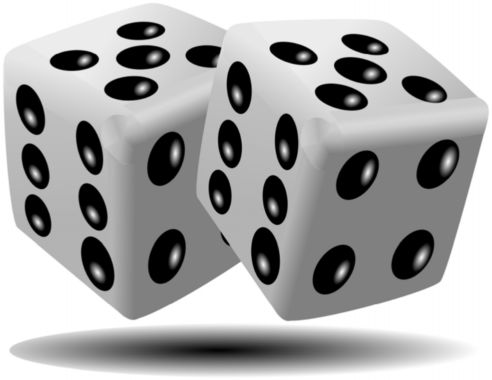
Téma 2:

Pravděpodobnostní vyjádření náhodných veličin

- Náhodná veličina
- Pravděpodobnost a rozdělení pravděpodobnosti
- Histogram
- Statistické momenty, kvantil

Náhodný jev, pravděpodobnost



Náhodný jev - opakovatelná činnost prováděná za stejných podmínek, jejíž výsledek je nejistý (různý) a závisí na náhodě.

Pravděpodobnost náhodného jevu je číslo, vyjadřující s jakou jistotou lze náhodný jev očekávat.

Pravděpodobnost náhodného jevu se obecně označuje reálným číslem od 0 do 1. Náhodný jev, který nemůže nastat, má pravděpodobnost 0. Náhodný jev jistý má naopak pravděpodobnost 1. Pravděpodobnost lze také uvádět v procentech (0 až 100 %).

Náhodný jev, pravděpodobnost

Příklady náhodných jevů: hod hrací kostkou, házení šipkou nebo loterie.

V teorii spolehlivosti konstrukcí: **porucha** - poruchový stav (failure) a stav, kdy je konstrukce **spolehlivá** (reliable).

Platí:

$$P_f + P_r = 1$$

kde P_f je **pravděpodobnost poruchy** a P_r je pravděpodobnost, že konstrukce bude zachovaná (je **spolehlivá**).

Náhodný jev, pravděpodobnost

Příklad: Hod mincí může mít dva výsledky („panna“ nebo „orel“) a oba jsou stejně pravděpodobné (pravděpodobnost, že padne „panna“, se rovná pravděpodobnosti, že padne „orel“). Vzhledem ke skutečnosti, že nejsou možné žádné jiné výsledky, pravděpodobnost, že padne „panna“ nebo „orel“ je 0,5 nebo 50 %.

Házení mincí je oblíbený způsob náhodného losování ze dvou možností se stejnou pravděpodobností (např. ve sportu pro rozhodování, na které straně hřiště začne příslušné mužstvo).

Označení „panna“ nebo „orel“ pochází z dob Rakouska-Uherska - na tolaru bývala na líci silueta Marie-Terezie („panna“) a na rubu státní znak s orlem.

Náhodný jev, pravděpodobnost

Příklad: Hod hrací kostkou má šest elementárních jevů, kdy může padnou jedno z čísel 1, 2, 3, 4, 5, 6 na stranách kostky.

Pak lze definovat např.:

Pravděpodobnost, že padne číslo 1: $P(1) = \frac{1}{6} = 0{,}1\overline{6}$ (1 ze 6 možností).

Pravděpodobnost, že padne sudé číslo: $P(2{,}4{,}6) = \frac{3}{6} = 0{,}5$ (3 ze 6 možností).

Pravděpodobnost, že padne číslo 7: $P(7) = 0$ (jev nemožný).

Pravděpodobnost, že padne číslo < 7: $P(1{,}2{,}3{,}4{,}5{,}6) = 1$ (jev jistý).

Náhodná veličina

Výsledky některých náhodných jevů lze vyjádřit číselně. Tato čísla tvoří obor hodnot **náhodné veličiny** (**náhodné proměnné**).

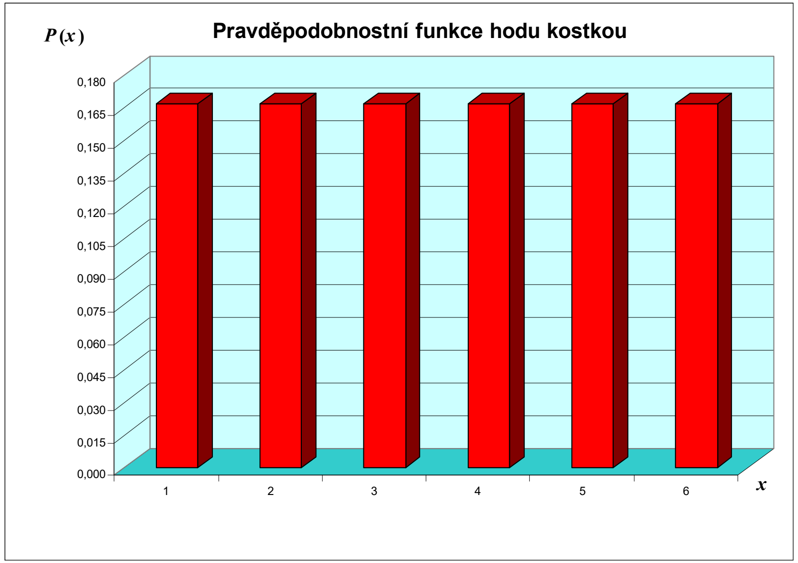
Náhodná veličina: proměnná, kterou je možné opakovaně měřit v čase. Její hodnoty při konstantních podmínkách závisí na náhodě, přičemž každá z těchto hodnot vystupuje s určitou pravděpodobností. Tyto hodnoty lze podrobit zpracování metodami teorie pravděpodobnosti nebo matematické statistiky.

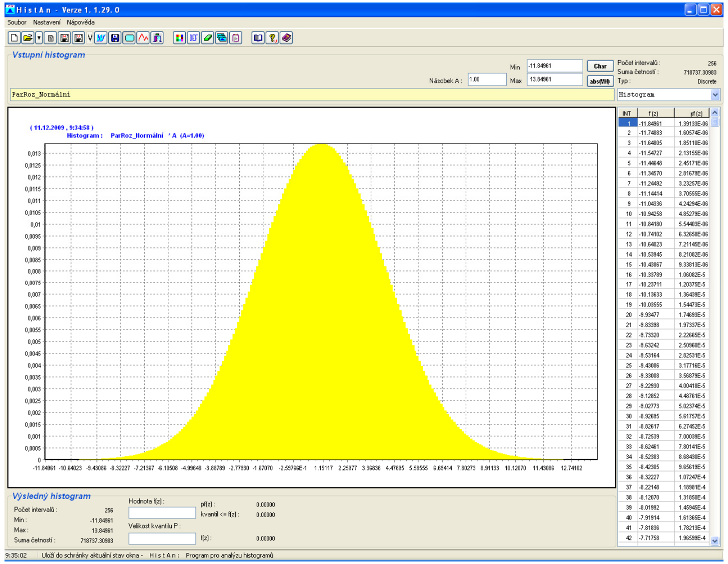
Náhodná veličina

Náhodná veličina může být:

- **diskrétní** – obor hodnot je konečná nebo nekonečná množina čísel,
- **spojitá** – obor hodnot je otevřený nebo uzavřený interval reálných čísel.

Rozdělení pravděpodobnosti: funkce, která přiřazuje náhodným veličinám pravděpodobnosti.





Diskrétní náhodná veličina

Náhodná veličina X je **diskrétní** (**discrete**), jestliže se prvky výběrového prostoru Ω zobrazí na ose reálných čísel jako samostatné body $x_1, x_2, \cdots x_k$, přičemž každý z těchto prvků má nenulovou pravděpodobnost.

Pravděpodobnostní funkce (**Probability mass function** - **PMF**) je funkce f, která udává pravděpodobnost pro odpovídající hodnotu x diskrétní náhodné proměnné X:

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

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

Distribuční funkce (**Cumulative Distribution Function** - **CDF**) je funkce F, která udává pravděpodobnost hodnoty náhodné proměnné X, která je menší nebo rovna zadané hodnotě x:

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

Spojitá náhodná veličina

Spojitou (**continuous**) náhodnou veličinu X lze popsat nezápornou funkcí f, nazývanou **hustota pravděpodobnosti** (**probability density function** - **PDF**), pro kterou za předpokladu $a < b$ platí:

$$f_X(x) \geq 0$$

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

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

kde Ω výběrový prostor náhodné proměnné.

