Praktická statistika

Petr Ponížil

Eva Kutálková

Zápis výsledků měření

Předpokládejme, že známe hodnotu napětí U = 238,9 V i její chybu 3,3 V.

Hodnotu veličiny zapíšeme na tolik míst, aby až poslední bylo zasaženo chybou.
Chybu píšeme na jednu platnou číslici a zokrouhlujeme nahoru.
Pouze, má-li chyba jako první číslici jedničku, uvádíme chybu na dvě místa.

U = (239 ± 4) V

Zapisujeme-li hodnotu z nějakých důvodů bez chyby, je třeba zapsat ji na tolik míst, aby pouze poslední mohlo být zasaženo chybou.

Zápis výsledků měření

Správně:

$U = (239 \pm 4)$ V

$U = (238{,}9 \pm 1{,}2)$ V

$U = (0{,}239 \pm 0{,}004)$ V

$U = (239 \pm 4).10^3$ V nebo $U = (239 \pm 4)$ kV

$c = 299\ 792\ 458$ m.s^{-1}

Špatně:

$U = (238{,}8 \pm 4)$ V

$U = (238{,}8 \pm 4{,}2)$ V

$U = (238 \pm 0{,}2)$ V

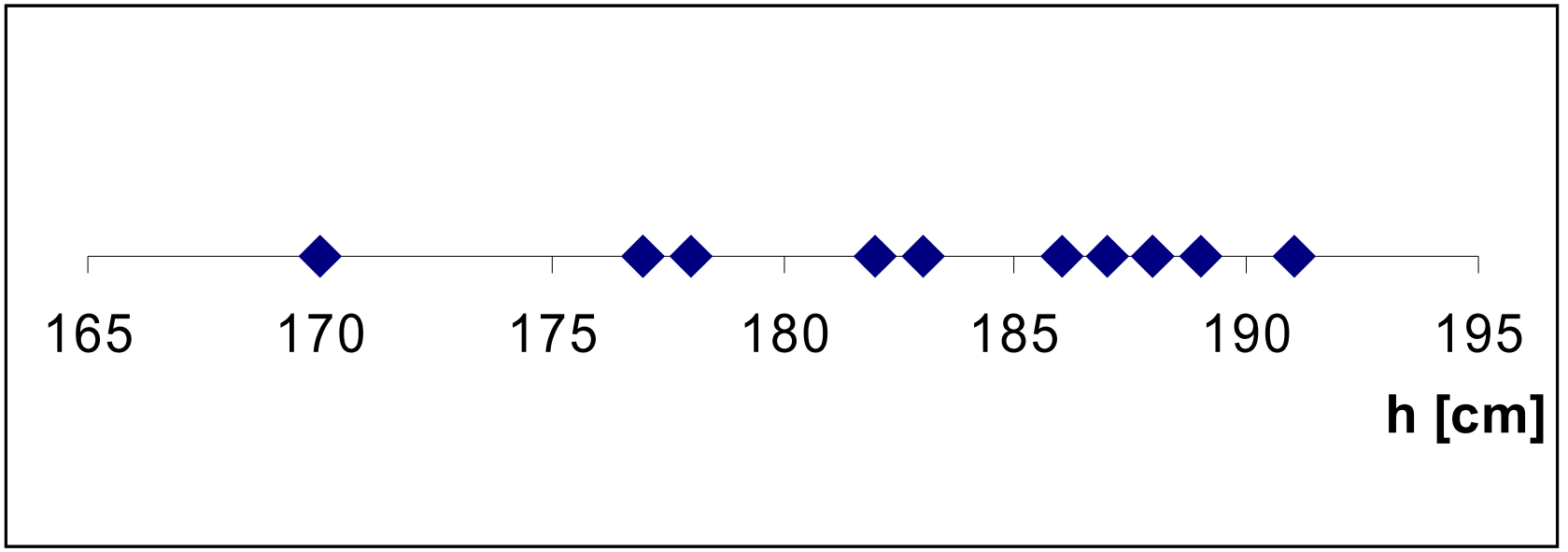
$U = (239\ 000 \pm 4\ 000)$ mV

$c = 300\ 000\ 000$ m.s^{-1}

Neroztříděná data

Výšky 12 studentů FT [cm]:
177 170 182 183 186 188 191 177 189 178 187 188

Diagram rozptýlení



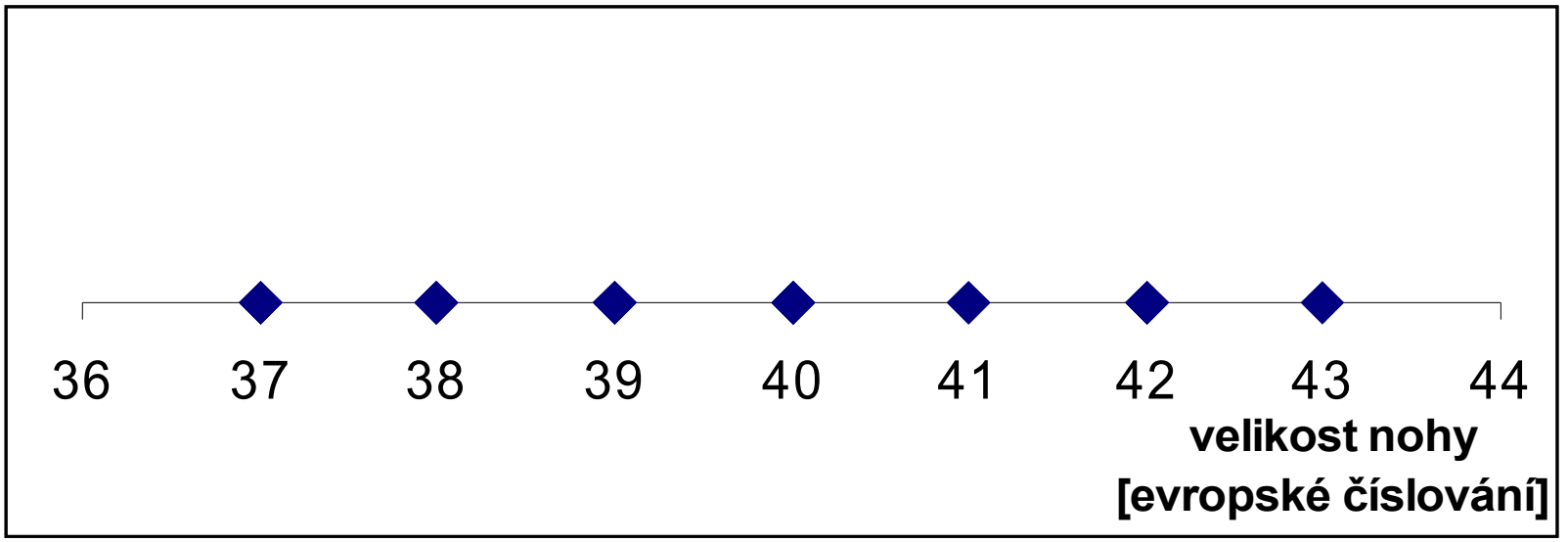
Používá se pro malý počet hodnot (do 30).

Bodové rozdělení četnosti

Velikost nohy 32 studentek FT [evropské číslování]

40 38 40 42 40 37 40 39 37 42 39 38 39 38 41 38
42 39 42 43 40 39 39 37 39 38 40 39 37 42 40 41

Diagram rozptýlení:



- tudy cesta nevede.

