

Přednáška IV.

Náhodná veličina, rozdělení pravděpodobnosti a reálná data

- ➔ Náhodná veličina
- ➔ Rozdělení pravděpodobnosti náhodných veličin
- ➔ Normální rozdělení a rozdělení příbuzná
- ➔ Transformace náhodných veličin



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



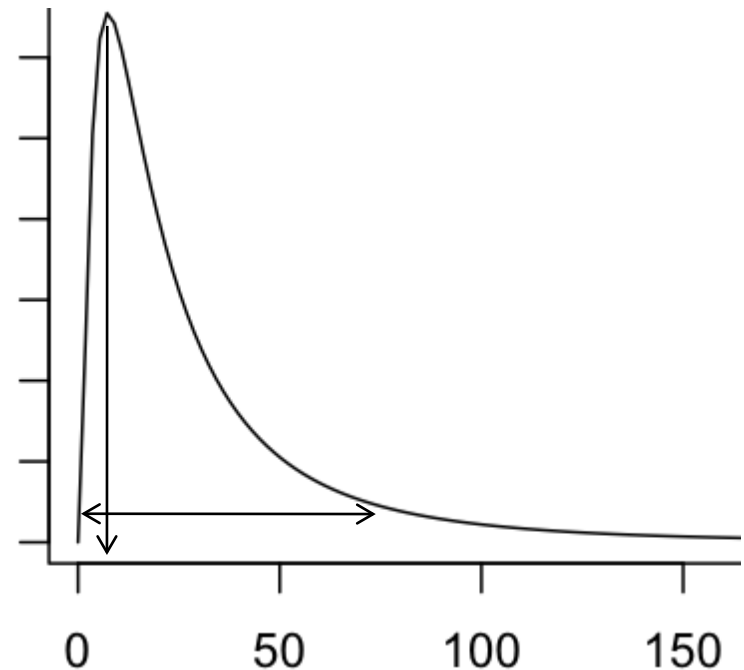
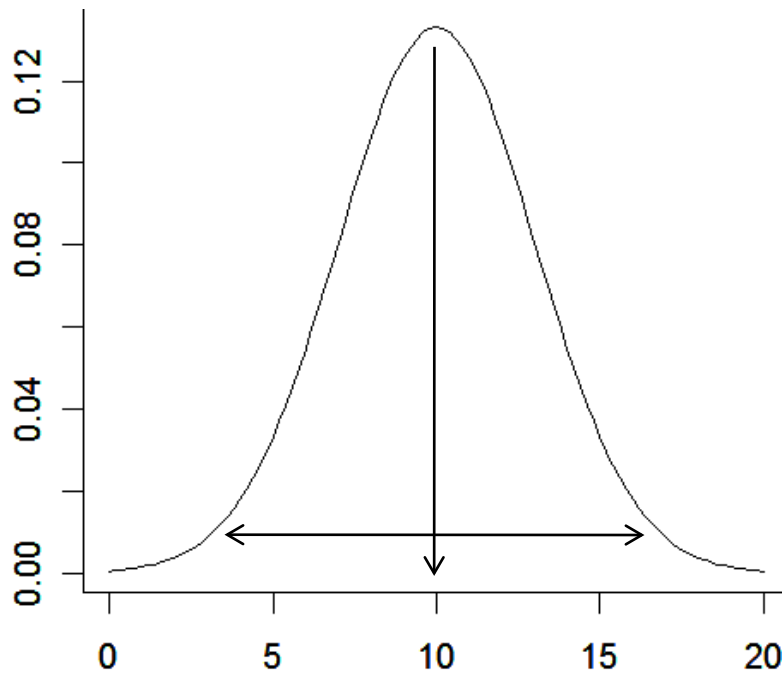
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Opakování – typy dat

- Jaké znáte typy dat?
- Uveďte příklady...

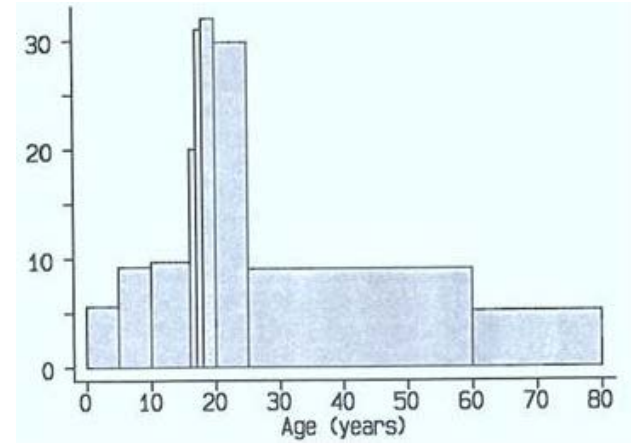
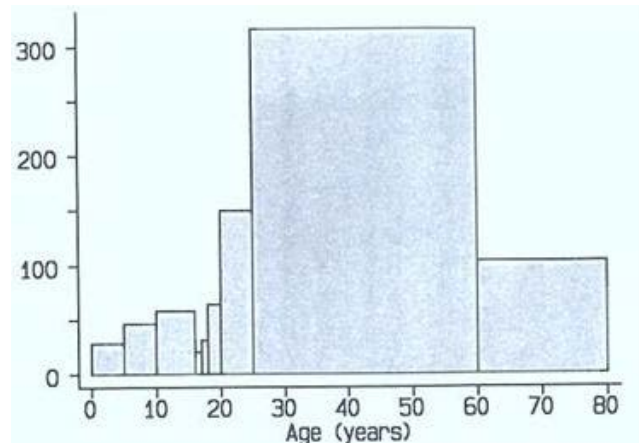
Opakování – popis dat

- Co chceme u dat popsat?
- Jak to můžeme udělat?