Distribuční funkce (**Cumulative Distribution Function** - **CDF**) je neklesající funkce:

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

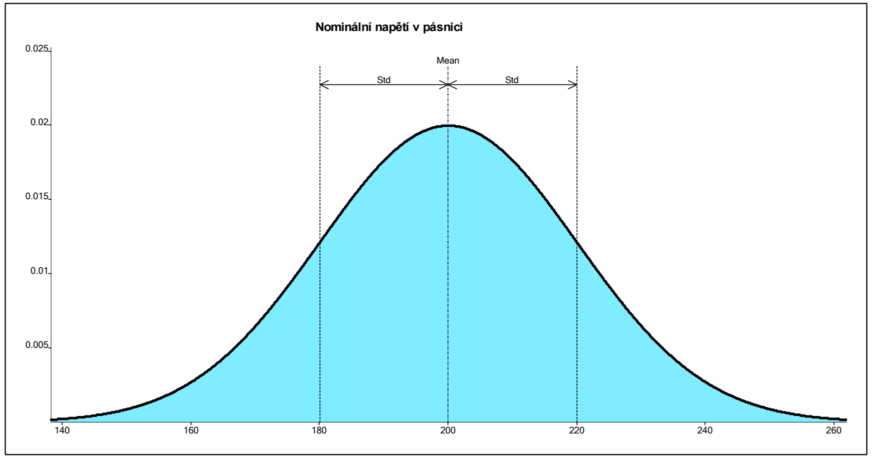
$$f_X(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

Rovněž platí:

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

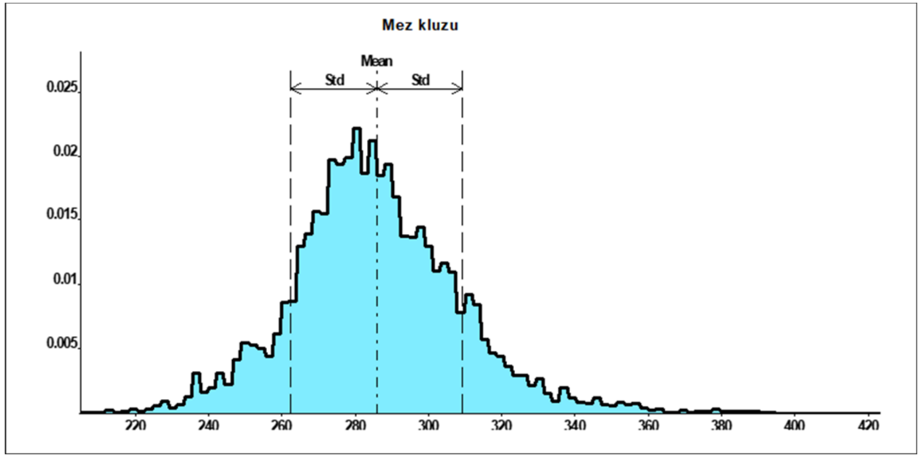
Rozdělení pravděpodobnosti

Parametrické rozdělení pravděpodobnosti – pravděpodobnosti jsou definovány analytickou funkcí – např. obecný vzorec funkce normálního (Gaussova) rozdělení pravděpodobnosti:



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

Parametry – charakteristiky funkce hustoty pravděpodobnosti náhodné proměnné (např. μ střední hodnota a σ směrodatná odchylka)

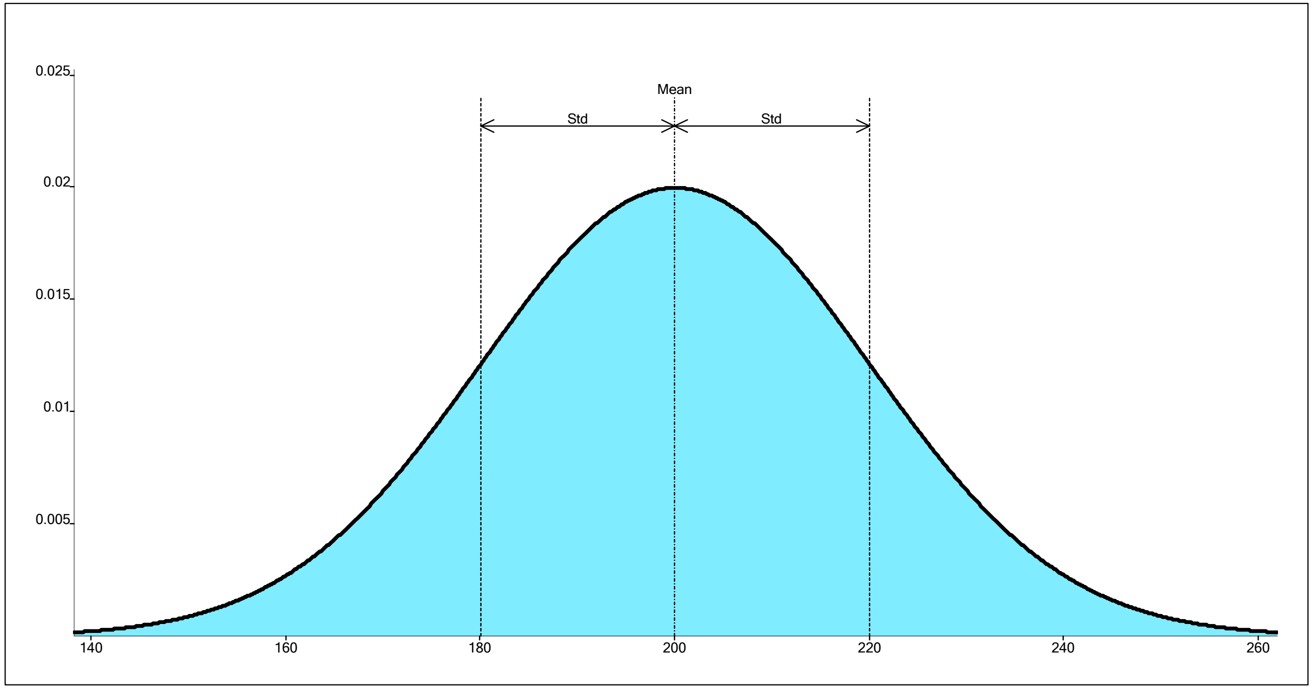


Neparametrické (empirické) **rozdělení pravděpodobnosti** – definováno na základě měření a monitoringu.

Normální (Gaussovo) rozdělení pravděpodobnosti

Obecný vzorec funkce **normálního (Gaussova) rozdělení pravděpodobnosti**:

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



Střední hodnota (mean value) μ:

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

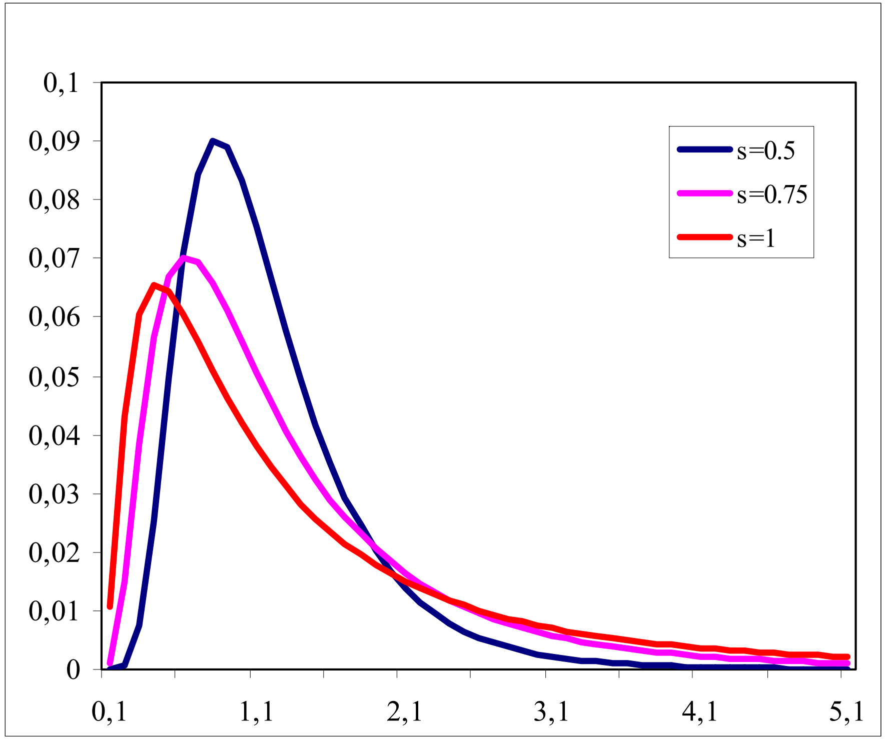
Směrodatná odchylka (standard deviation) σ:

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

Log-normální rozdělení pravděpodobnosti

Obecný vzorec funkce **logaritmicko-normálního** (**log-normálního**) **rozdělení pravděpodobnosti**:

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



Střední hodnota (mean value) μ:

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

Směrodatná odchylka (standard deviation) σ:

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

Histogram / Neparametrické rozdělení pravděpodobnosti

Histogram je grafické vyjádření rozdělení empiricky získaných dat pomocí sloupcového grafu s vhodným počtem sloupců (intervalů, tříd) stejné šířky. Výška sloupců vyjadřuje četnost dat nebo pravděpodobnost výskytu v intervalu.