Bodové rozdělení četnosti

Velikost nohy 32 studentek FT [evropské číslování]

40 38 40 42 40 37 40 39 37 42 39 38 39 38 41 38
42 39 42 43 40 39 39 37 39 38 40 39 37 42 40 41

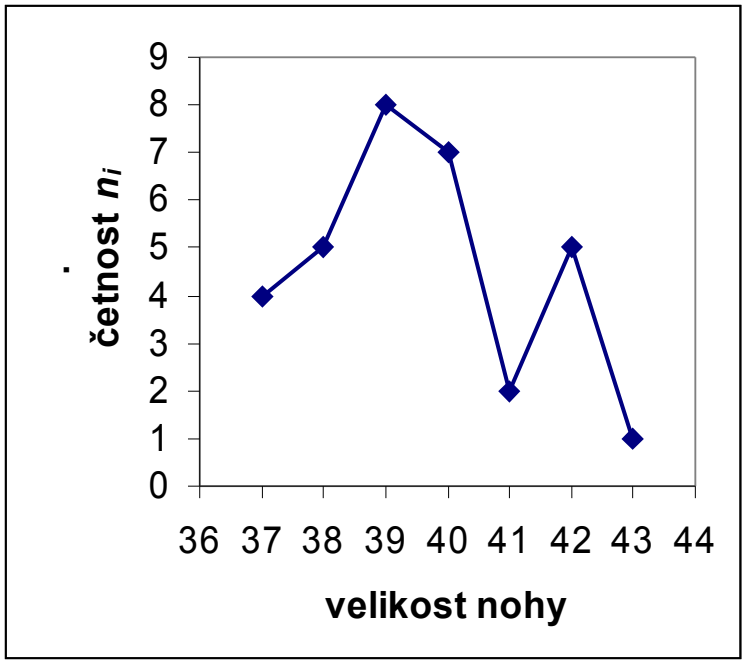
Hodnota znaku x_i	*Absolutní četnost n_i*	*Relativní četnost p_i*	*Absolutní kumulativní četnost N_i*	*Relativní kumulativní četnost P_i*
37	4	0,125	4	0,125
38	5	0,156	9	0,281
39	8	0,250	17	0,531
40	7	0,219	24	0,750
41	2	0,063	26	0,813
42	5	0,156	31	0,969
43	1	0,031	32	1,000
Σ	**32**	**1,000**		

Bodové rozdělení četnosti

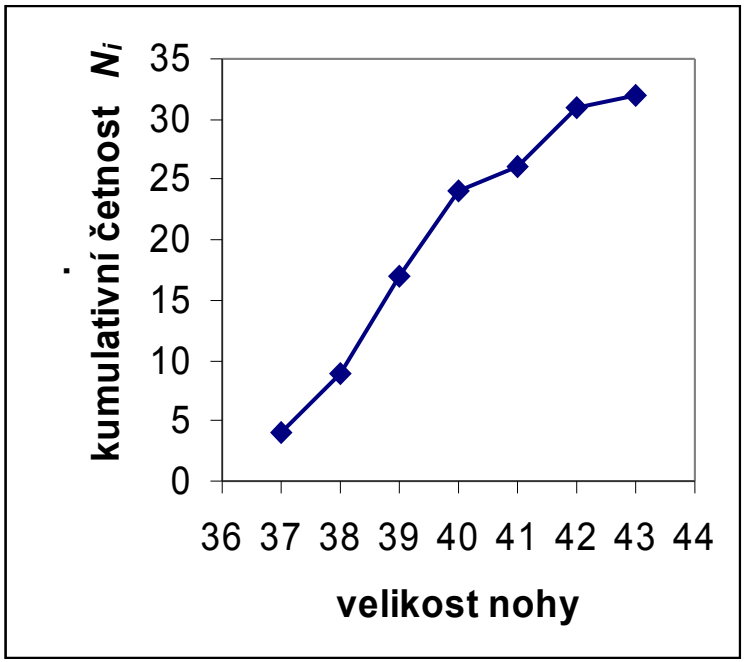
Velikost nohy 32 studentek FT [evropské číslování]

40 38 40 42 40 37 40 39 37 42 39 38 39 38 41 38
42 39 42 43 40 39 39 37 39 38 40 39 37 42 40 41

Polygon četností



Součtová křivka

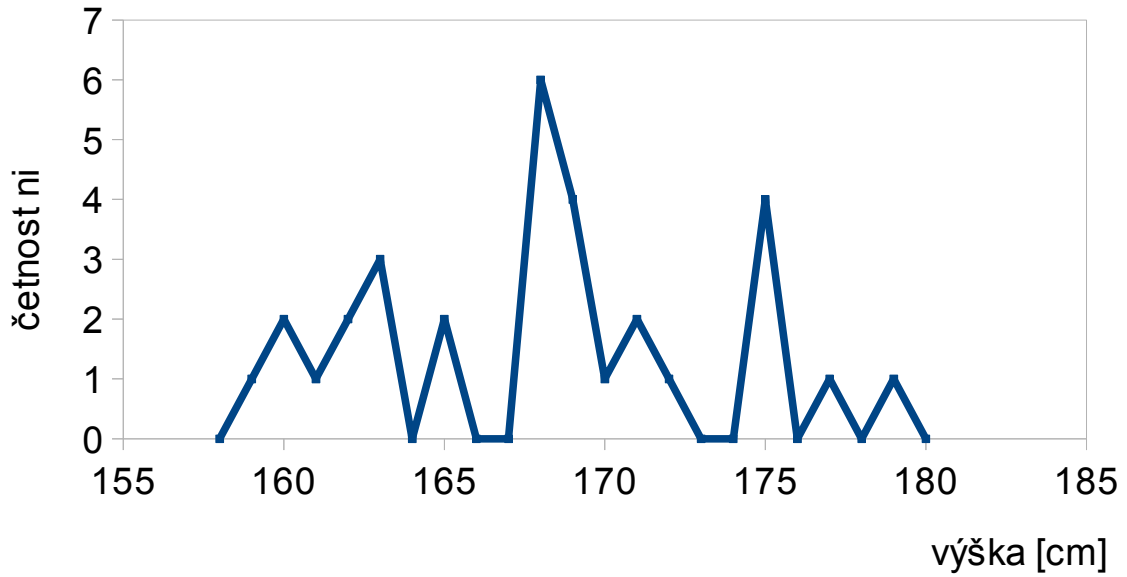


Bodové rozdělení četnosti

Výška 32 studentek FT [cm]:

172	169	161	171	165	169	169	165	168	175	163
168	175	169	168	171	168	163	168	175	159	168
162	163	162	179	160	177	169	160	175	170	171

polygon četností



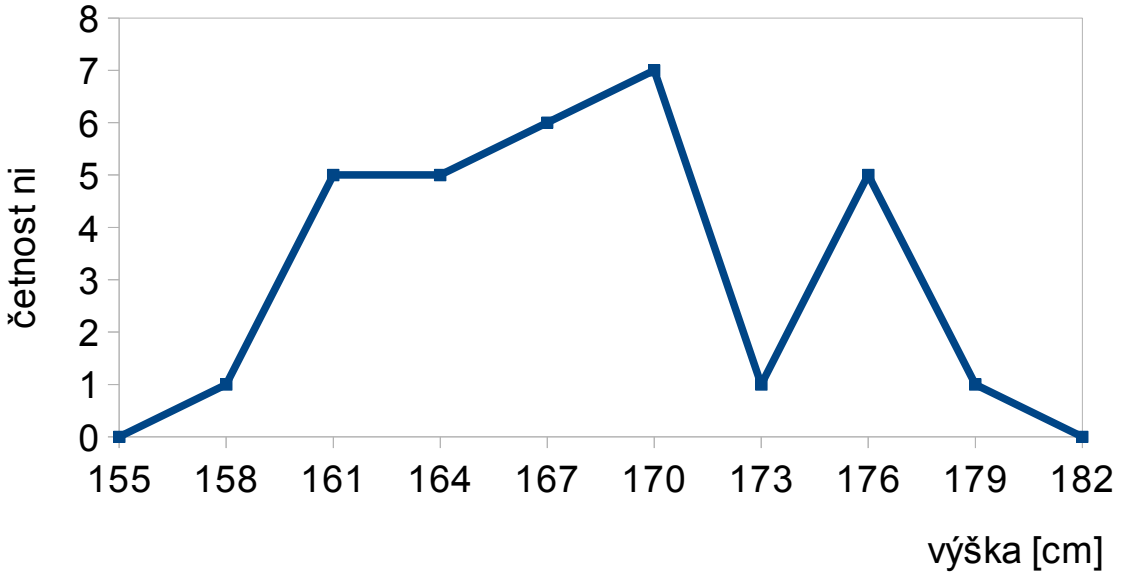
- tudy cesta nevede.

