applies:

$$f_X(x) \geq 0$$

$$\int_\Omega f_X(x)\mathrm{d}x = 1$$

where Ω is definition field of a random variable.

Cumulative Distribution Function (**CDF**) is a non-decreasing function:

$$F_X(x) = P(X \leq x)$$

It applies:

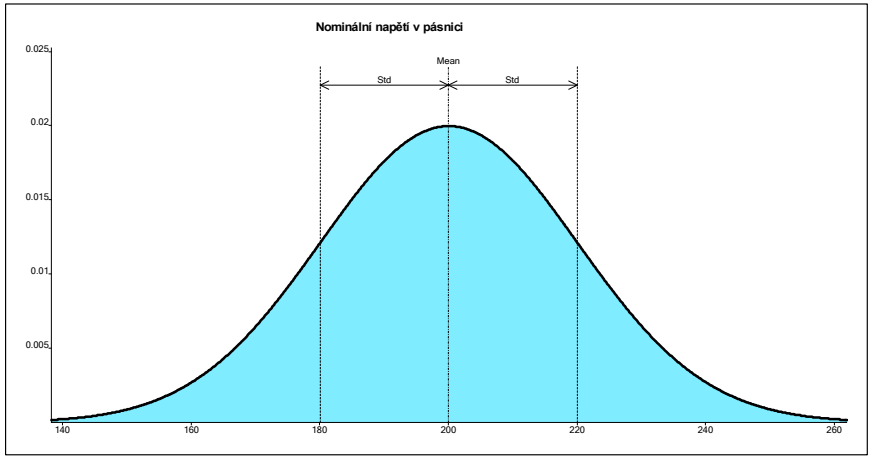
$$F_X(-\infty) = 0 \qquad F_X(\infty) = 1$$

$$f_X(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

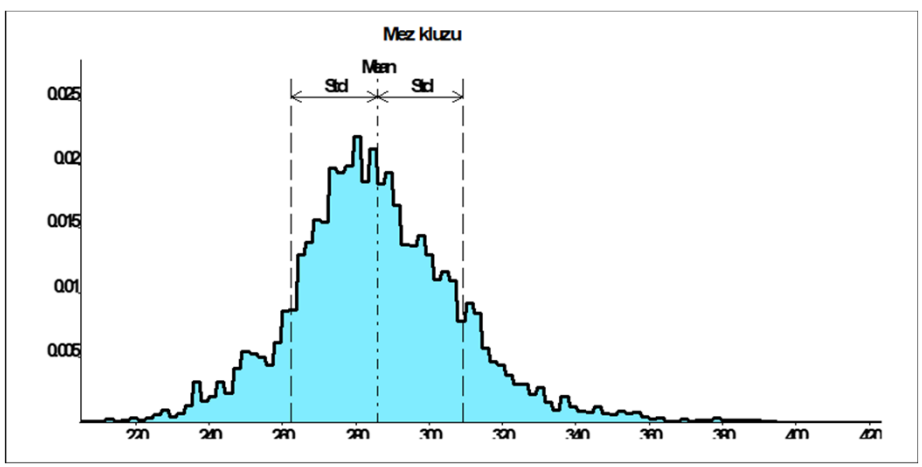
Probability Distribution

Parametric probability distribution - probabilities defined by analytical function – e.g., common expression of normal (Gaussian) probability distribution:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Parameters - characteristics of random variable probability distribution
(e.g., μ mean value and σ standard deviation)

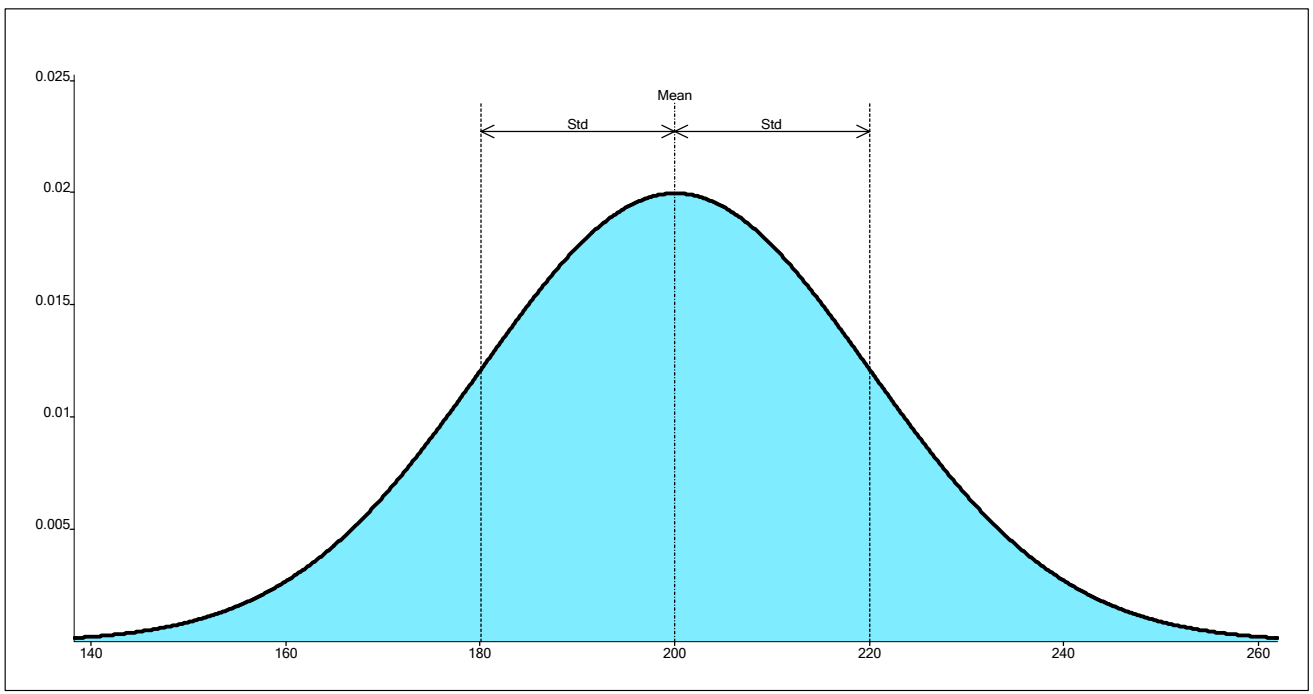


Non-parametric (empirical) **probability distribution** - definition based on measurements (often long-term)

Normal (Gaussian) Probability Distribution

Common expression of **normal** (**Gaussian**) **probability distribution**:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ … **mean value**