Carl Pearson
(1857-1936)

Sestrojení histogramu: **1)** Výpočet **rozsahu naměřených** hodnot (např. max-min), který se rozdělí do k intervalů. **2)** Výpočet kolik naměřených hodnot spadá do příslušného intervalu.

Nesprávná šířka intervalů (nesprávný počet k) může snížit informační hodnotu histogramu. Pro optimální počet intervalů k slouží několik postupů, které jsou závislé na počtu naměřených hodnot n v souboru dat:

Sturgesovo pravidlo: $k = \lceil 1 + \log_2 n \rceil$

Riceovo pravidlo: $k = \lceil 2\sqrt[3]{n} \rceil$

Např. pro $n = 100$:

$$k = \lceil 1 + \log_2 100 \rceil = \lceil 7.643856 \rceil = \mathbf{8}$$

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

Poznámka: $\lceil x \rceil$ ... zaokrouhlení nahoru na nejbližší celé číslo

Distribuční funkce / Histogram

Distribuční funkce spojité náhodné veličiny X je nezáporná funkce:

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

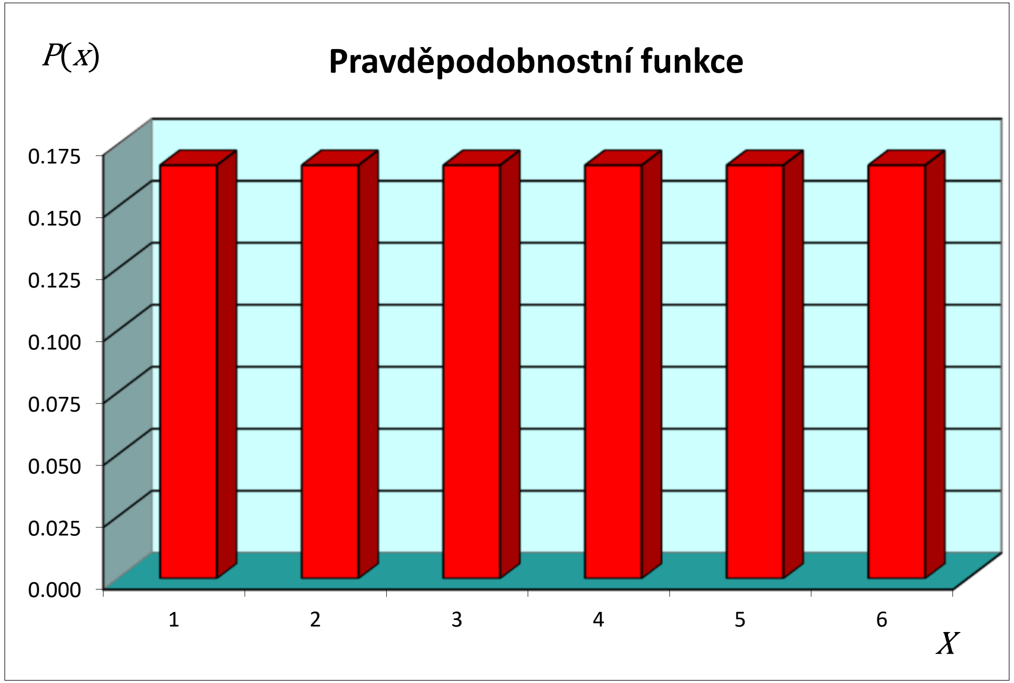
Platí (pro $a < b$):

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

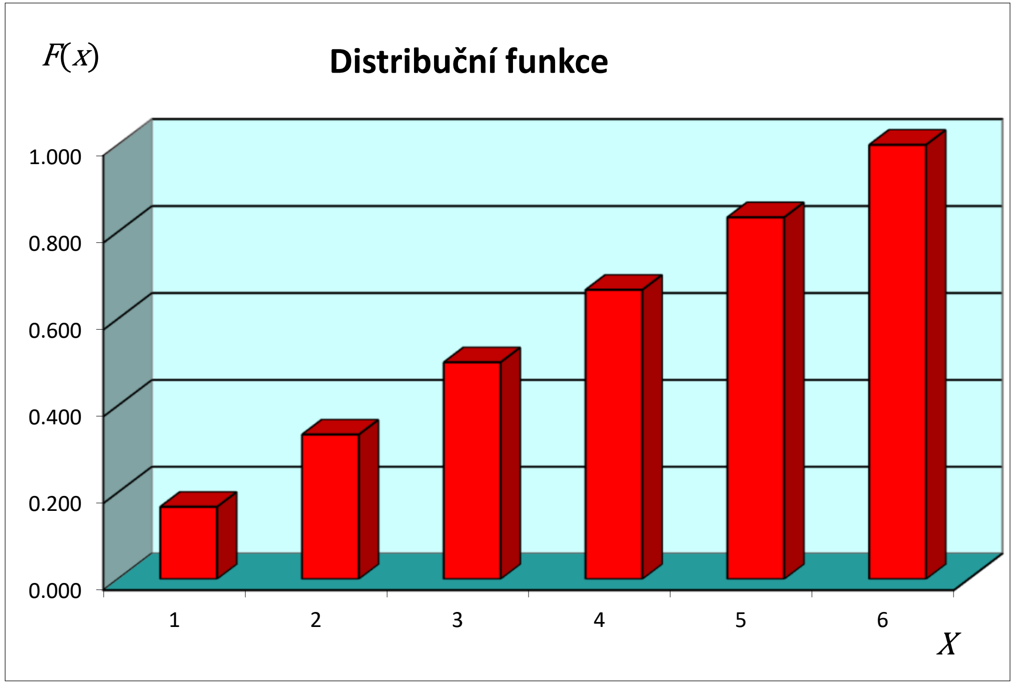
Pokud je X **diskrétní** náhodná veličina:

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

Distribuční funkce / Histogram

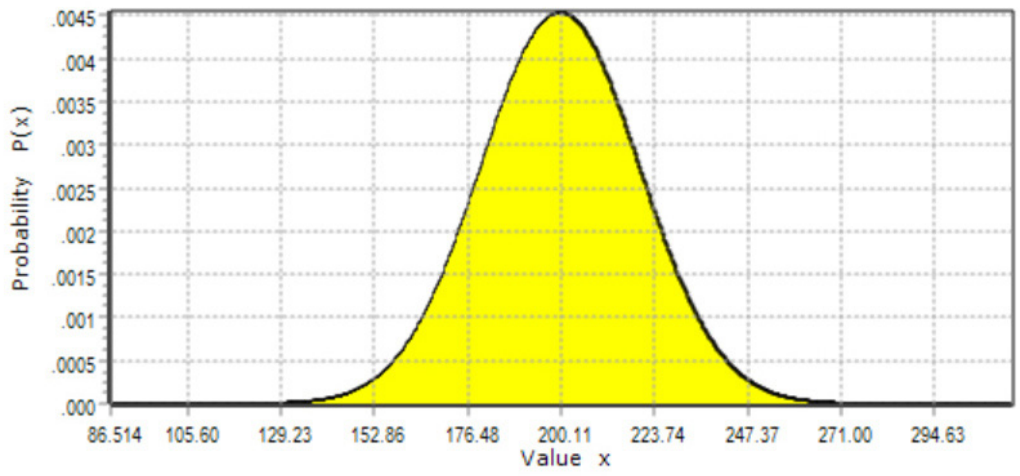


Histogram diskrétní náhodné proměnné (graf pravděpodobnostní funkce) - hod hrací kostkou