Intervalové rozdělení četnosti

Výška 32 studentek FT [cm]:

172	169	161	171	165	169	169	165	168	175	163
168	175	169	168	171	168	163	168	175	159	168
162	163	162	179	160	177	169	160	175	170	171

polygon četností



Doporučený počet tříd

$k \approx 2{,}46\left(N-1\right)^{0{,}4}$ nebo $k \approx \sqrt{N}$

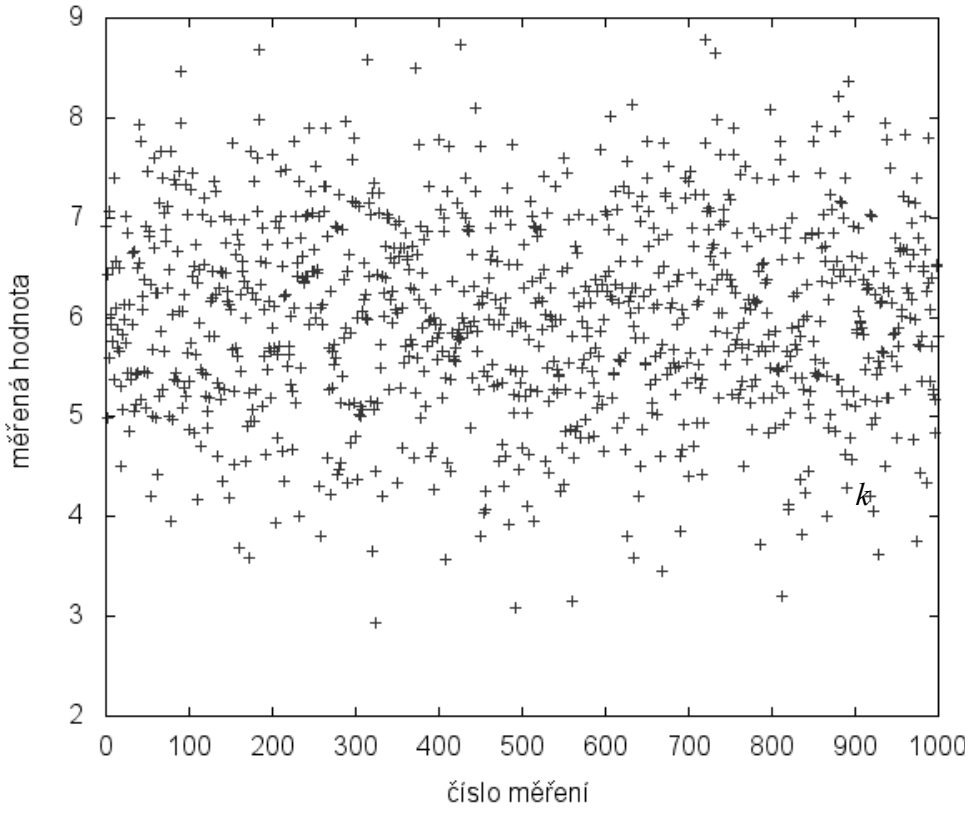
Absolutní a relativní četnost měření

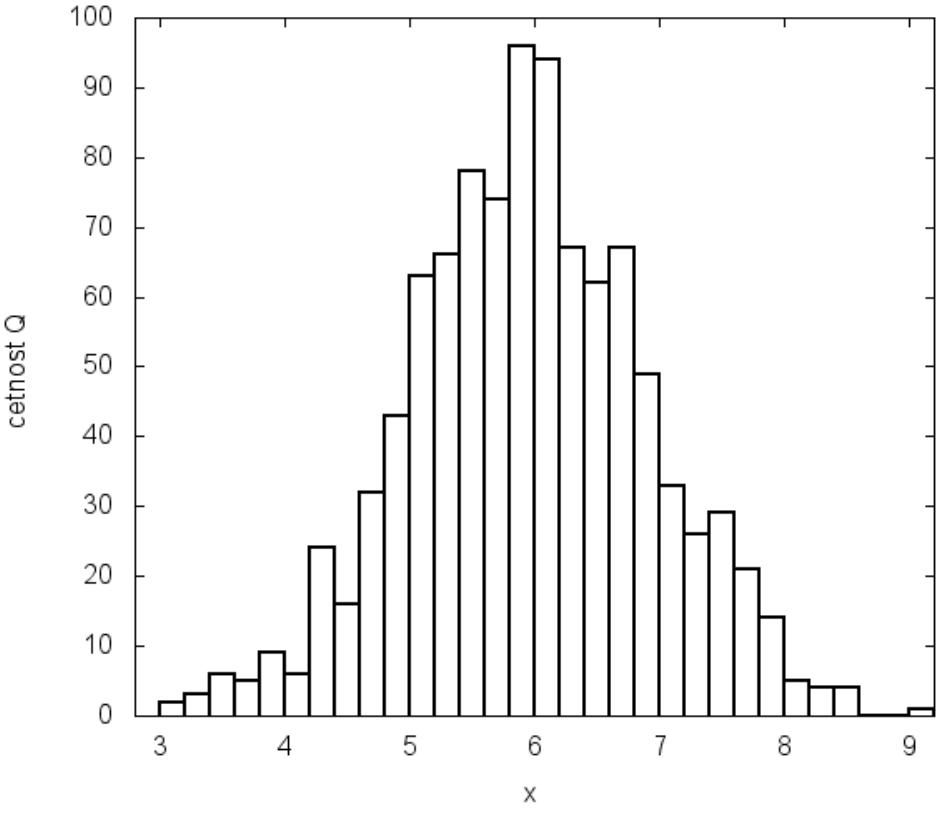
U spojité veličiny je pravděpodobnost výskytu dané hodnoty x nekonečně malá **P(x)** → **0**.
Je proto lepší zjišťovat pravděpodobnost výskytu hodnoty x v nějakém intervalu od x_1 do x_2 → ozn. $P(x_1, x_2)$.

Počet výsledků spadajících do daného intervalu nazýváme (absolutní) četnost Q.
Výhodnější je používat relativní četnost q = Q/N, kde N je počet měření.

Histogram četnosti měření

Četnosti měření graficky znázorníme histogramem.