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

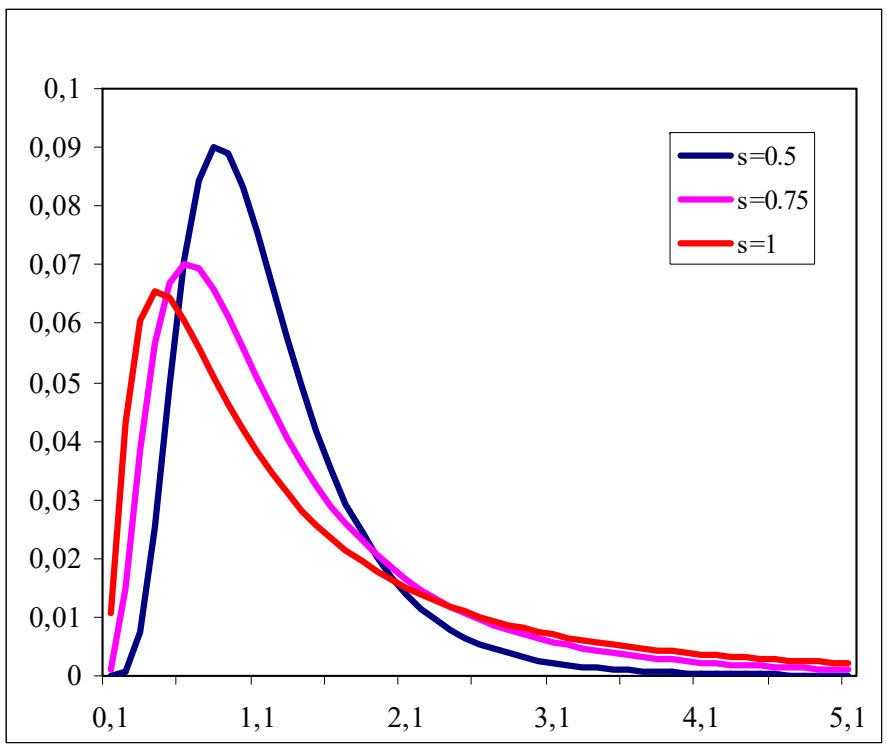
σ … **standard deviation**

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Log-Normal Probability Distribution

Common expression of **log-normal probability distribution**:

$$f(x|\mu,\sigma) = \frac{1}{x\sqrt{2\pi}\sigma} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}$$



μ … **mean value**

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

σ … **standard deviation**

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\ln(x_i) - \mu)^2}$$

Histogram / Non-Parametric Probability Distribution

Histogram is an approximate representation of the distribution of numerical data.

Carl Pearson
(1857-1936)

Construction of histogram: the first step is calculating the **range** of values - dividing the entire range of values into a **series of intervals**. Then counting how many values fall into each interval.

There are several rules for determining the optimal **number of classes k** depending on the **number of values n** in the file, for example:

($\lceil x \rceil$ … round up to the nearest integer) e.g., for $n = 100$

Sturges' rule: $k = \lceil 1 + \log_2 n \rceil$ $k = \lceil 1 + \log_2 100 \rceil = \lceil 7.643856 \rceil = \mathbf{8}$

Rice rule: $k = \lceil 2\sqrt[3]{n} \rceil$ $k = \lceil 2\sqrt[3]{100} \rceil = \lceil 9.283178 \rceil = \mathbf{10}$

Cumulative Distribution Function / Histogram

The **cumulative distribution function** of a real-valued continuous random variable X is the function given by:

$$F_X(x) = P(X \leq x)$$

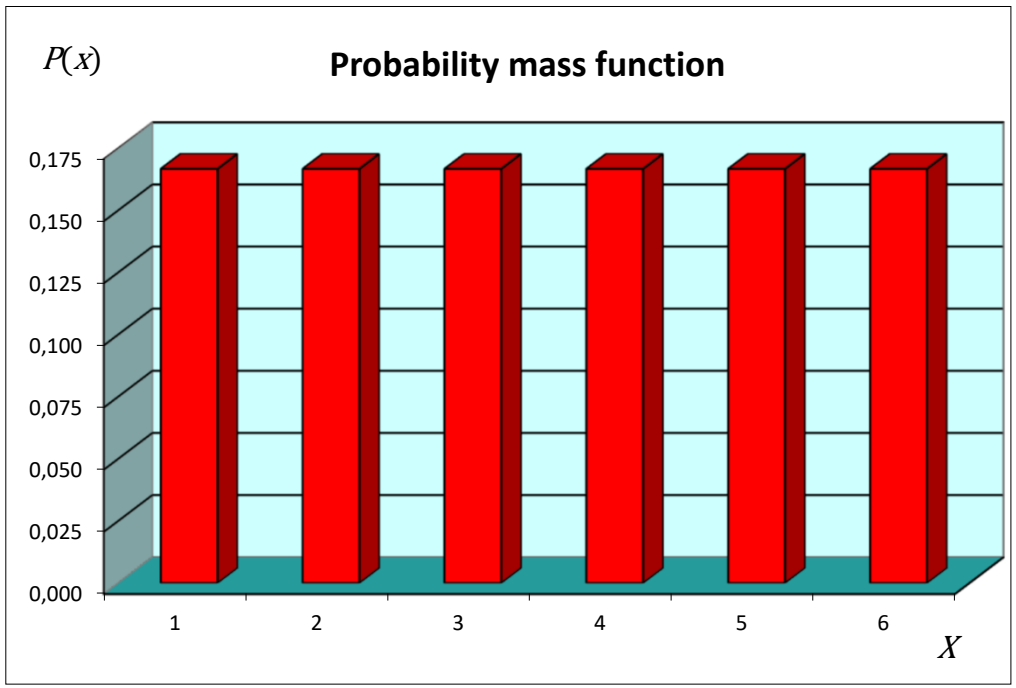
It applies:
(for $a < b$)

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\mathrm{d}x$$

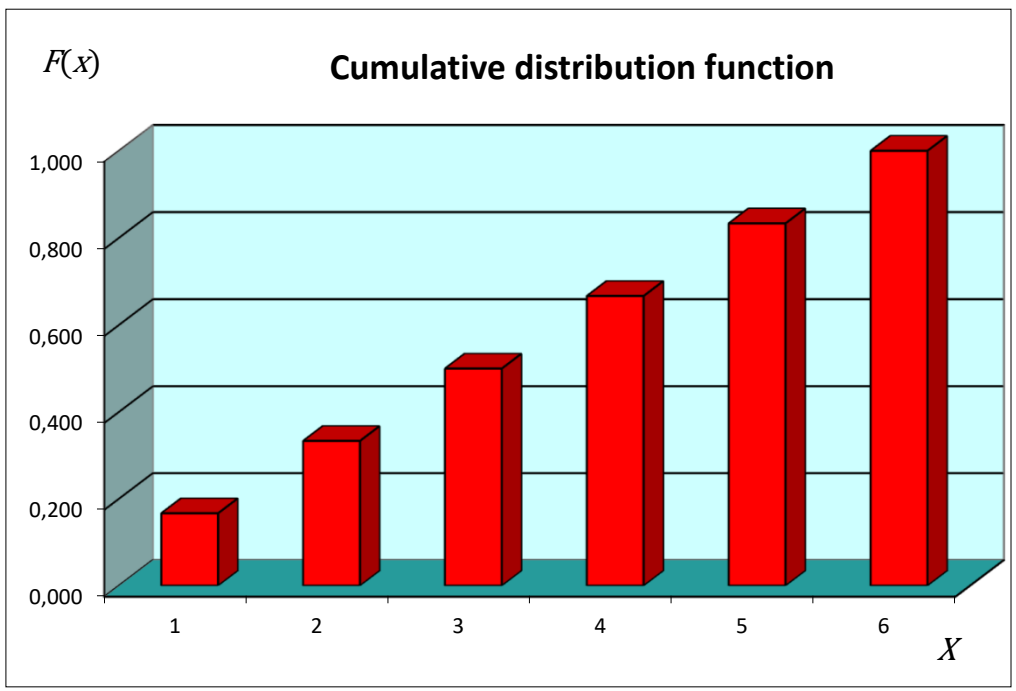
If X is **pure discrete** random variable:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i)$$