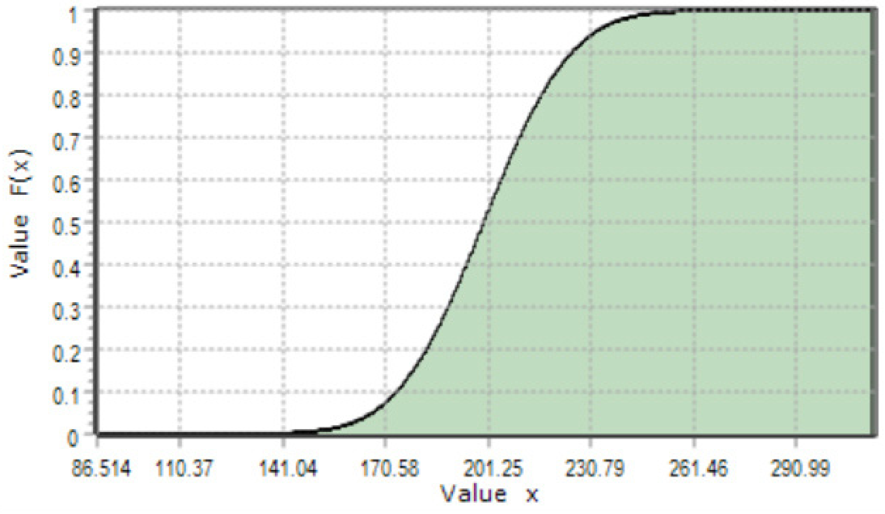
Odpovídající **distribuční funkce**

Distribuční funkce / Histogram



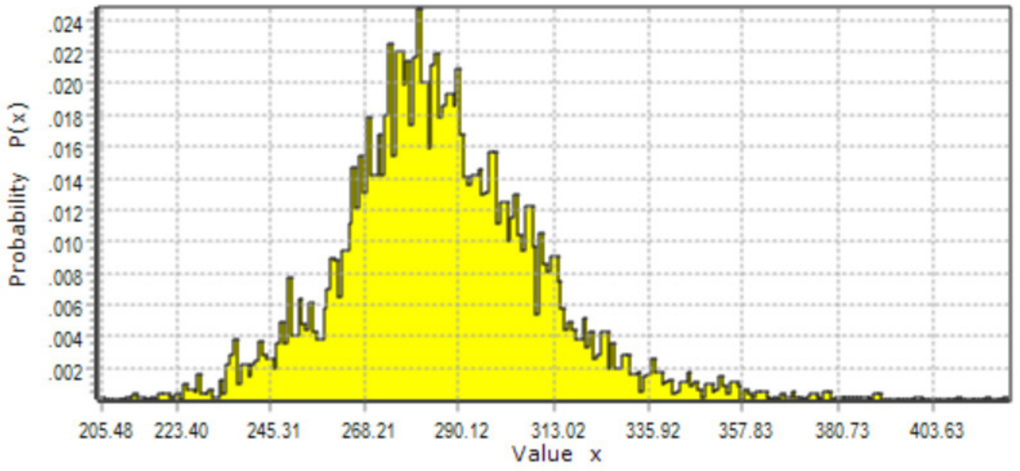
Histogram diskretizované **spojité náhodné proměnné** s parametrickým rozdělením pravděpodobnosti

Normální rozdělení pravděpodobnosti, 1000 intervalů



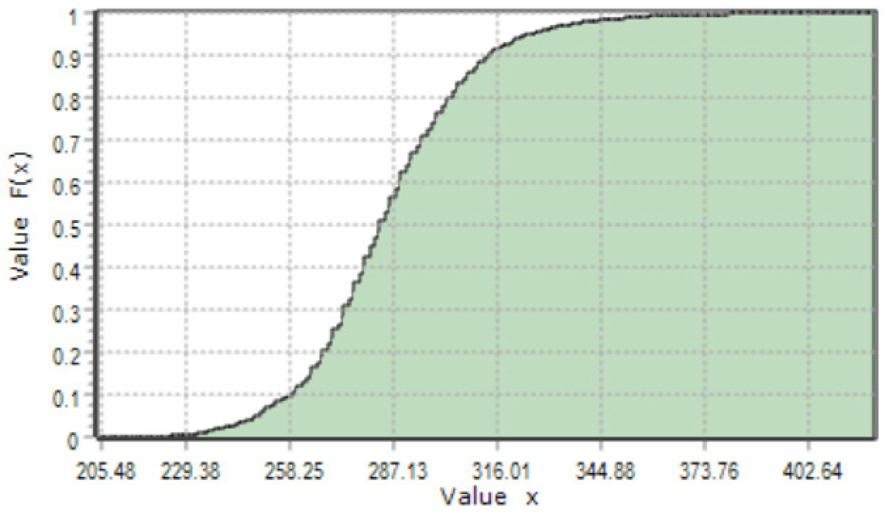
Odpovídající **distribuční funkce**

Distribuční funkce / Histogram



Histogram diskretizované **spojité náhodné proměnné** s neparametrickým (empirickým) rozdělením pravděpodobnosti

Napětí na mezi kluzu oceli, 217 intervalů



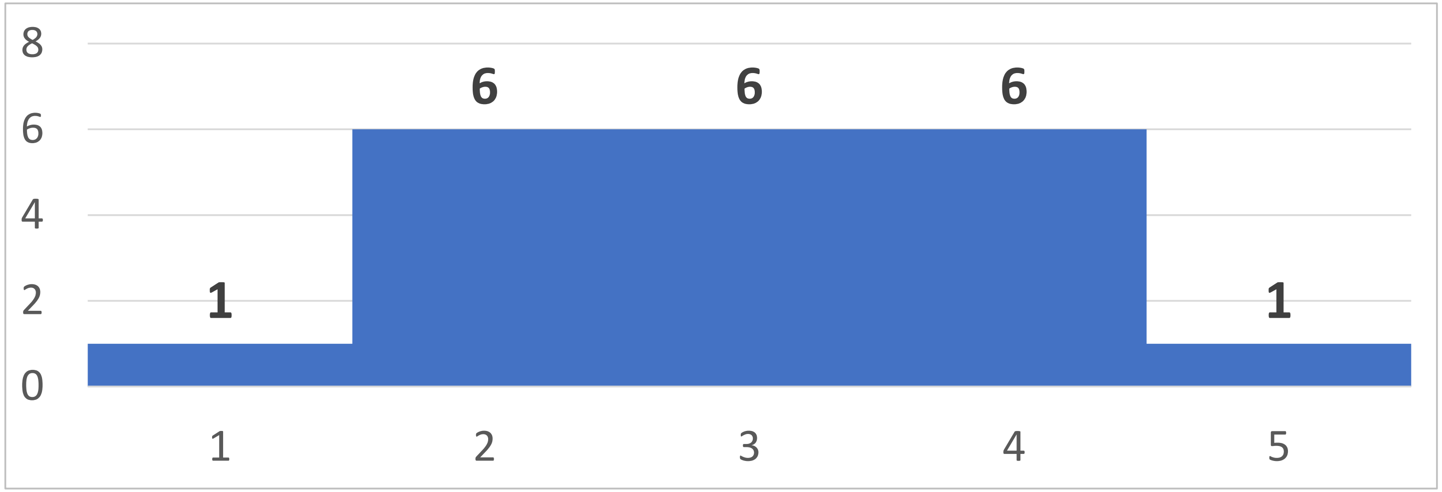
Odpovídající **distribuční funkce**

Příklad 1

Sestrojte **histogram** z naměřených hodnot:

Číslo měření i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Naměřená hodnota x_i	5	2	4	4	1	2	3	2	3	3	4	4	2	3	2	3	4	3	4	2

Výsledný histogram diskrétní náhodné proměnné
(svislá osa – **četnost výskytu naměřených hodnot**)

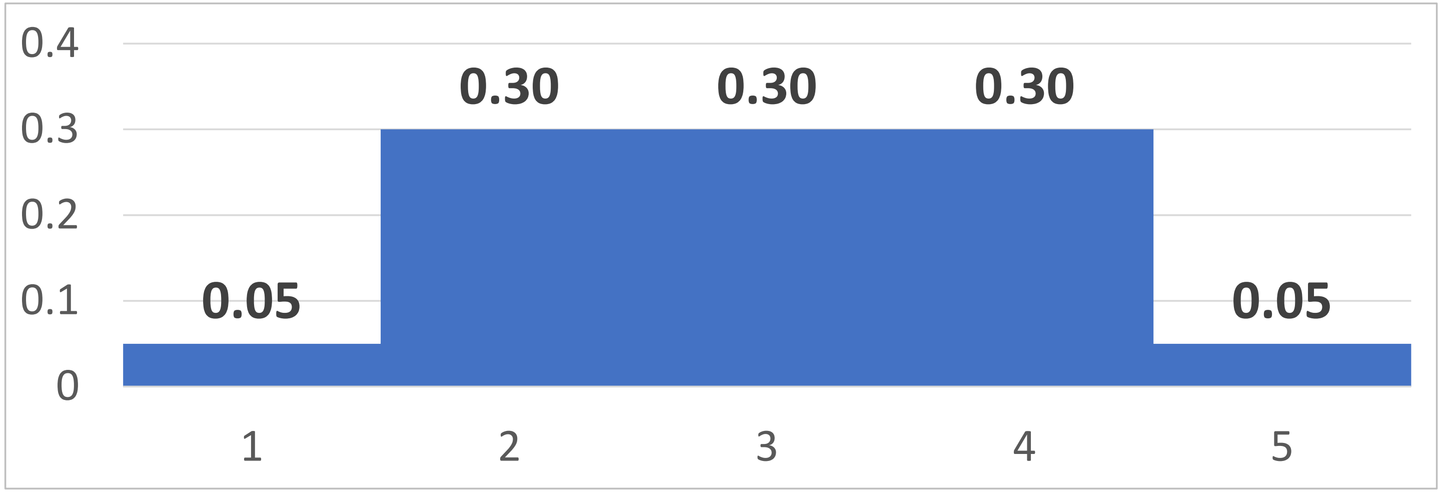


Příklad 1

Výsledný histogram diskrétní náhodné proměnné z příkladu 1 může být vyjádřen rovněž s využitím **pravděpodobností výskytu naměřených hodnot**.