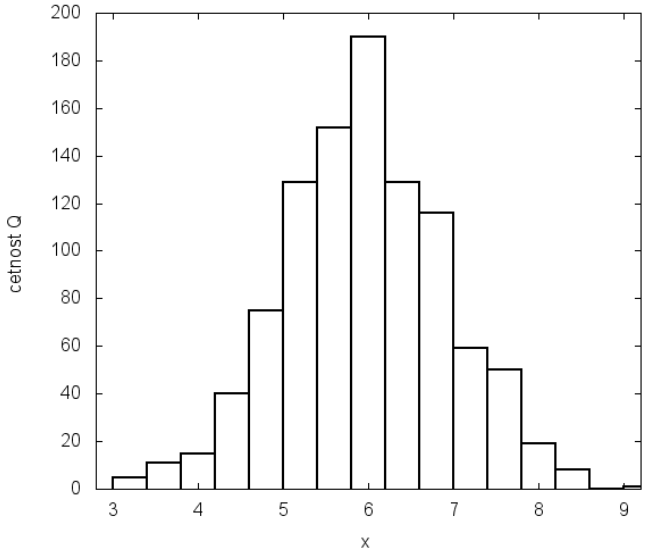


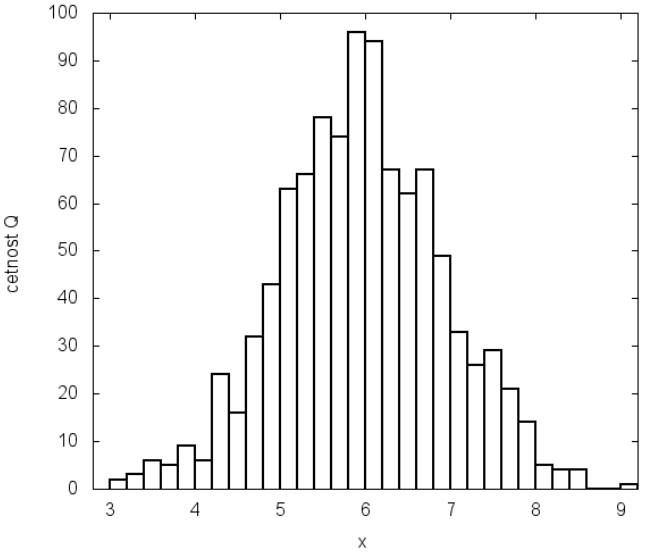


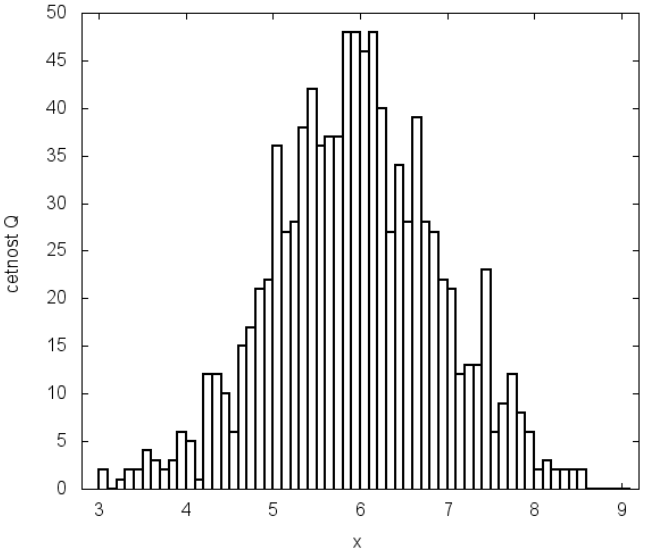
Doporučený počet tříd histogramu

$$k \approx 2{,}46\left(N-1\right)^{0{,}4} \text{ nebo } k \approx \sqrt{N}$$

Počet tříd histogramu



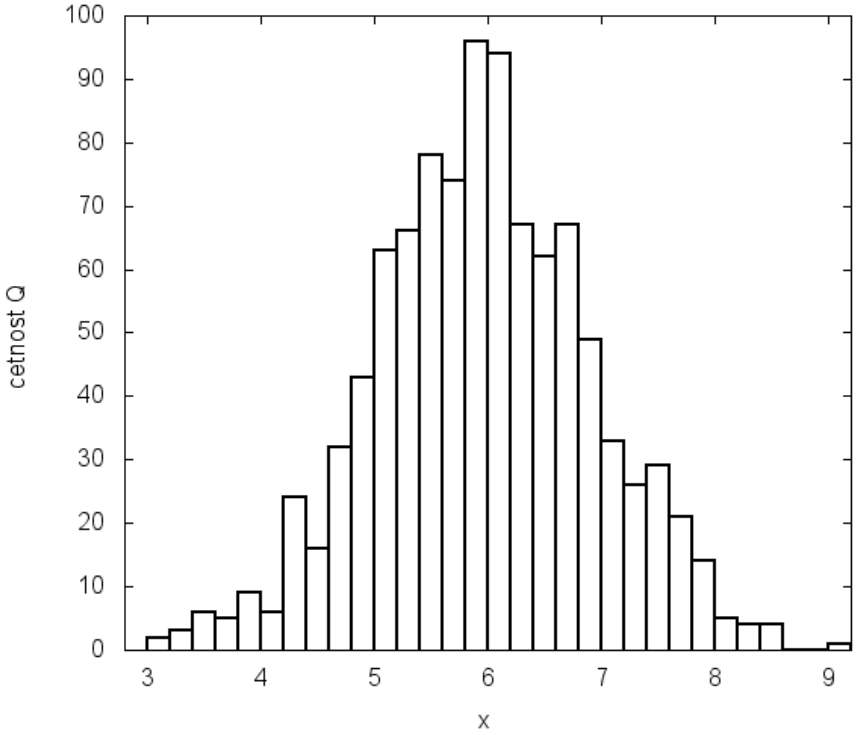


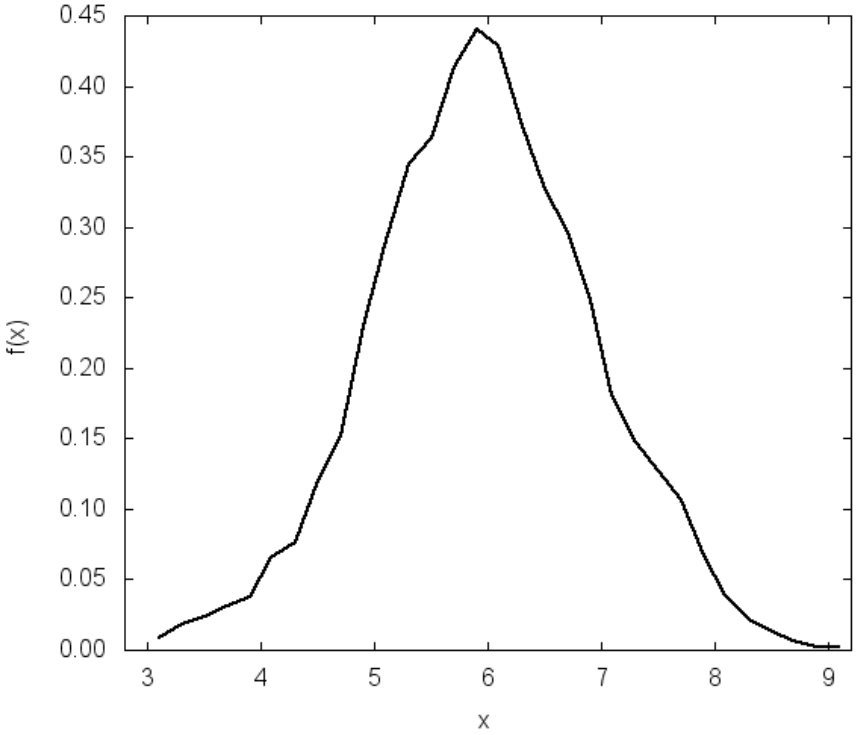


Doporučený počet tříd histogramu

$k \approx 2{,}46\left(N-1\right)^{0{,}4}$ nebo $k \approx \sqrt{N}$

Hustota pravděpodobnosti

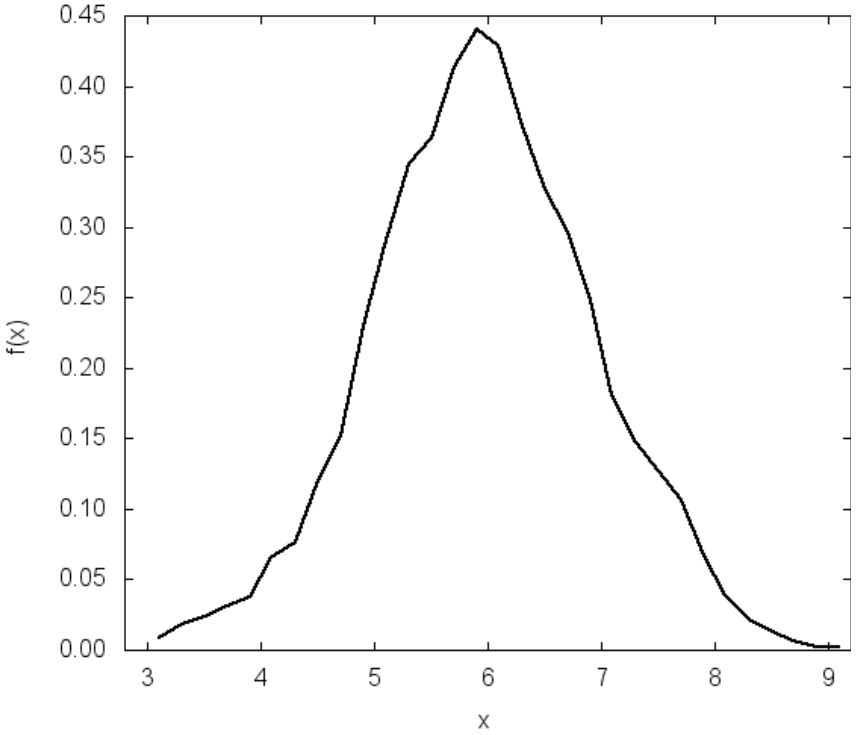


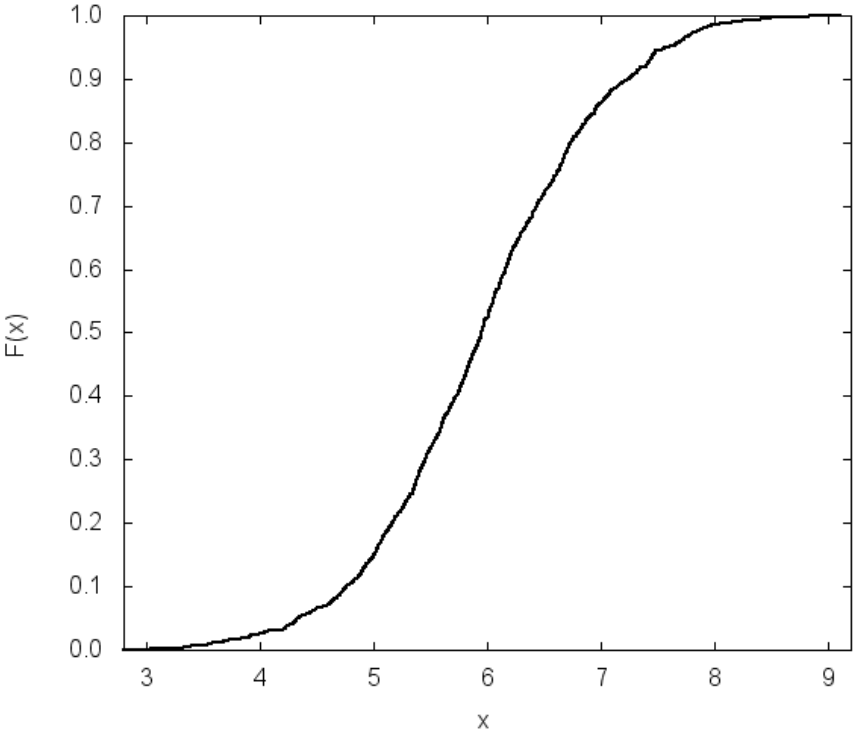


$$f(x) = \lim_{\Delta x \to 0} \frac{P(x, x+\Delta x)}{\Delta x}$$

$$P(a, b) = \int_a^b f(x)\,\mathrm{d}x$$

Distribuční funkce



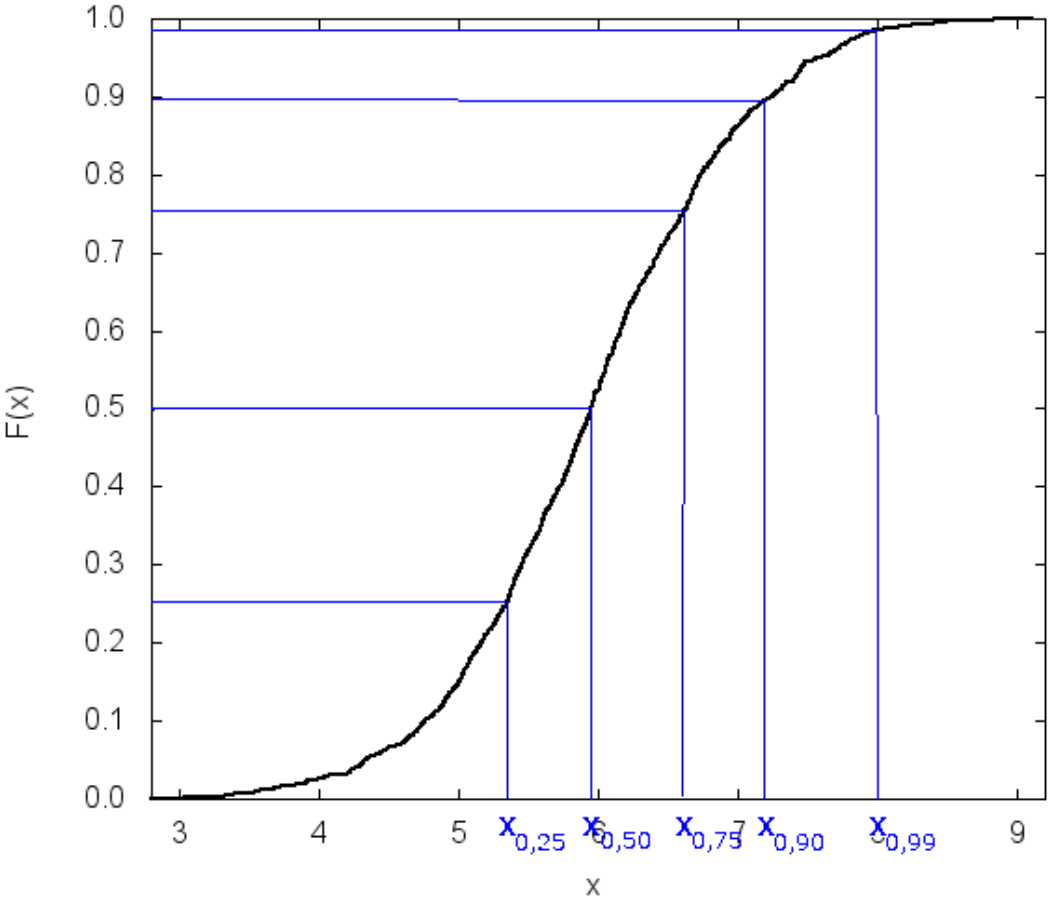


$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