Cumulative Distribution Function / Histogram

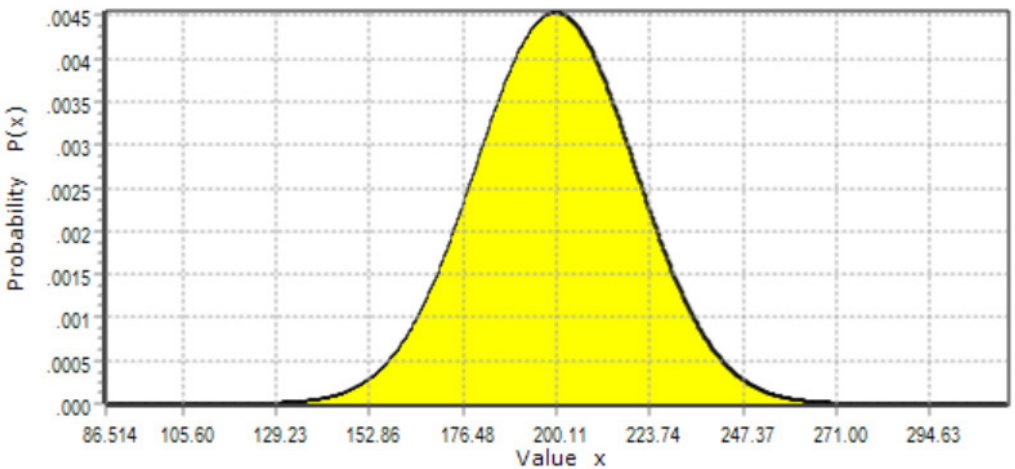


Histogram of pure discrete random variable
(random dice roll phenomenon)



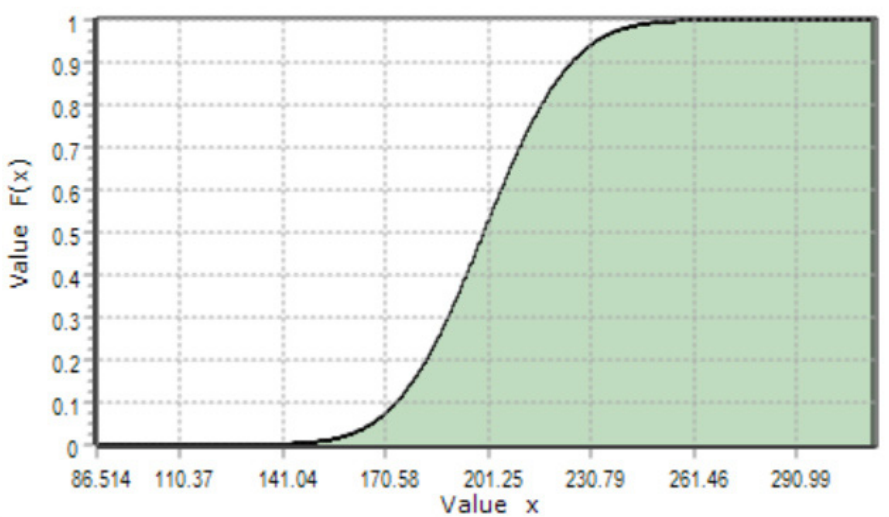
Corresponding **cumulative distribution function**

Cumulative Distribution Function / Histogram



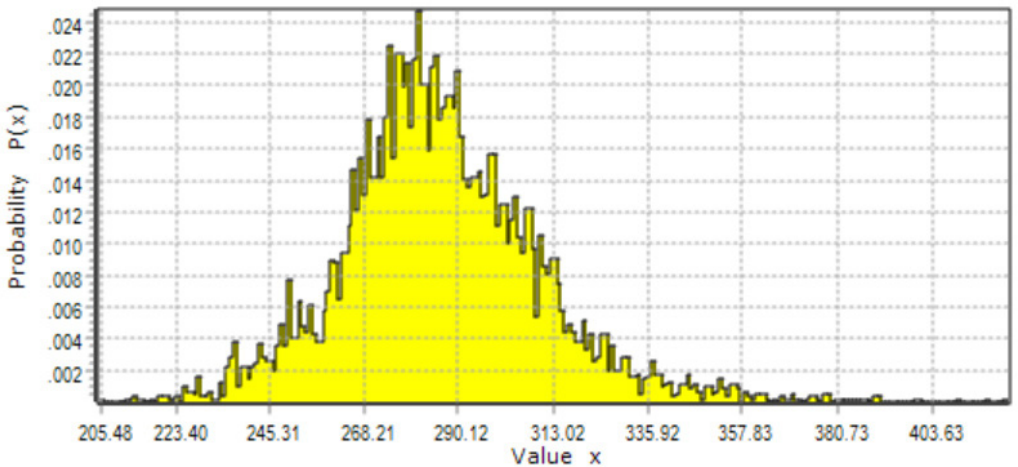
Histogram of discretized continuous random variable with parametric probability distribution

Normal probability distribution, 1000 intervals



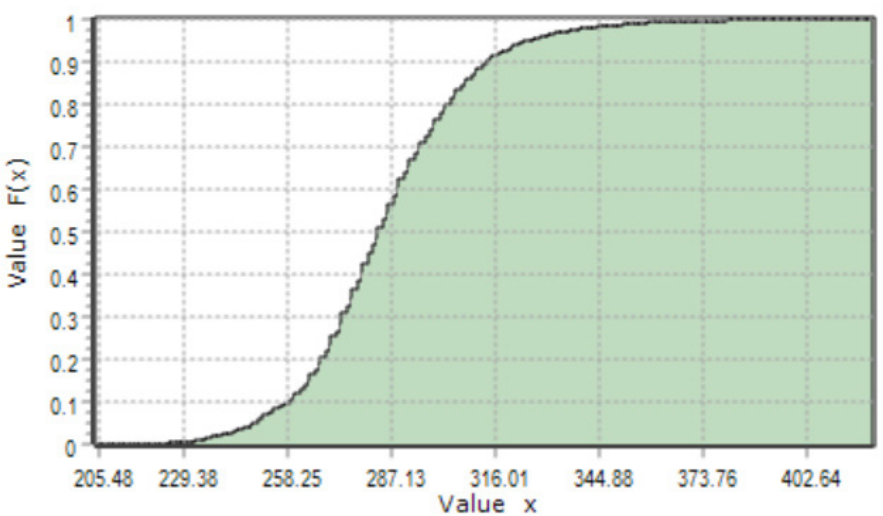
Corresponding **cumulative distribution function**

Cumulative Distribution Function / Histogram



Histogram of discretized continuous random variable with non-parametric (empirical) probability distribution