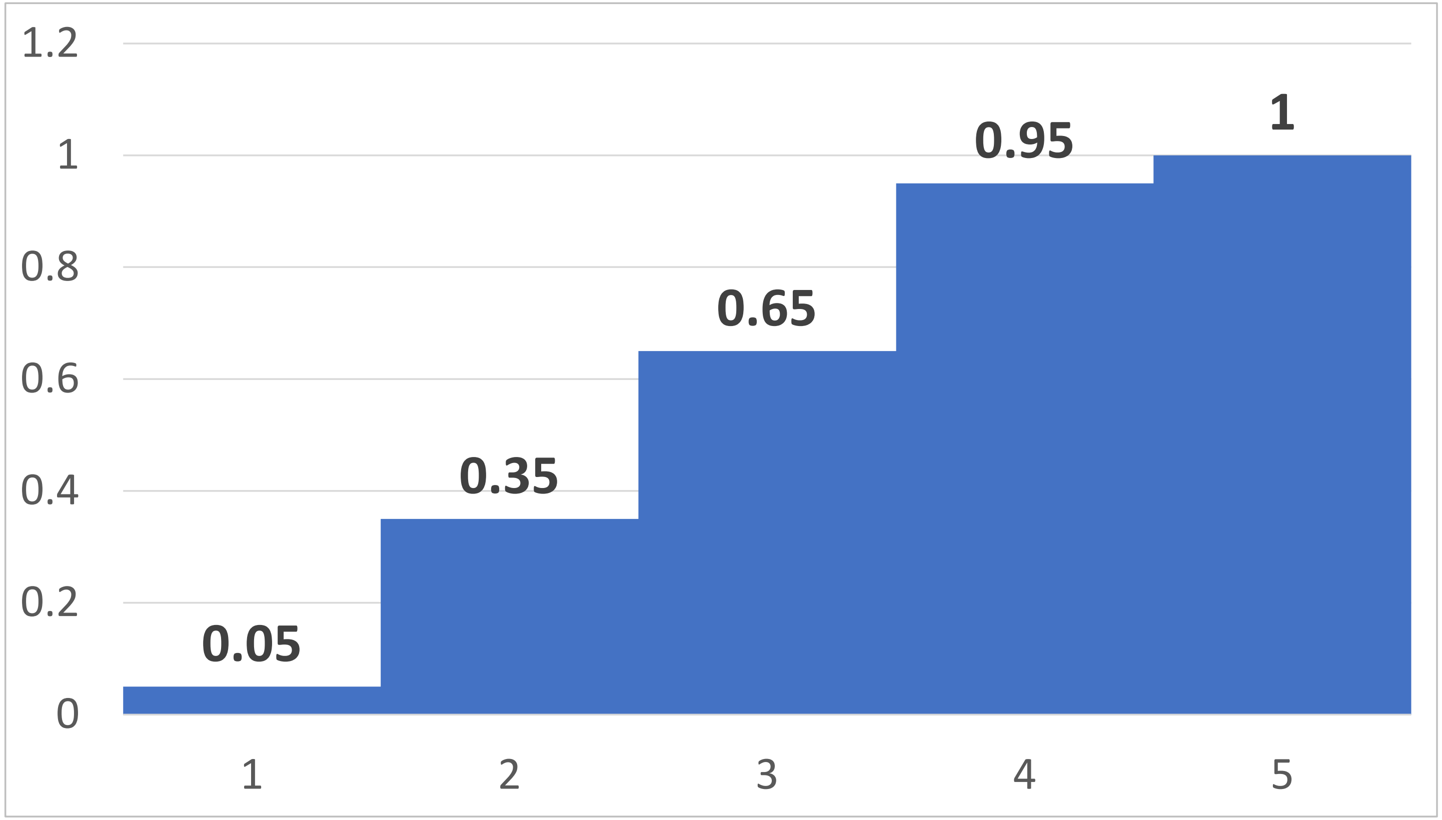
Hodnota x_i náhodné proměnné X	**1**	**2**	**3**	**4**	**5**	**Součet**
Četnost výskytu	1	6	6	6	1	**20**
Pravděpodobnost výskytu	1/20	3/10	3/10	3/10	1/20	**1**

Výsledný histogram diskrétní náhodné proměnné (svislá osa – **pravděpodobnost výskytu naměřených hodnot**)



Příklad 1

Sestrojte graf **distribuční funkce** diskrétní náhodné proměnné z příkladu 1:



Výsledný graf **distribuční funkce** diskrétní náhodné veličiny z příkladu 1

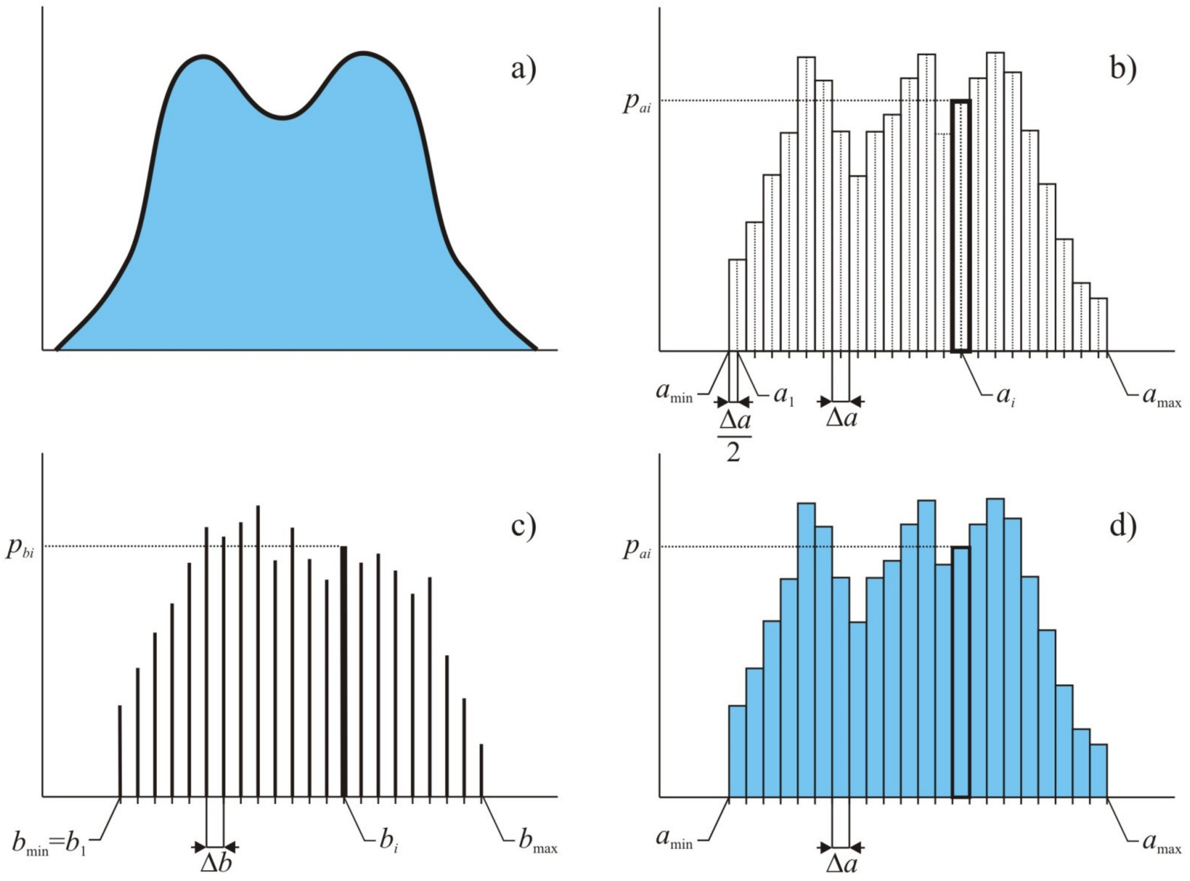
Aproximace rozdělení pravděpodobnosti

a) Původní (**original**) rozdělení pravděpodobnosti spojité náhodné veličiny

b) Diskrétní (**discrete**) aproximace původního rozdělení pravděpodobnosti

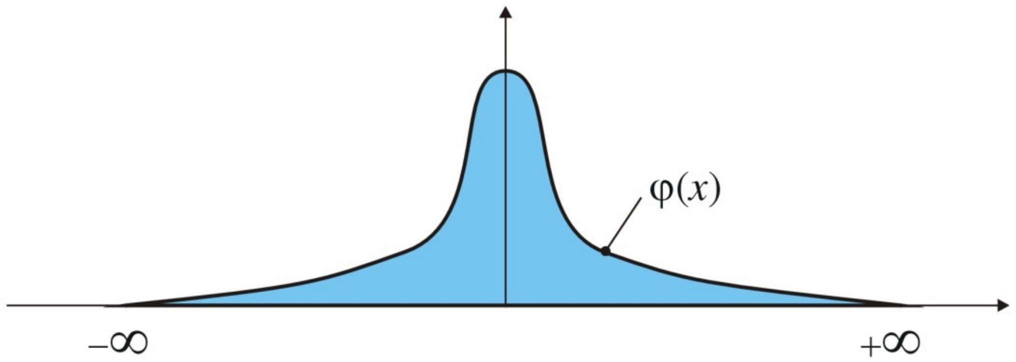
c) Čistě diskrétní (**pure discrete**) aproximace původního rozdělení pravděpodobnosti

d) Po částech rovnoměrná (**piece-wice uniform**) aproximace původního rozdělení pravděpodobnosti