Funkce hustoty pravděpodobnosti nebo distribuční funkce v sobě nesou kompletní informaci o náhodém rozdělení.

Charakteristiky rozdělení - kvantily

Kvantil x_p je hodnota znaku, pro kterou platí, že 100p % jednotek uspořádaného souboru má hodnotu menší nebo rovnu x_p.



25% kvantil $x_{0,25}$ - dolní kvartil,

75% kvantil $x_{0,75}$ - horní kvartil,

50% kvantil $x_{0,50}$ - **medián**,

90% kvantil $x_{0,90}$ - 9. decil,

99% kvantil $x_{0,99}$ - 99. percentil

Aritmetický průměr

Aritmetický průměr **spojité** náhodné veličiny *x*

$$E(x) = \int_{\text{def.obor}} x\, f(x)\, \mathrm{d}\, x$$

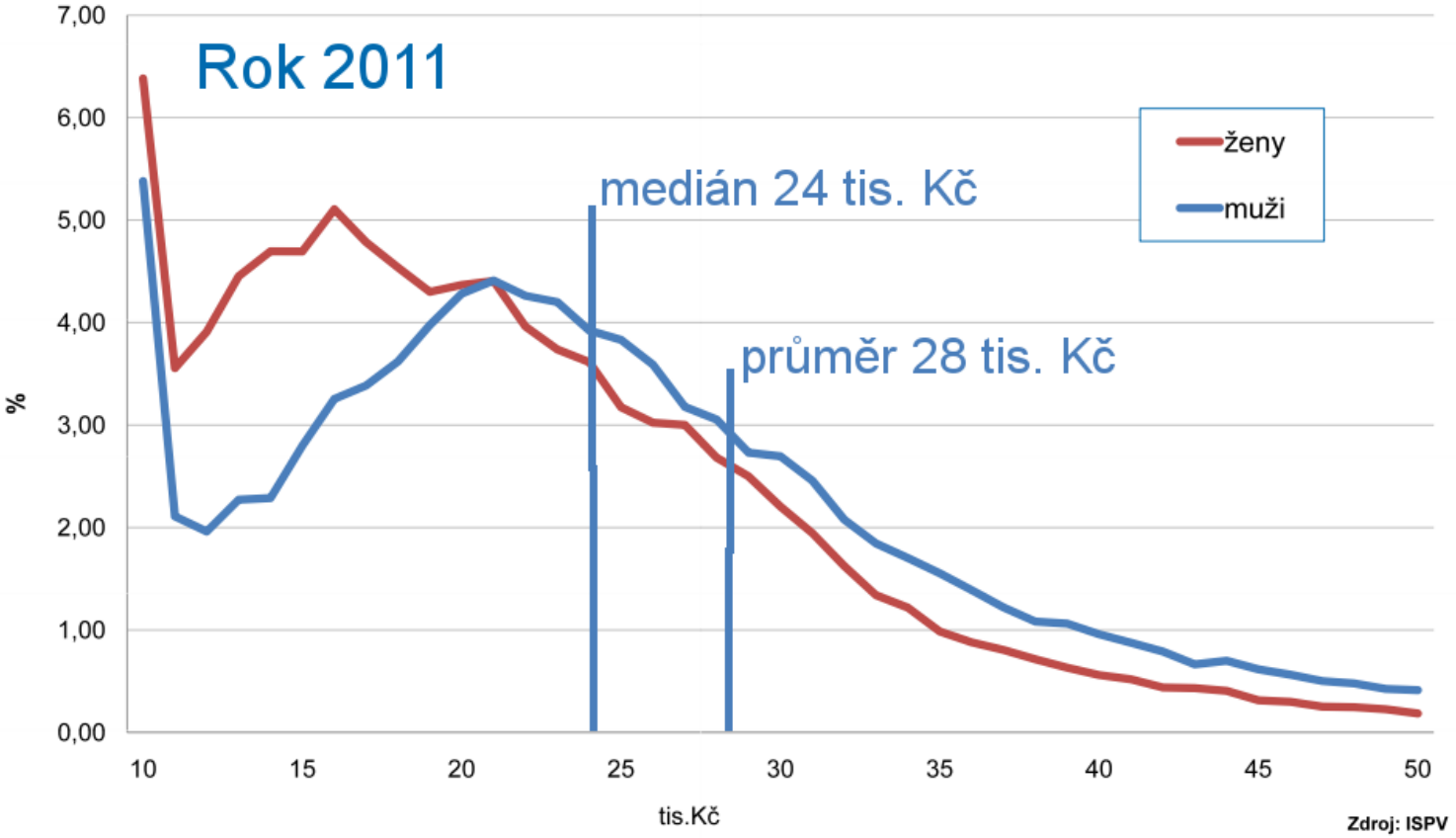
Aritmetický průměr **diskrétní** náhodné veličiny *x*

$$E(x) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Ukážeme dále, že nejlepším odhadem střední hodnoty μ je aritmetický průměr $E(x)$