Yield stress of steel, 217 intervals



Corresponding **cumulative distribution function**

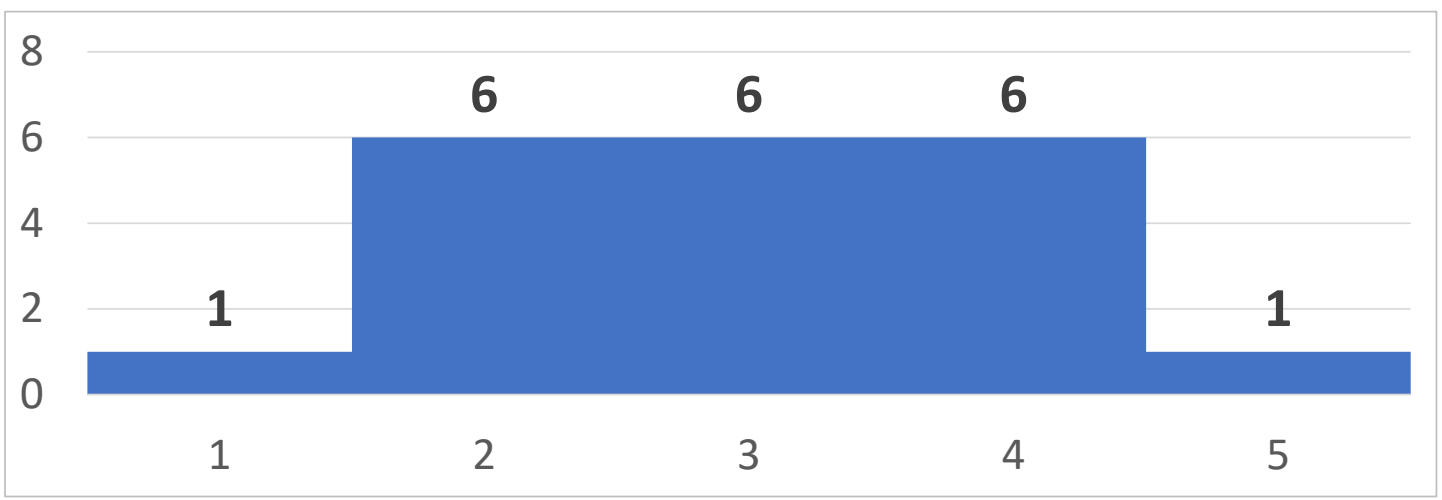
Example 1

Construct a histogram from the measured values:

Measurement number i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Measured value x_i	5	2	4	4	1	2	3	2	3	3	4	4	2	3	2	3	4	3	4	2

Resulting histogram of pure discrete random variable

(vertical axis - **frequency**)

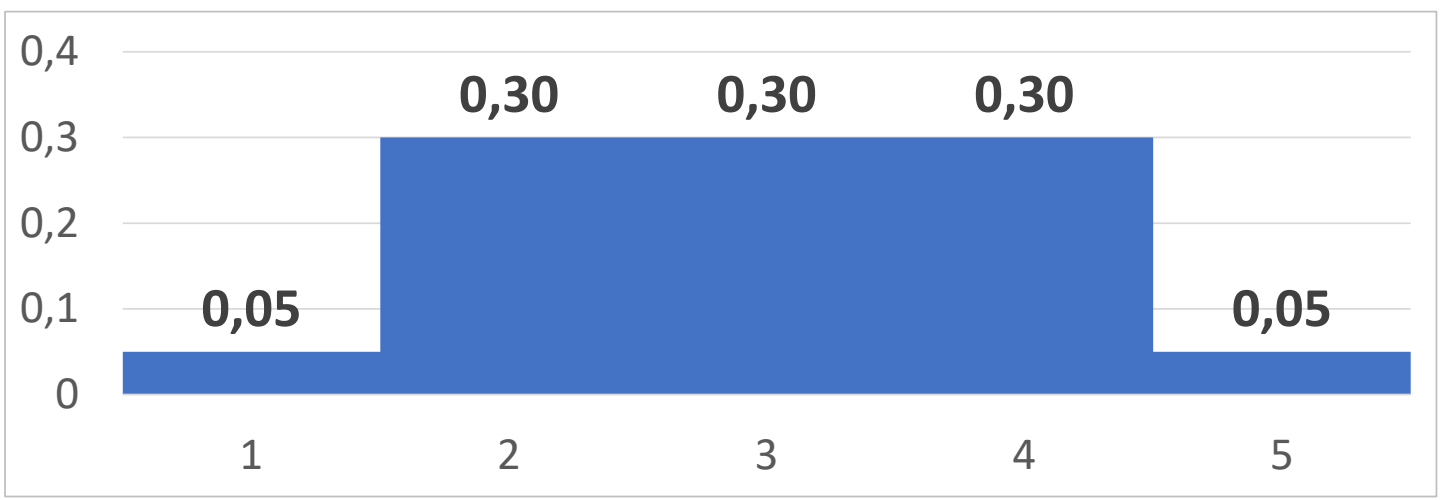


Example 1

Resulting histogram of pure discrete random variable from Example 1 can be described also using **probabilities**.

Value x_i of random variable X	1	2	3	4	5	Total
Frequency	1	6	6	6	1	**20**
Probability	1/20	3/10	3/10	3/10	1/20	**1**

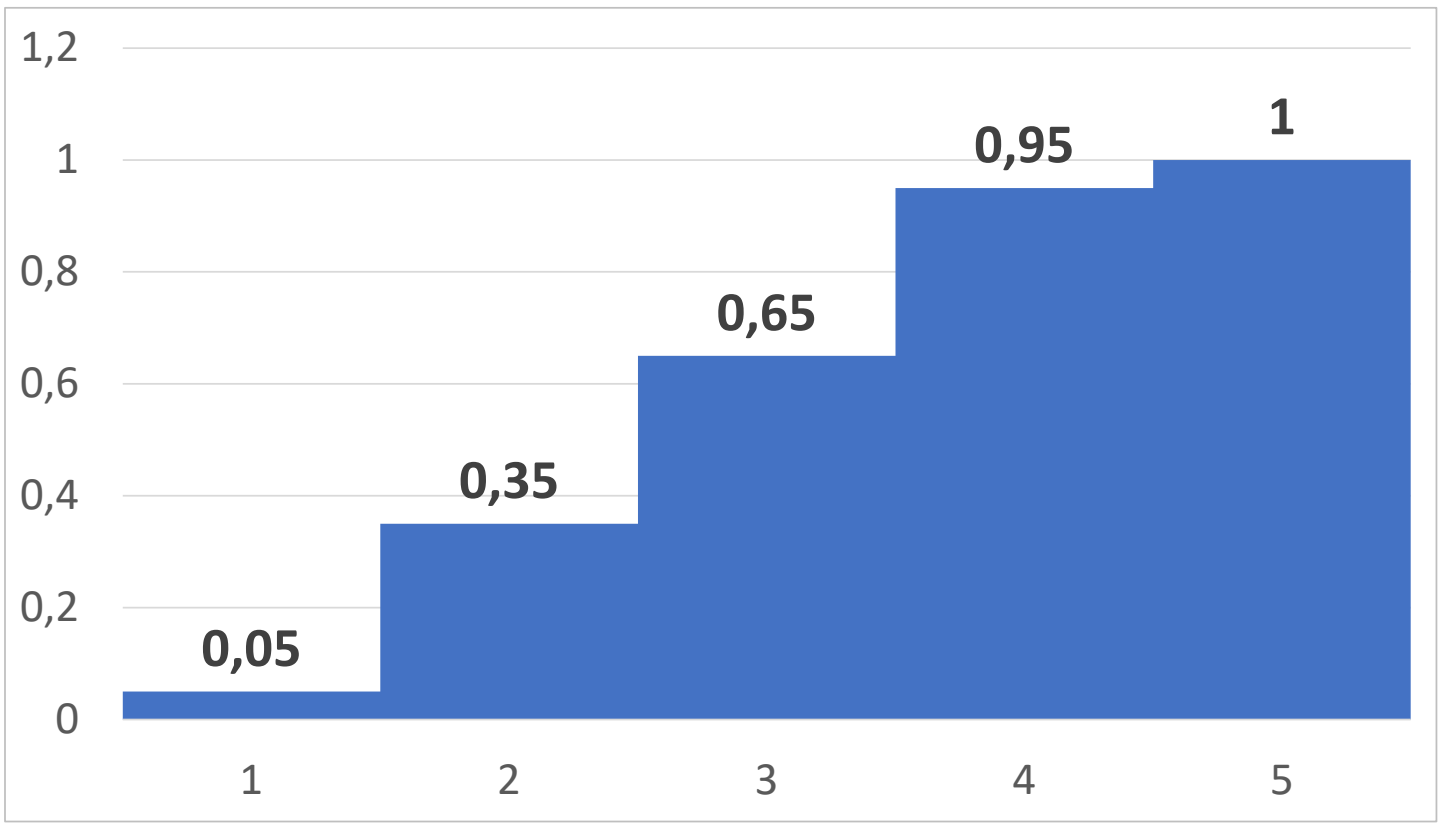
Resulting histogram of pure discrete random variable
(vertical axis - **probability**)