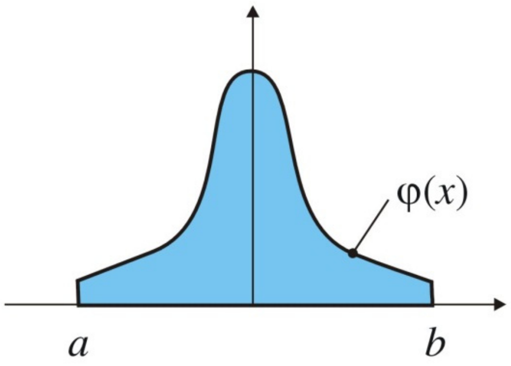


Omezený rozsah rozdělení pravděpodobnosti

Neomezený rozsah rozdělení pravděpodobnosti spojité náhodné veličiny



Omezený rozsah rozdělení pravděpodobnosti spojité náhodné veličiny



Omezený rozsah rozdělení pravděpodobnosti

Omezení rozsahu náhodné veličiny z důvodu počítačové interpretace:

Rozsah datových typů:

Celočíselné typy:

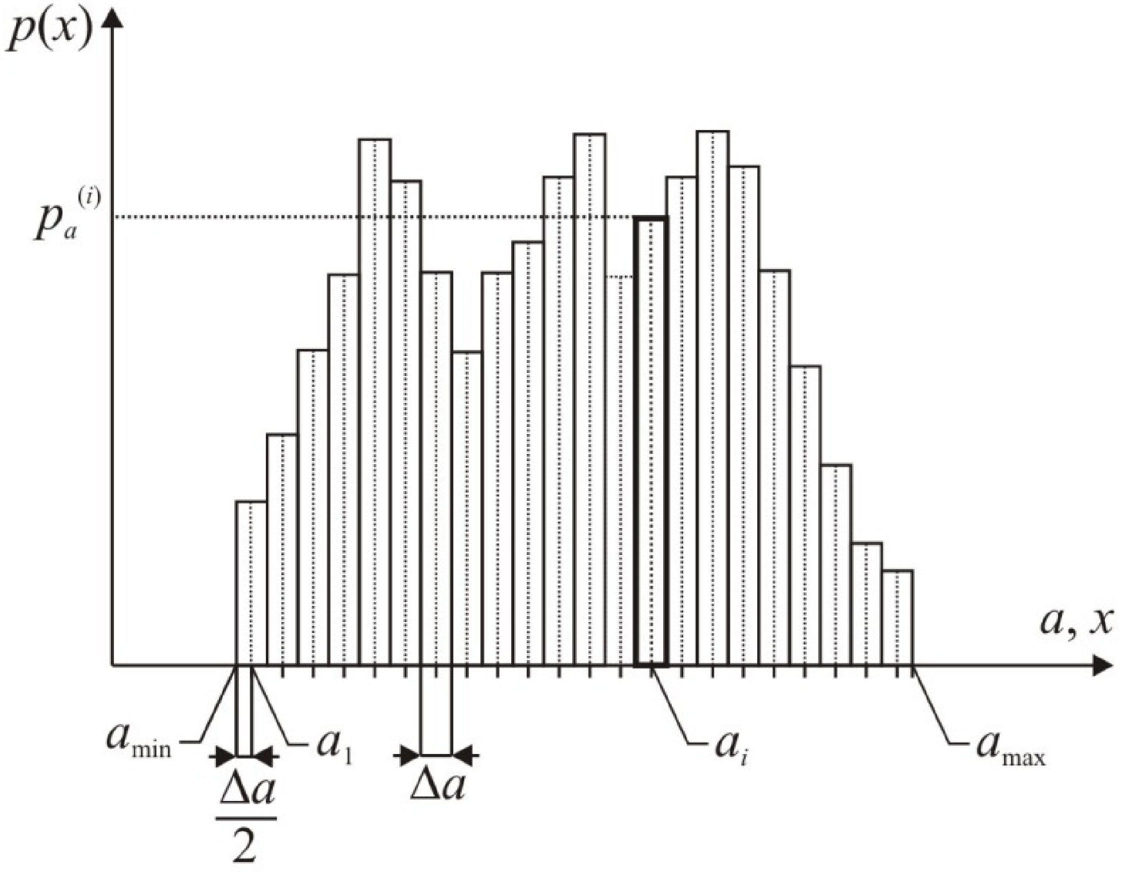
Byte (8 bitů – 1 byte)	0 až 255
Integer (16 bitů – 2 bytů)	-32,768 až +32,767
Word (16 bitů – 2 bytů)	0 až 65,535
LongInt (32 bitů – 4 bytů)	-2,147,483,648 až 2,147,483,647

Číselné typy s plovoucí desetinnou čárkou:

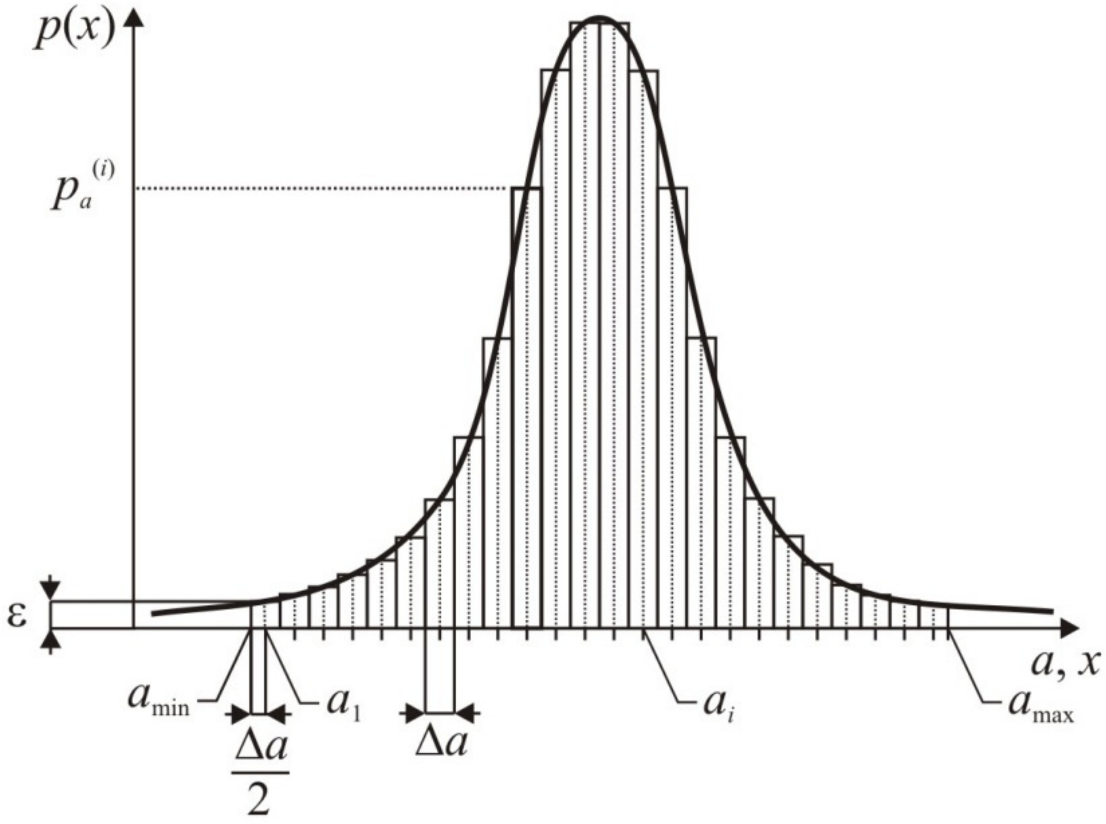
Float (32 bitů = 4 bytů)	$\pm 3.4 \cdot 10^{-38}$ až $3.4 \cdot 10^{38}$
Double (64 bitů = 8 bytů)	$\pm 1.7 \cdot 10^{-308}$ až $1.7 \cdot 10^{308}$
Long double (80 bitů = 10 bytů)	$\pm 3.4 \cdot 10^{-4932}$ až $3.4 \cdot 10^{4932}$