$$\mu = E(x)$$

Charakteristiky polohy



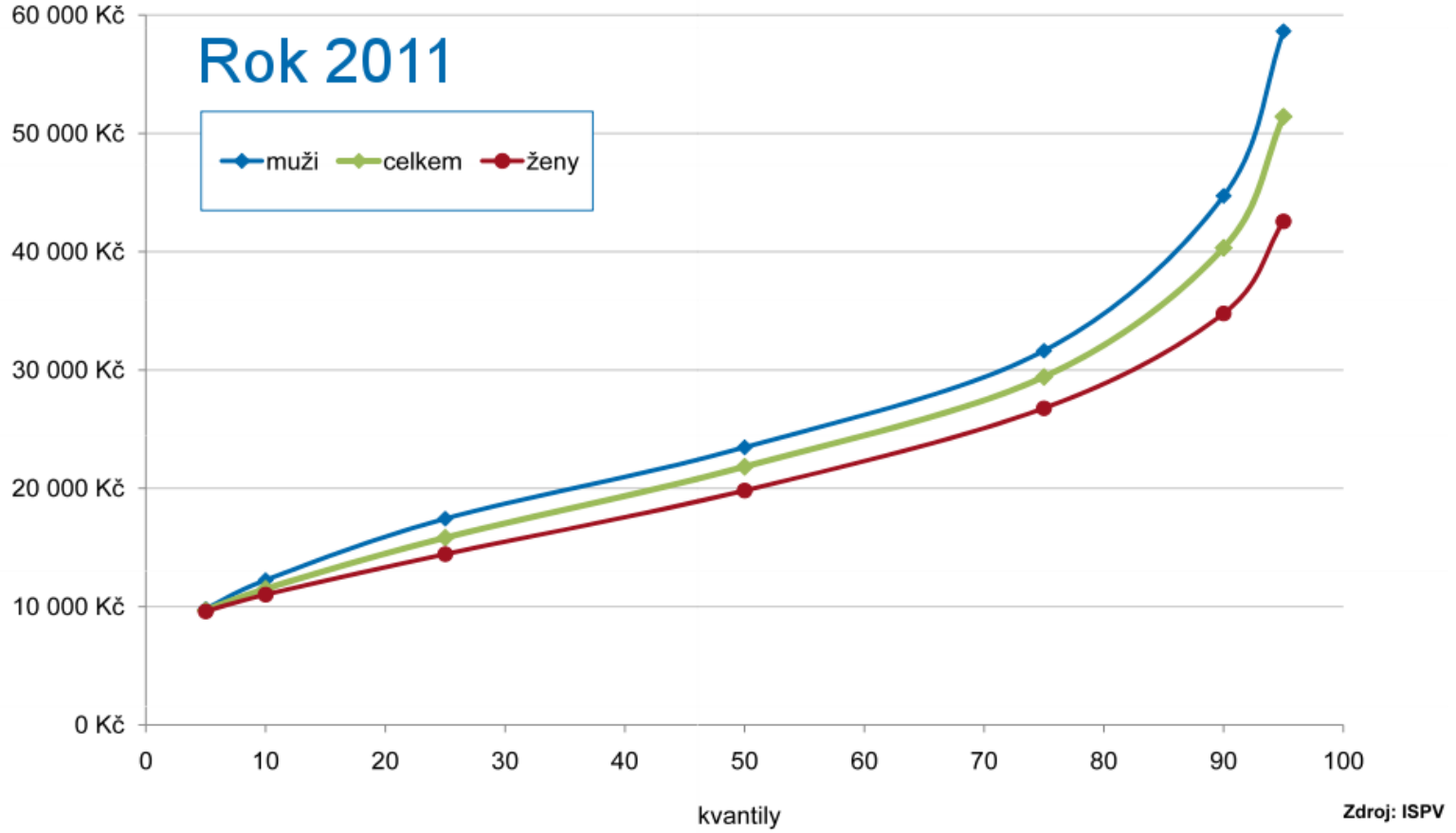
Aritmetický průměr
- součet hodnot vydělený jejich počtem.

Medián
- kvantil $x_{0,50}$
- stejně hodnot pod mediánem jako nad ním.

Modus
- nejpravděpodobnější hodnota.

Nejčastější mzda žen je 16 tisíc, mužů 21 tisíc hrubého

Distribuční funkce mezd



Třetina mužů vydělává méně než prostřední žena,
2/3 žen méně než prostřední muž

Jak je zkonstruovaná distribuční funkce mezd?
Jak odstranit rozdíly?

Charakteristiky variability

Variabilita popisuje rozsah souboru
- rozdíl maximální a minimální hodnoty
- kvartilové rozpětí (rozdíl horního a dolního kvartilu)
- decilové rozpětí (rozdíl devátého a prvního decilu)
- percentilové rozpětí (rozdíl 99. a 1. percentilu)

Charakteristiky variability

Nejčastěji používanou charakteristikou variability (tj. míry odchylky jednotlivých hodnot od střední hodnoty) je **rozptyl** náhodné veličiny *x*.
Pro **spojitou** náhodnou veličinu

$$D(x) = \int_{\text{def.obor}} \{x - E(x)\}^2 f(x)\,\mathrm{d}x$$

Pro **diskrétní** náhodnou veličinu

$$D(x) = \frac{1}{N}\sum_{i=1}^{N} \{x_i - E(x)\}^2$$

Odmocnina z rozptylu se nazývá směrodatná odchylka a značí se σ

$$\sigma = \sqrt{D(x)}$$

Jako charakteristiku variability (nepříliš dobrou) lze použít i minimální a maximální hodnotu.

Charakteristiky koncentrace

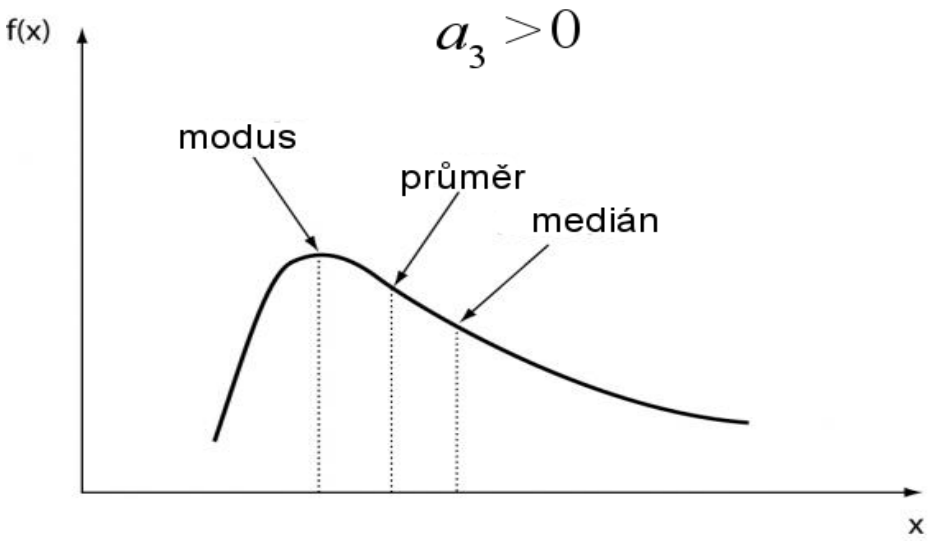
Kromě střední hodnoty charakterizující polohu rozdělení a směrodatné odchylky charakterizující variabilitu (šířku) rozdělení existují i charakteristiky koncentrace informující o „hustotě“ dat.
Budeme používat koeficient šikmosti (**šikmost**) a koeficient špičatosti (**špičatost**).
Tyto koeficienty informují o tvaru rozdělení.
Šikmost o jeho souměrnosti.
Špičatost o koncentraci prostředních hodnot.

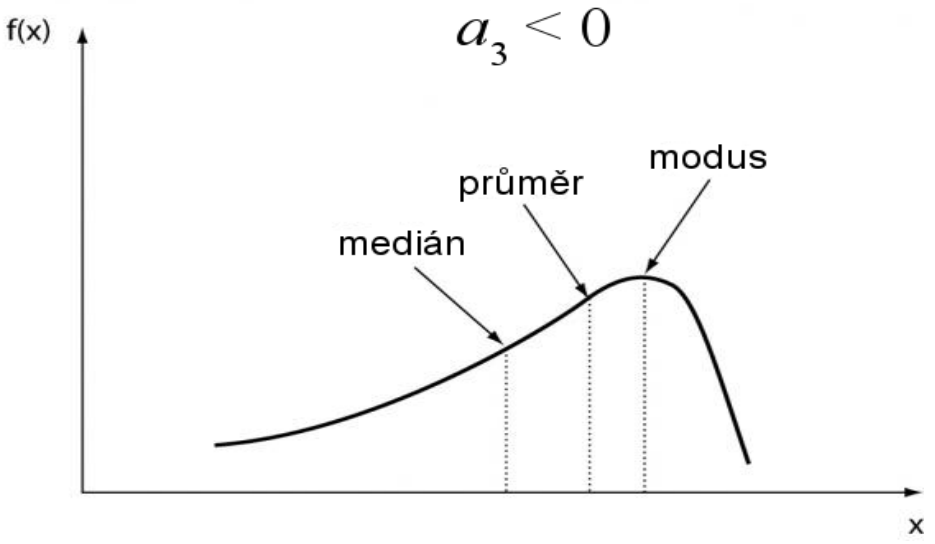
Šikmost

Šikmost je charakteristika rozdělení náhodné veličiny, která popisuje jeho symetrii.