Approximation of Probability Distributions

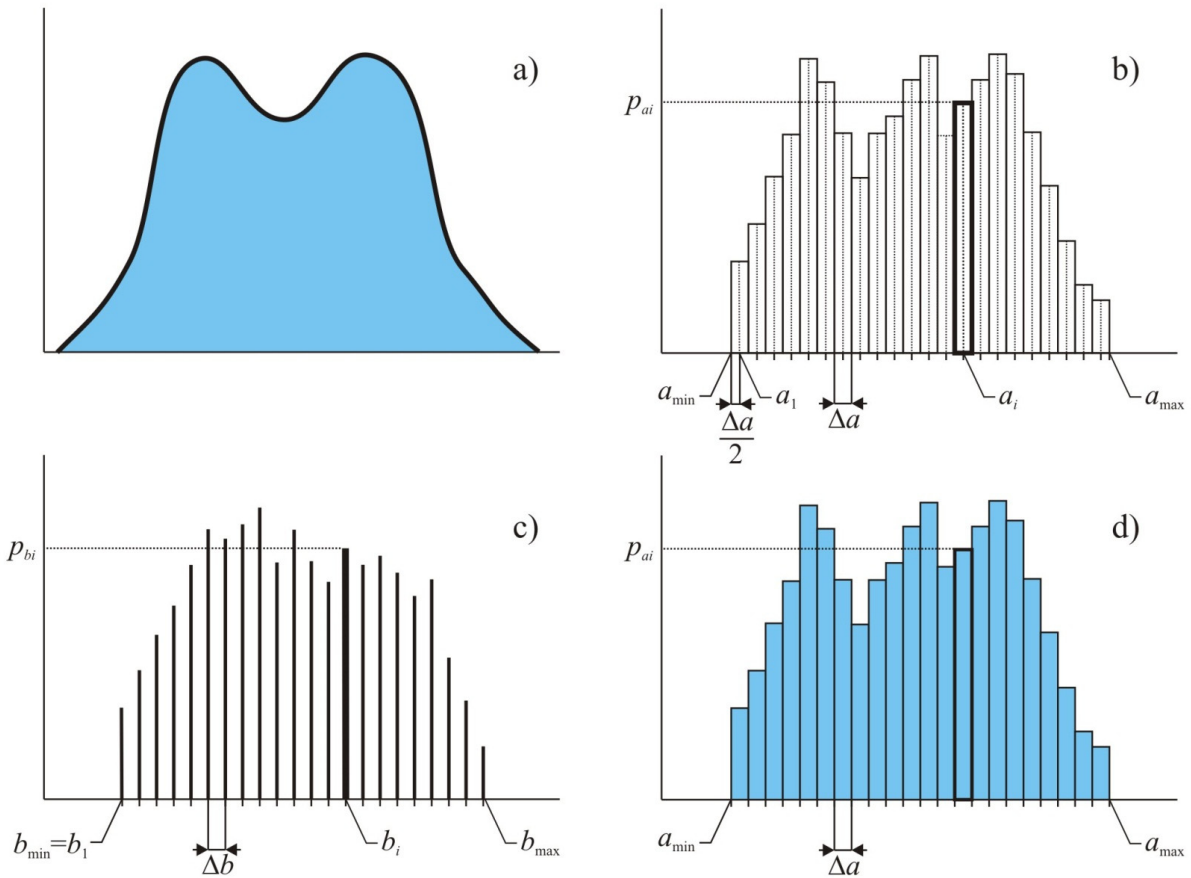
Construct a chart of cumulative distribution function:

Resulting chart of **cumulative distribution function** of pure discrete variable from Example 1



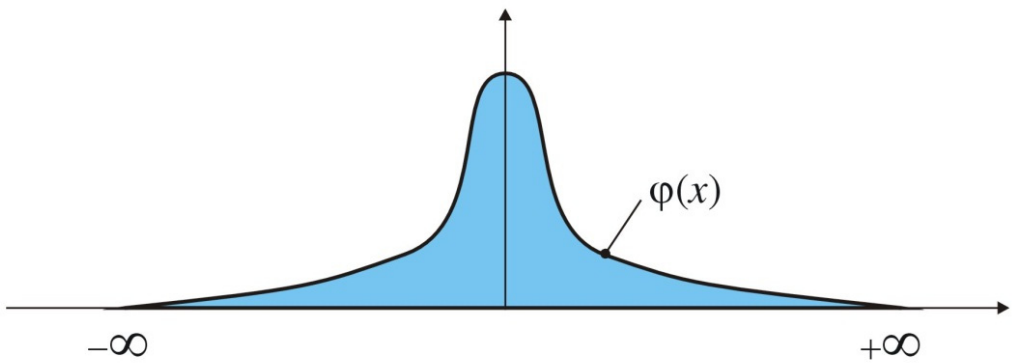
Approximation of Probability Distributions

a) Original approximation,
b) Discrete approximation,
c) Pure discrete approximation,
d) Piece-wise uniform approximation

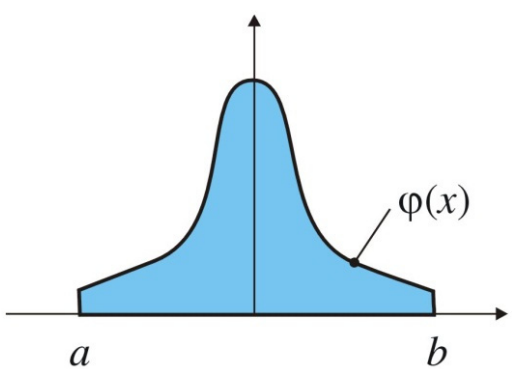


Limited Range of a Random Variable

Unlimited range of the continuous random variable



Limited range of the continuous random variable



Limited Range of a Random Variable

Limitation of the range of the probability distribution domain due to computer interpretation:

Range of data types:

Integer types:

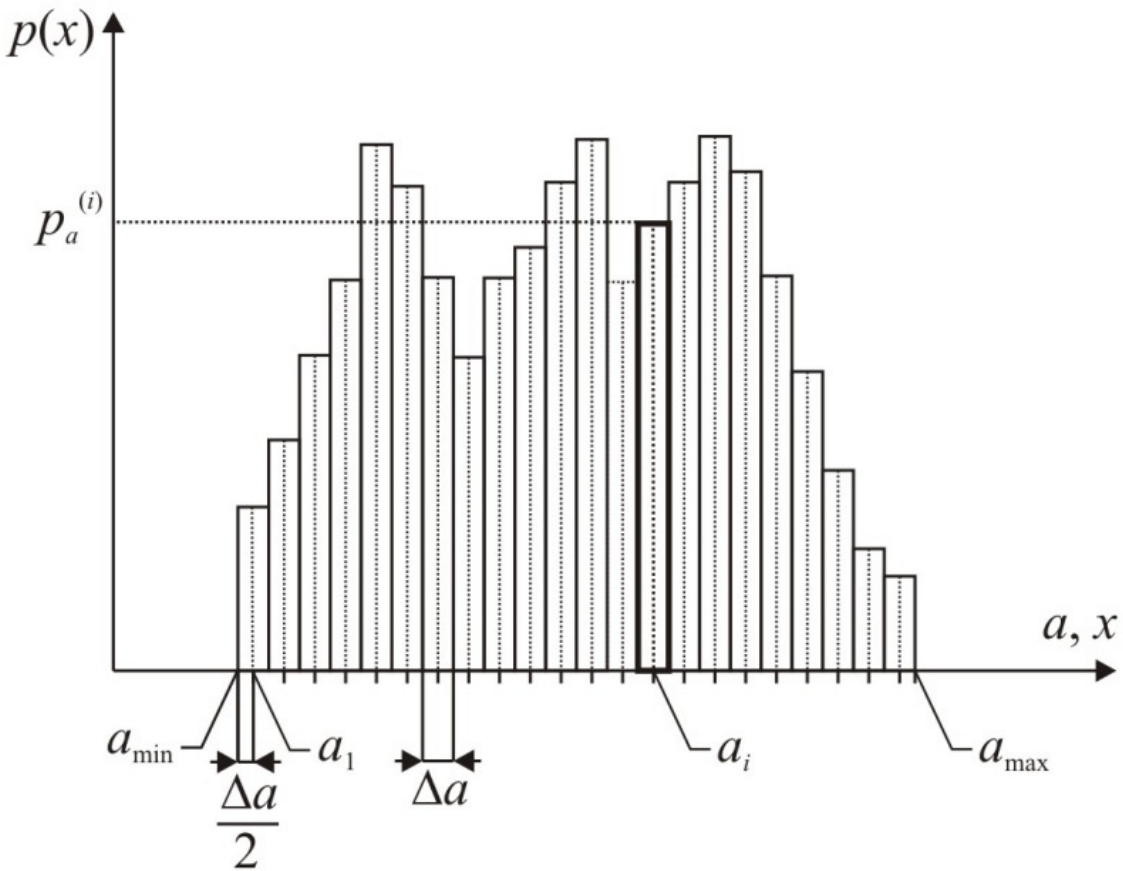
Byte (8 bits – 1 byte)	0 to 255
Integer (16 bits – 2 bytes)	$-32{,}768$ to $+32{,}767$
Word (16 bits – 2 bytes)	0 to $65{,}535$
LongInt (32 bits – 4 bytes)	$-2{,}147{,}483{,}648$ to $2{,}147{,}483{,}647$

Floating point types:

Float (32 bits – 4 bytes)	$\pm\ 3.4 \cdot 10^{-38}$ to $3.4 \cdot 10^{38}$
Double (64 bits – 8 bytes)	$\pm\ 1.7 \cdot 10^{-308}$ to $1.7 \cdot 10^{308}$
Long double (80 bits – 10 bytes)	$\pm\ 3.4 \cdot 10^{-4932}$ to $3.4 \cdot 10^{4932}$

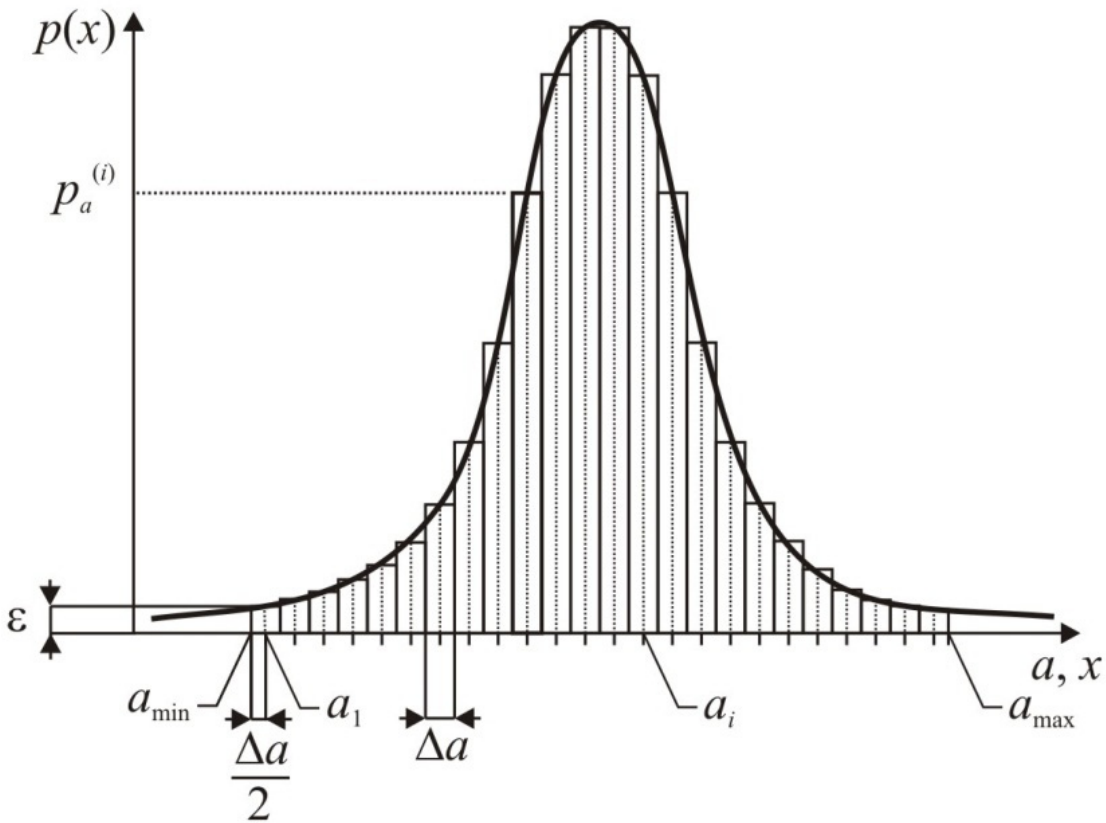
Histogram of Random Variable

Histogram of discretized continuous random variable with non-parametric (empirical) probability distribution