Histogram náhodné proměnné

Histogram spojité náhodné veličiny s diskrétní (**discrete**) aproximací **neparametrického** (empirického) **rozdělení pravděpodobnosti**



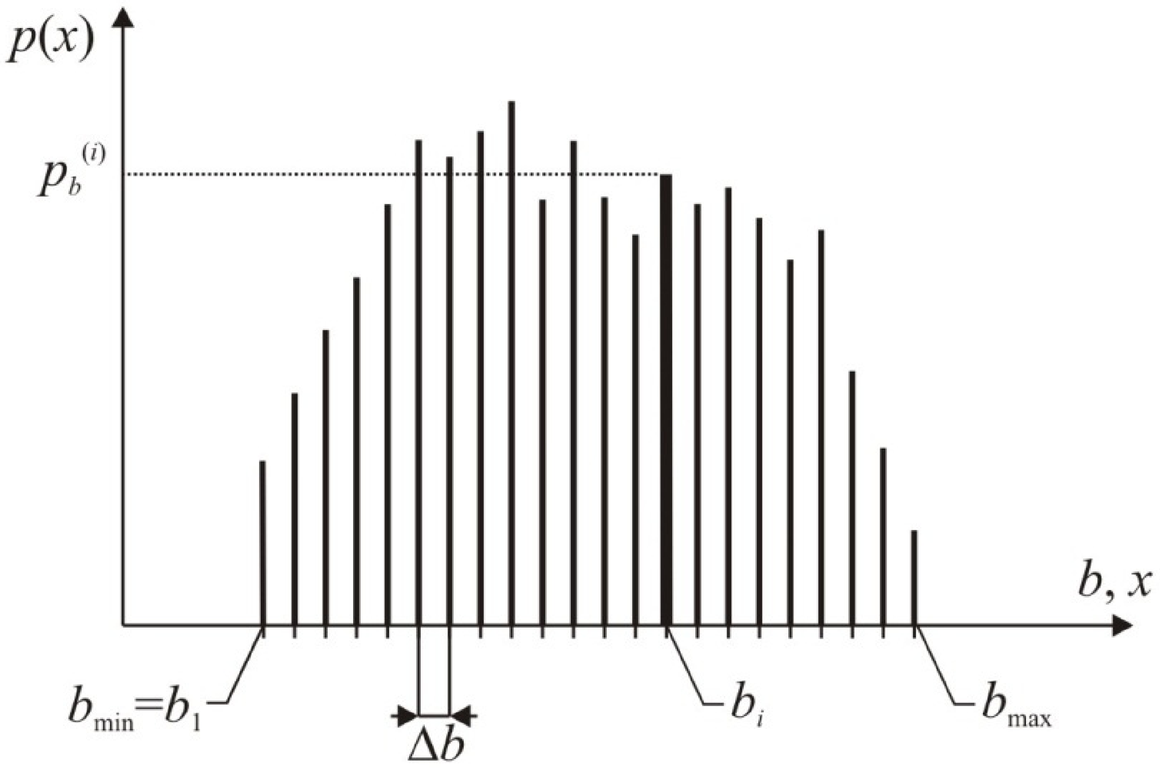
Histogram náhodné proměnné



Histogram spojité náhodné veličiny s diskrétní (**discrete**) aproximací omezeného **parametrického rozdělení pravděpodobnosti**

Histogram náhodné proměnné

Histogram diskrétní náhodné veličiny s čistě diskrétní (**pure discrete**) aproximací **neparametrického** (empirického) **rozdělení pravděpodobnosti**



Structura datového souboru / Definice histogramu

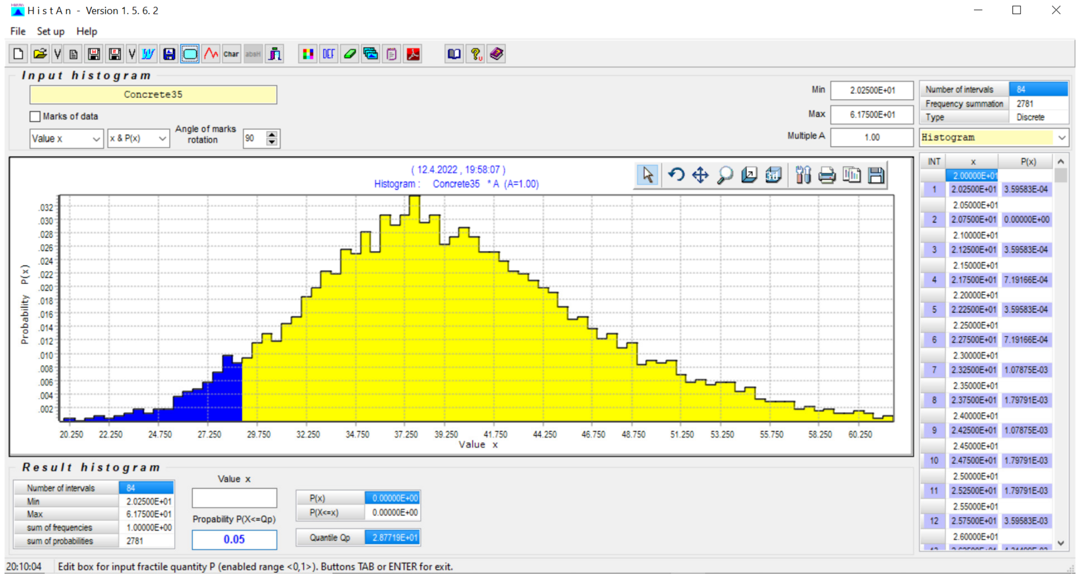
Textový soubor s příponou ***.dis** (distribution), který obsahuje údaje v následujícím tvaru:

```
[Description] (1. oddíl datového souboru)
Identification= volitelný popis datového souboru
Type= Pure Discrete | Discrete | Continuous (typ aproximace)

[Parameters] (2. oddíl datového souboru)
Min= minimální hodnota náhodné proměnné
Max= maximální hodnota náhodné proměnné
Bins= celkový počet intervalů (tříd)
Total= součet četností ve všech intervalech

[Bins] (3. oddíl datového souboru)
četnost v 1. intervalu
četnost ve 2. intervalu
atd. ...
```

Programový nástroj HistAn

Program umožňuje **tvorbu** a **detailní analýzu histogramů** náhodných proměnných:

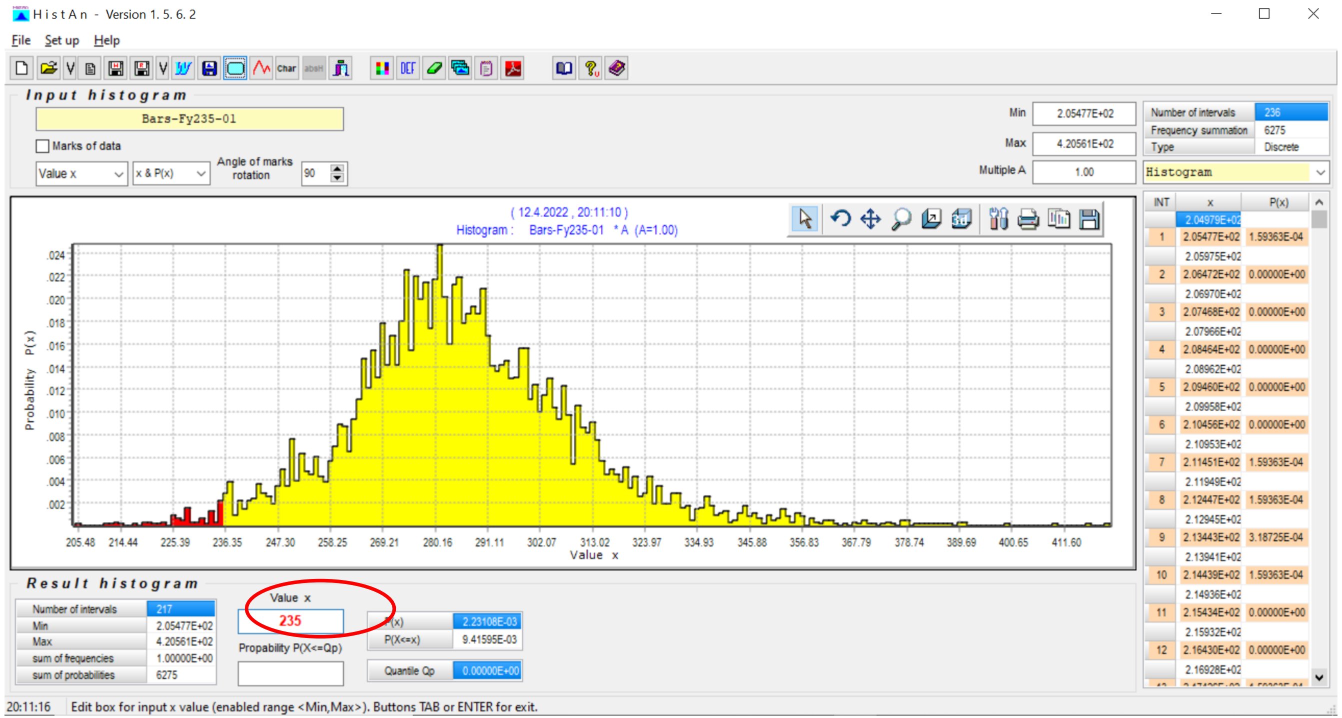
- **Minimální** a **maximální hodnoty** náhodné proměnné
- **Počet tříd** (intervalů) v histogramu a četnosti v nich definované
- **Jednoduché pravděpodobnostní výpočty** s histogramy (např. stanovení hodnoty p-kvantilu)
- Určení **kombinace** (součtu) několika vstupních **histogramů**
- Vytvoření histogramů s **parametrickým rozdělením** pravděpodobnosti
- Zpracování **naměřených** (prvotních) **dat**