Šikmost a_3 je definována vztahem:

$$a_3 = \frac{\sum_{i=1}^{n} (x_i - \mu)^3}{N \sigma^3}$$





pro $a_3 = 0$ je rozdělení symetrické

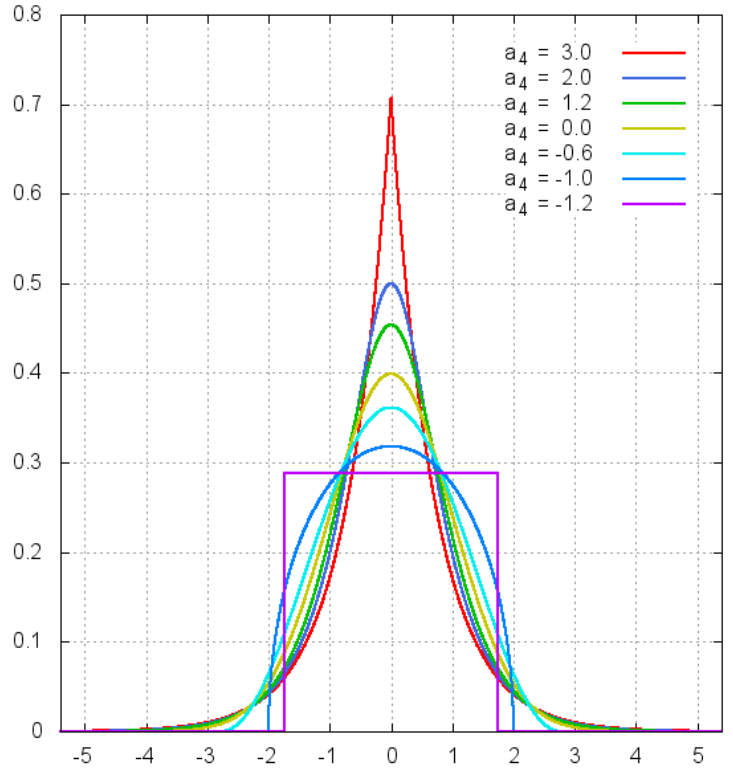
Špičatost

Špičatost (koeficient špičatosti) popisuje koncentraci rozdělení kolem středu.

- normální rozdělení má nulovou špičatost
- při kladné špičatosti je křivka hustoty pravděpodobnosti špičatější než u normálního rozdělení
- při záporné špičatosti je křivka hustoty pravděpodobnosti plošší než u normálního rozdělení

Šikmost a_4 je definována vztahem:

$$a_4 = \frac{\sum_{i=1}^{n} (x_i - \mu)^4}{n\sigma^4} - 3$$



Výpočet šikmosti a špičatosti v Excelu

V Excelu existuje pro výpočet šikmosti funkce a_3^* =SKEW()
a pro výpočet špičatosti funkce a_4^*=KURT().
Bohužel jsou tyto funkce definovány pomocí jiných vzorců:

$$a_3^* = \frac{N}{(N-1)(N-2)} \frac{\sum_{i=1}^{n} (x_i - \mu)^3}{\sigma^3}$$

$$a_4^* = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^{n} (x_i - \mu)^4}{\sigma^4} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Hodnoty lze přepočítat podle vztahů:

$$a_3 = \frac{(N-2)}{\sqrt{(N(N-1))}} a_3^*$$

$$a_4 = \frac{(N-2)(N-3)}{N^2-1} a_4^* - \frac{6}{N+1}$$

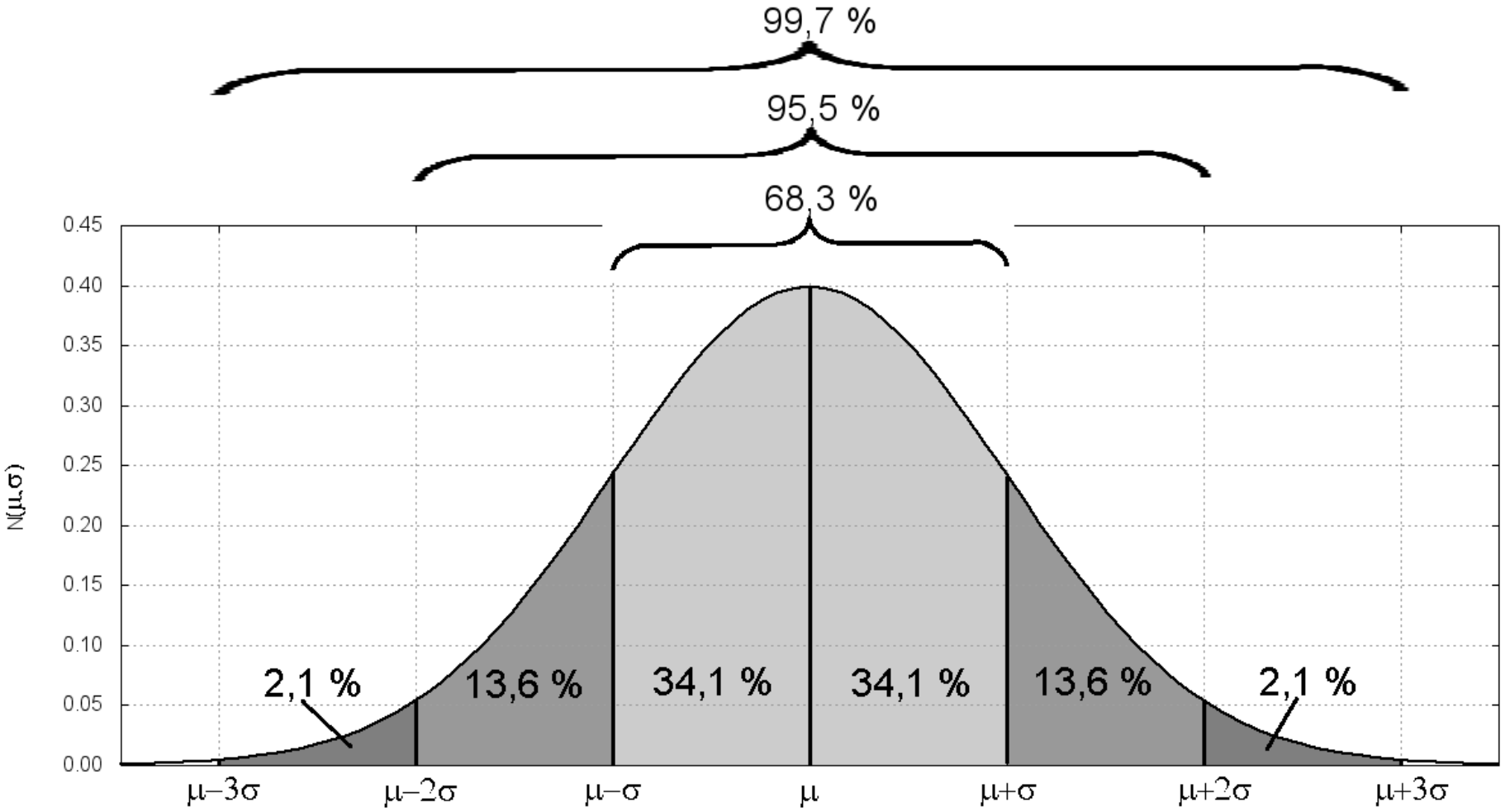
Normální (Gaussovo) rozdělení

Normální (Gaussovo) rozdělení popisuje vlastnosti náhodné spojité veličiny, která vzniká složením různých náhodných vlivů, které jsou navzájem nezávislé, kterých je velký počet a každý z nich ovlivňuje výsledek jen málo.