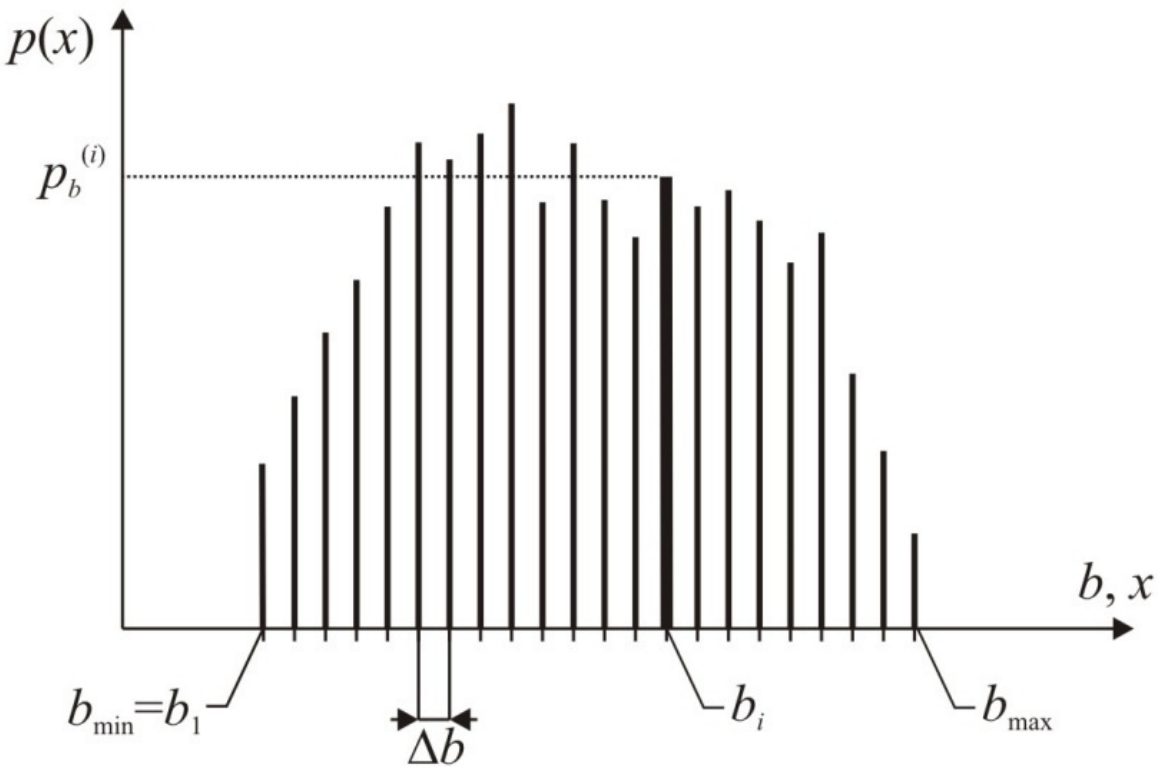
Histogram of Random Variable

Histogram of discretized continuous random variable with parametric probability distribution



Histogram of Random Variable

Histogram of pure discrete random variable



Structure of the Data File / Histogram Definition

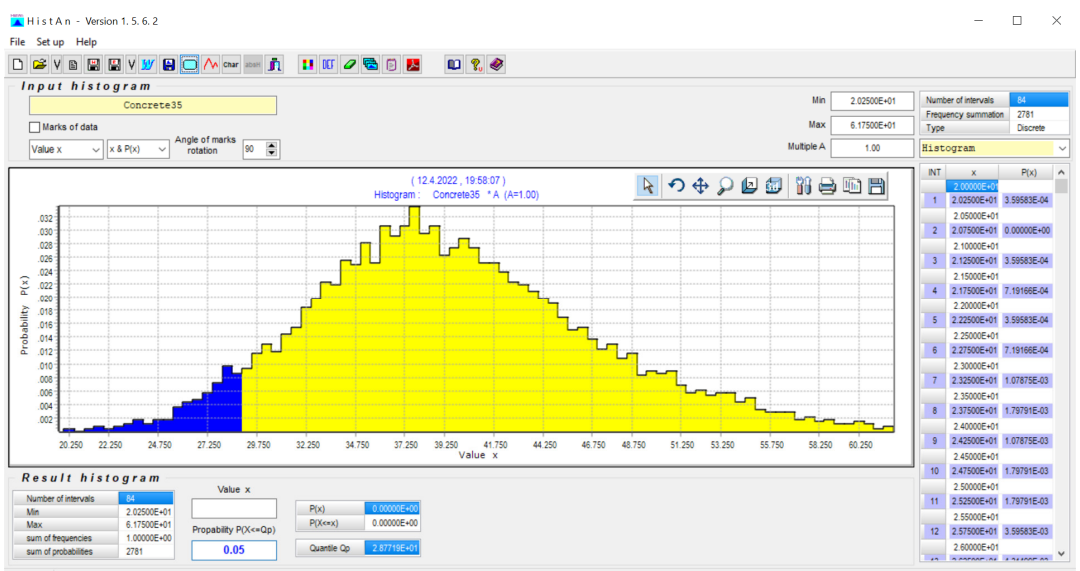
A **text file** with the extension ***.dis** (distribution), which contains data in the following form:

```
[Description] (1st section of the data file)
Identification= Optional data file description
Type= Pure Discrete | Discrete | Continuous (Histogram
type of random variable)

[Parameters] (2nd section of the data file)
Min= Minimum value of a random variable
Max= Maximum value of a random variable
Bins= Total number of classes in the histogram
Total= Sum of the frequencies in all classes

[Bins] (3rd section of the data file)
frequency in 1st class
frequency in 2nd class
etc. ...
```

HistAn Software Tool

Program for more detailed **analysis of input histograms**:

- **Minimum** and **maximum values** of a random variable
- **Number of histogram classes** (intervals) and frequencies defined in them
- **Simple probabilistic calculations** with histograms (determination of p-quantile and probability of exceedingly the determined value of a random variable)
- Determining the **combination of several input histograms**
- Creation of **histograms with parametric distribution**
- Processing of **measured raw data**