Programový nástroj HistAn

Detailní analýza histogramu definujícího **napětí na mezi kluzu oceli S235**:

- Výpočet pravděpodobnosti $P(X \leq x)$ pro zadanou hodnotu x náhodné proměnné X. Pro zadanou hodnotu $x = 235\ \mathrm{MPa}$ je **výsledná pravděpodobnost** $P(X \leq x) =$ $= 9{,}41595 \cdot 10^{-3} = 0{,}94\ \%$.



Pracovní plocha programu **HistAn**

Statistické momenty náhodné proměnné

Praktický význam mají 4 momentové charakteristiky náhodné veličiny:

1. střední hodnota μ

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

2. rozptyl D

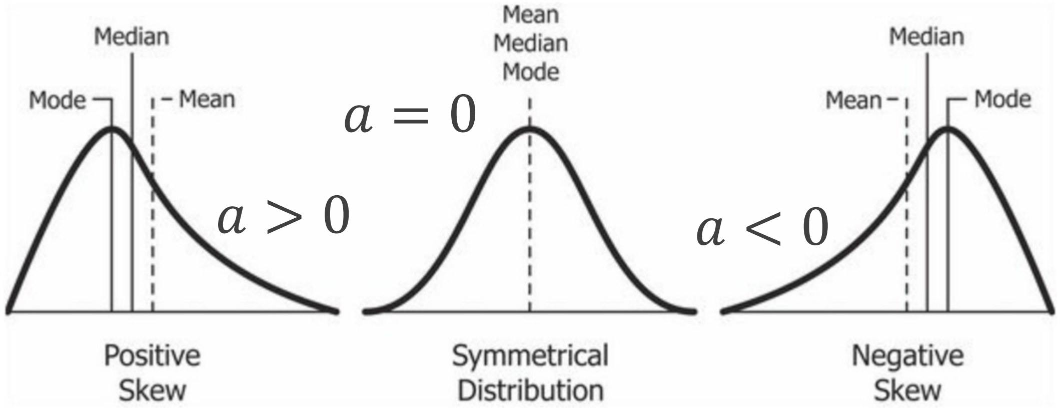
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} = \sqrt{D}$$

σ… **směrodatná odchylka**

3. šikmost (skewness) a

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

(asymetrie rozložení hodnot náhodné proměnné)



4. špičatost (kurtosis) b

(koncentrace hodnot náhodné proměnné v okolí její střední hodnoty)

Kvantil

Kvantil lze obecně chápat jako body, které rozsah spojité náhodné proměnné rozdělují do spojitých intervalů se **stejnou pravděpodobností**.

Běžné kvantily mají speciální jména: **kvartily** (čtyři skupiny dat se stejnou pravděpodobností), **decily** (deset skupin) a **percentily** (100 skupin).

p-**kvantil** x_p pro zadanou hodnotu pravděpodobnosti p je hodnota spojité náhodné proměnné X, pro kterou platí: $F(x_p) = P(X \leq x_p) = p$. Tato hodnota náhodné proměnné rozděluje soubor dat na dvě části: s pravděpodobností výskytu $p \cdot 100\ \%$ a $(1-p) \cdot 100\ \%$. Nazývá se $p \cdot 100\ \%$ kvantil (např. pro $p = 0{,}05$ se jedná o **5% kvantil** s označením $x_{0,05}$).

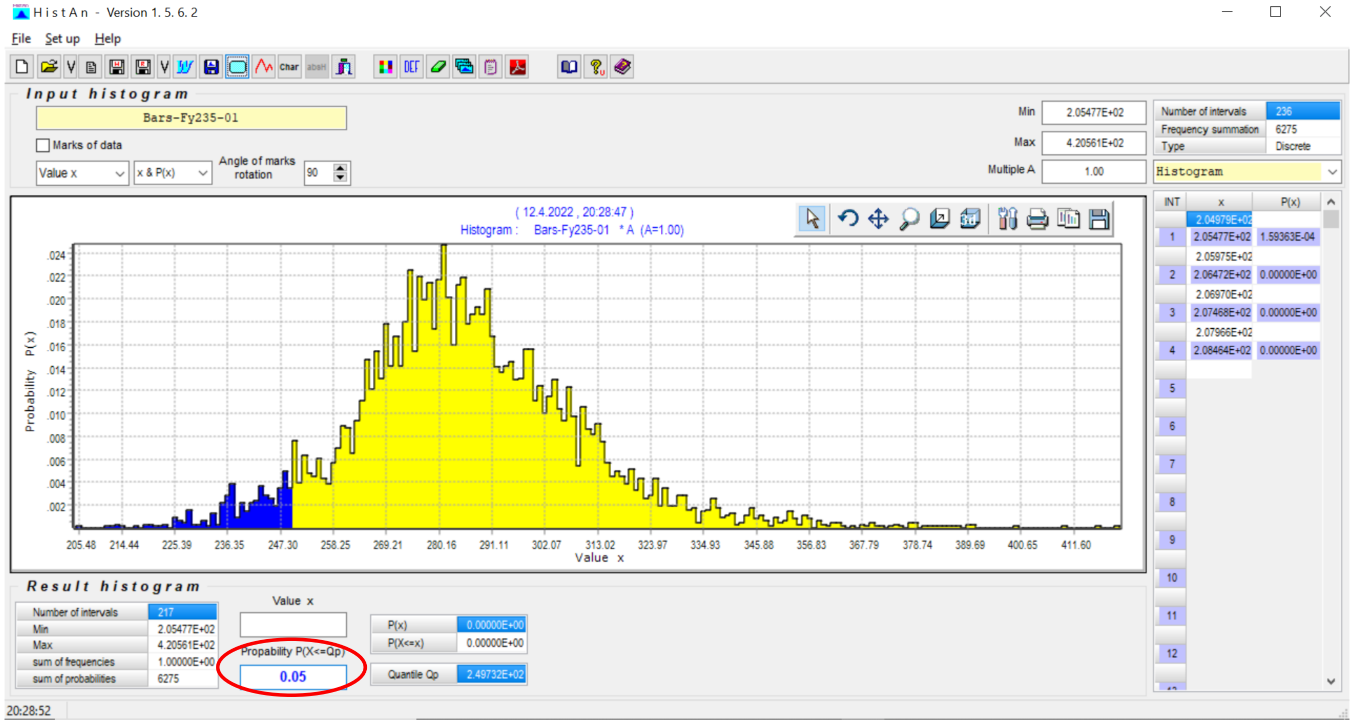
Medián ($x_{0,5}$) = 50% kvantil, rozděluje soubor dat na dvě části - 50% hodnot je menších než medián a 50% hodnot větších než medián (nebo stejných).

Programový nástroj HistAn

Detailní analýza histogramu definujícího **napětí na mezi kluzu oceli S235**:

- Výpočet **pěti-procentního kvantilu** $x_{0,05}$ pro zadanou pravděpodobnost $p = 0{,}05 = 5\ \%$ (platí $P(X \leq x_{0,05}) = 0{,}05$). **Výsledný kvantil** $x_{0,05} = 249{,}732\ \mathrm{MPa}$.

Poznámka: Soubor naměřených hodnot lze rozdělit do 2 skupin: 5% vzorků má hodnotu napětí na mezi kluzu $f_y \leq 249{,}732$ MPa (modrá část histogramu), 95% vzorků naopak $f_y > 249{,}732$ MPa (žlutá část histogramu).



Pracovní plocha programu **HistAn**