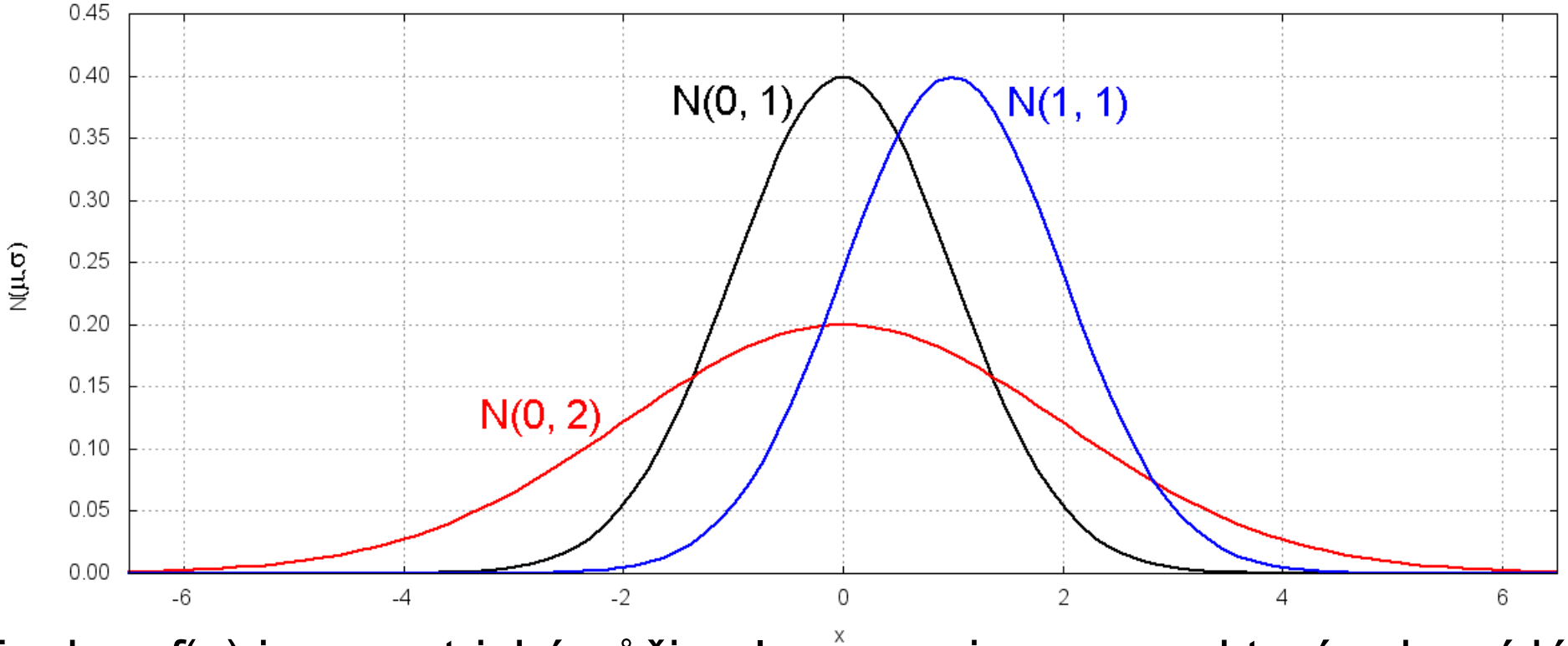
Hustota pravděpodobnosti má v tomto případě tvar

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Normální (Gaussovo) rozdělení



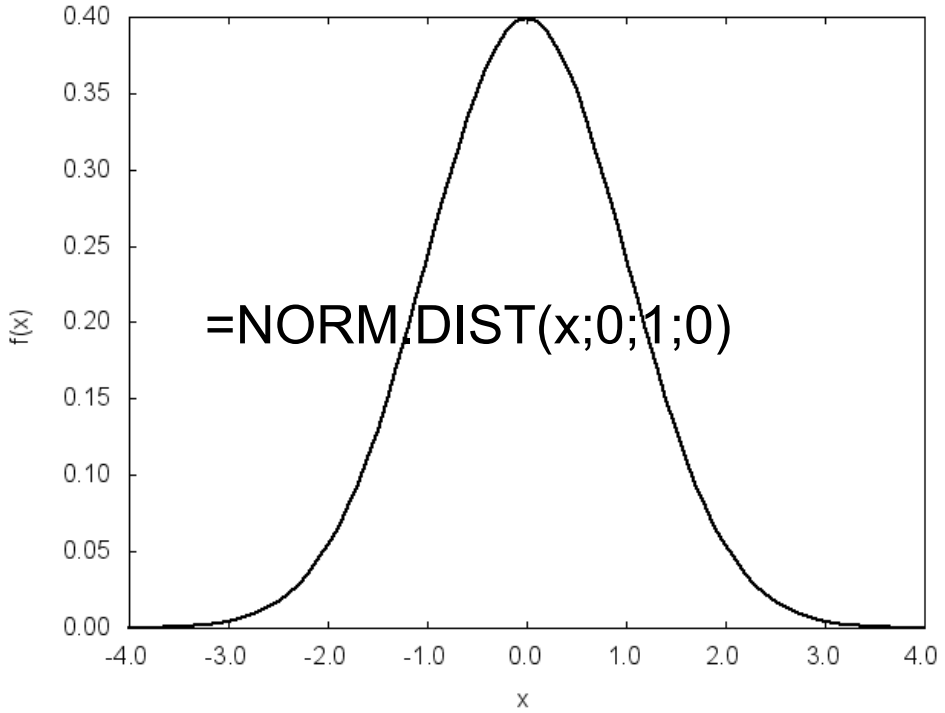
Normální (Gaussovo) rozdělení

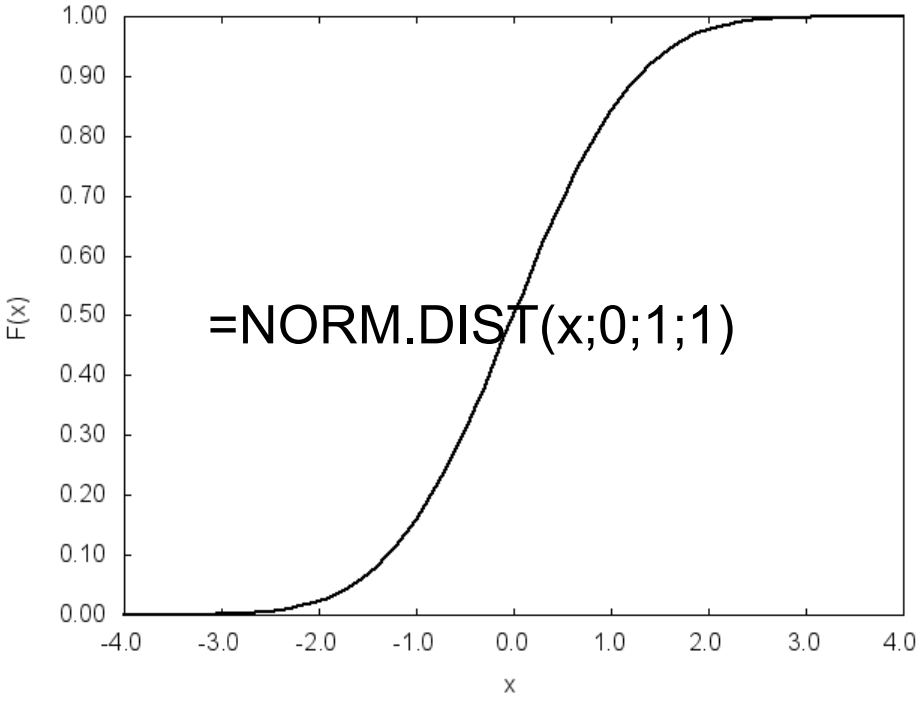


Funkce *f*(*x*) je symetrická vůči poloze maxima *x* = μ, které odpovídá současně i střední hodnotě náhodné proměnné.
S rostoucí hodnotou σ se křivka rozšiřuje a klesá její funkční hodnota v maximu v souladu s požadavkem, aby plocha pod křivkou zůstávala jednotková. Roste tak rozptyl hodnot. Hodnota σ se proto nazývá směrodatná odchylka (resp. střední kvadratická odchylka). Přesně jde o tzv. pološířku křivky mezi inflexními body.

Normální (Gaussovo) rozdělení - výpočet

Hodnoty hustoty pravděpodobnosti i distribuční funkce lze spočítat pomocí funkce
=NORM.DIST(x;střed_hod;sm_odch;S) (Excel 2007)
nebo =NORMDIST(x;střed_hod;sm_odch;S) (Excel 2003 a OOCalc)
x - hodnota, pro kterou zjišťujeme hodnotu rozdělení.
střed_hod - střední hodnota.
sm_odch - směrodatná odchylka rozdělení.
S - je-li NEPRAVDA, vrací hustotu pravd., je-li PRAVDA, vrací distribuční funkci.





Normální (Gaussovo) rozdělení - výpočet

Příklad:
Výšky studentek FT mají normálni rozdělení se střední hodnotou 168 cm a směrodatnou odchylkou 6 cm.
Odhadněte, kolik studentek FT má výšku mezi 158 a 166 cm.

=NORM.DIST(158;168;6;1) vrací 0.047
=NORM.DIST(166;168;6;1) vrací 0.369