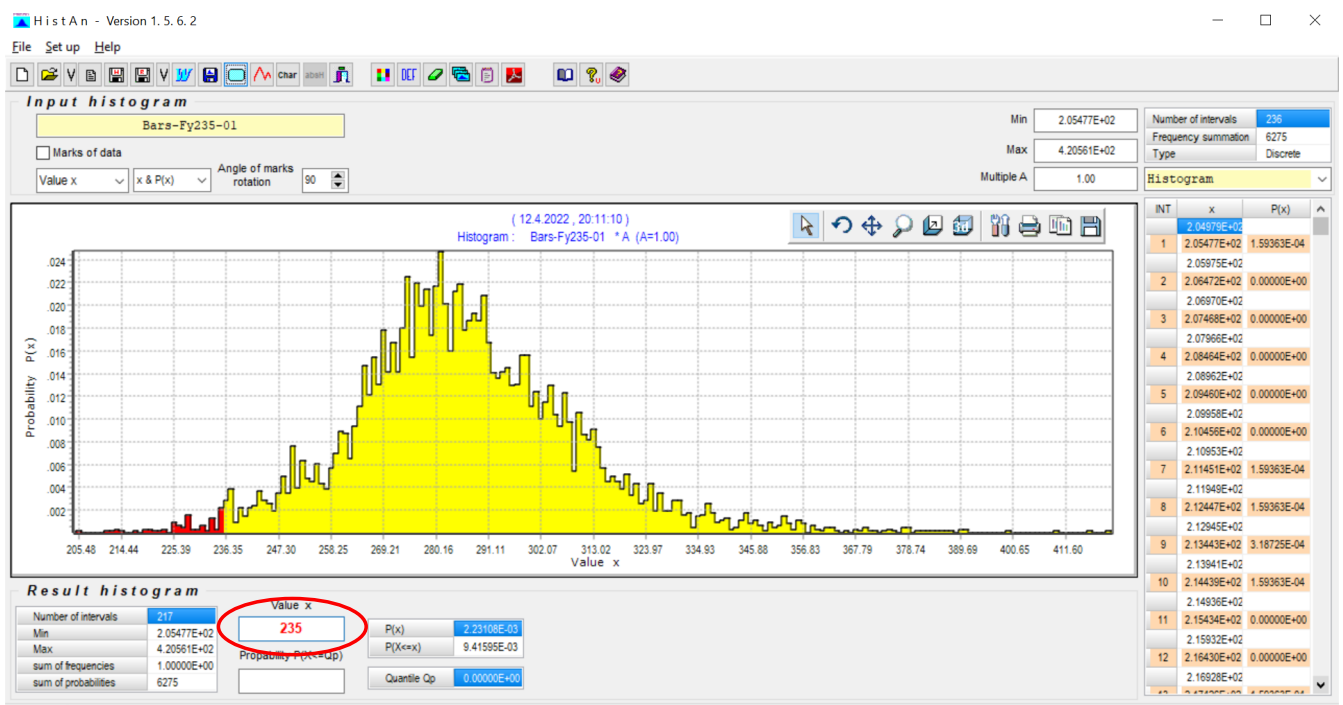


Desktop of the **HistAn software tool**

HistAn Software Tool

Detailed analysis of input histogram of the **yield stress of the steel S235**:

- Calculation of probability of exceedingly the determined value of the yield stress (value of random variable $x = 235\ \mathrm{MPa}$, resulting probability is $P(X \leq x) = 9.41595 \cdot 10^{-3}$)



Desktop of the **HistAn software tool**

Statistical Moments of Random Variable

The **moments of a random variable** (or of its distribution) are expected values of powers or related functions of the random variable.

μ … **mean value**

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

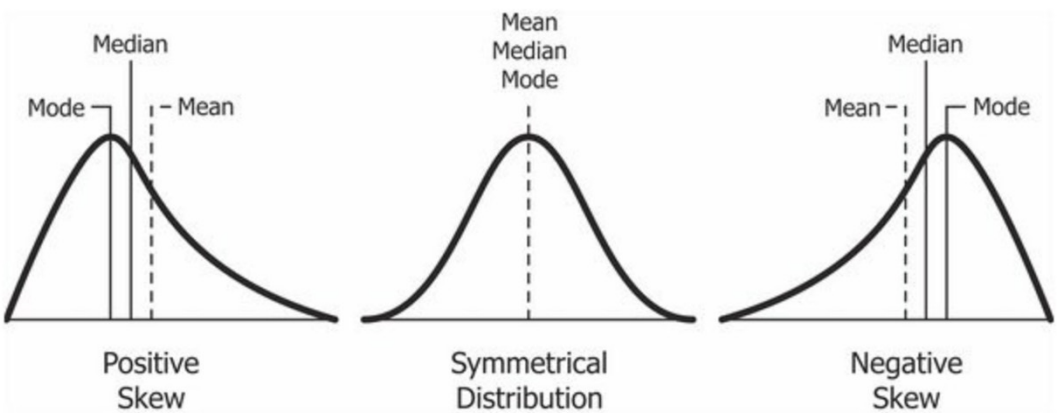
σ … **standard deviation**

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

a … **skewness**

(asymmetry of the distribution of the values around their average)

$$a = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3}$$



b … **kurtosis**

(concentration of the values around their average)

Quantiles

Quantiles are cut points dividing the range of a probability distribution into continuous intervals with **equal probabilities**.

Common quantiles have special names, such as **quartiles** (four groups), **deciles** (ten groups), and **percentiles** (100 groups).

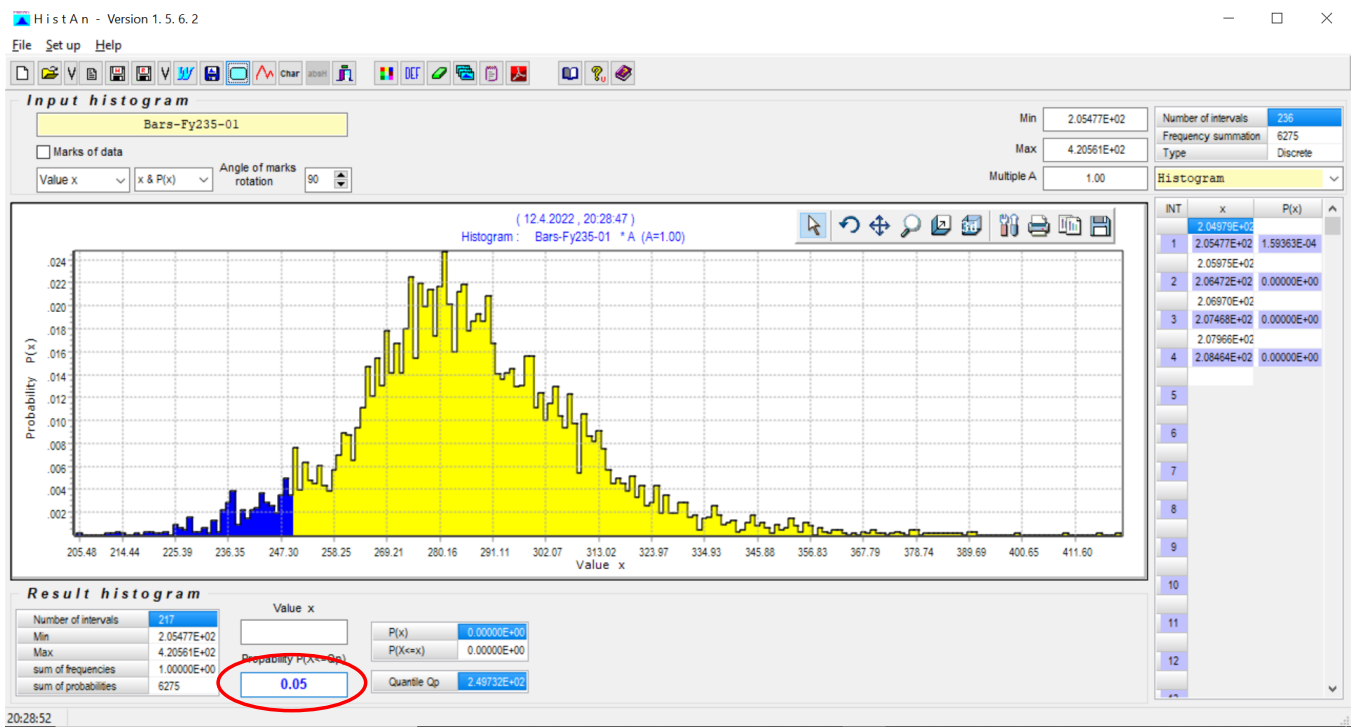
p-quantile - the quantile of the random variable X, which separates $p \cdot 100\ \%$ of the smaller values from the rest of the set, i.e., from $(1-p) \cdot 100\ \%$ values, is called the $p \cdot 100\ \%$ quantile (e.g., for $p = 0.05$ it is **5% quantile**) and denote it x_p (e.g., $x_{0.05}$).

Median ($x_{0.5}$) = 50% quantile, divides the data set so that half (50%) values are less than the median and half (50%) values are greater than median (or equal).

HistAn Software Tool

Detailed analysis of input histogram of the **yield stress of the steel S235**:

- Calculation of **five percent quantile $x_{0.05}$** of the yield stress (value of the specified probability $P(X \leq x_{0.05}) = 0.05$, resulting quantile $\boldsymbol{x_{0.05} = 249.732\ \mathrm{MPa}}$)



Desktop of the **HistAn software tool**