Opakování – který histogram je správný a proč?

- Chceme pomocí histogramu vykreslit počty zraněných při automobilových haváriích na předměstí Londýna v roce 1985. Data máme zadána jako počty v daných věkových kategoriích.



1. Náhodná veličina

Pojem náhodná veličina

- Číselné vyjádření výsledku náhodného pokusu. Matematicky je to funkce, která každému elementárnímu jevu ω z Ω přiřadí hodnotu $X(\omega)$ z nějaké množiny možných hodnot.

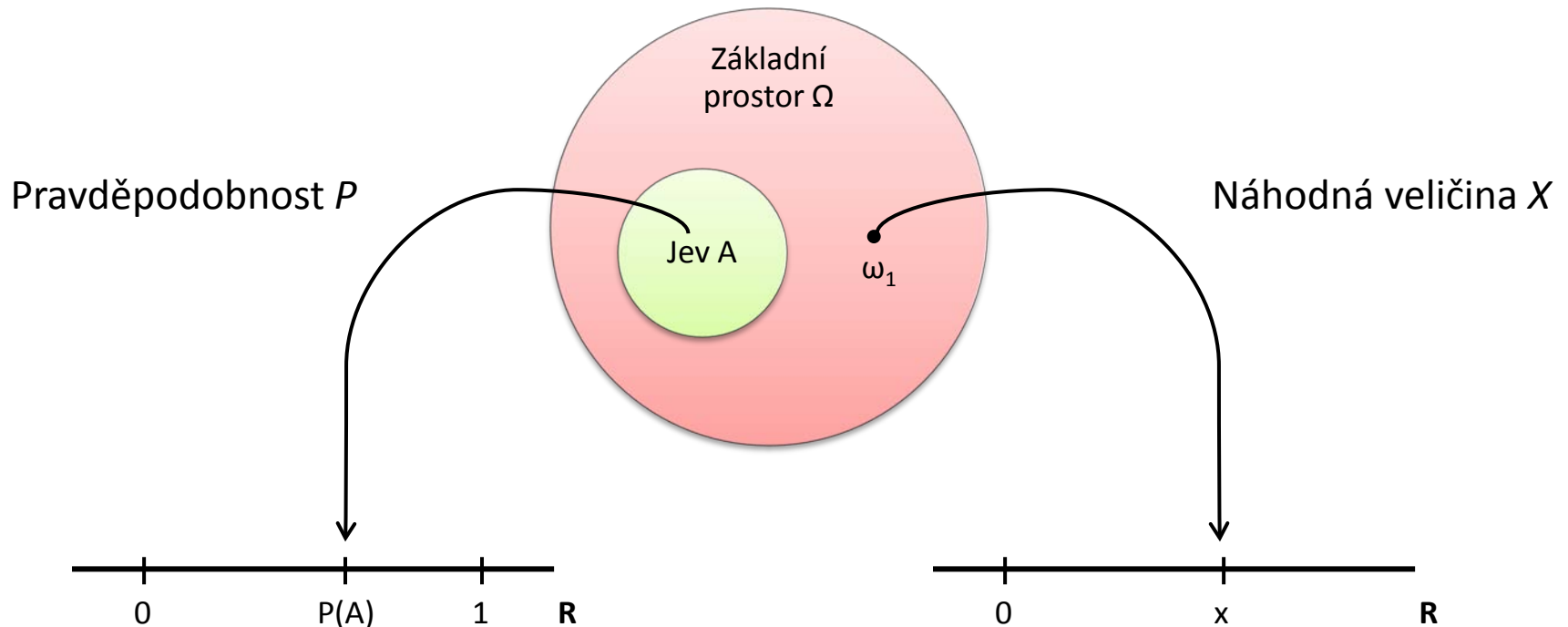
$$X : \Omega \rightarrow R$$

- Náhodná veličina se netýká pouze kvantitativních proměnných. Číselné vyjádření výsledku náhodného pokusu může popisovat i pohlaví.
- Chování náhodné veličiny lze popsat pomocí rozdělení pravděpodobnosti:
 - Funkce zadaná analyticky
 - Výčet možností a příslušných pravděpodobností



Význam náhodných veličin

- ➔ **Množina Ω často není známa** (může být i nekonečná) a nejsme tak schopni ji popsat. Náhodná veličina převádí Ω na čísla, se kterými se pracuje lépe.
- ➔ Neznáme-li Ω , nejsme schopni popsat ani X , ale **jsme schopni ho pozorovat**.



Pravděpodobnostní chování náhodné veličiny

→ Pravděpodobnostní chování náhodné veličiny je jednoznačně popsáno tzv. **rozdělením pravděpodobnosti** náhodné veličiny .

→ **Rozdělením náhodné veličiny X** definované na prostoru Ω s pravděpodobností P **rozumíme předpis**, který jednoznačně určuje všechny pravděpodobnosti typu

$$P_X(B) = P(X \in B) = P(\omega_i \in \Omega : X(\omega_i) \in B)$$

pro každou $B \subset \mathbb{R}$.

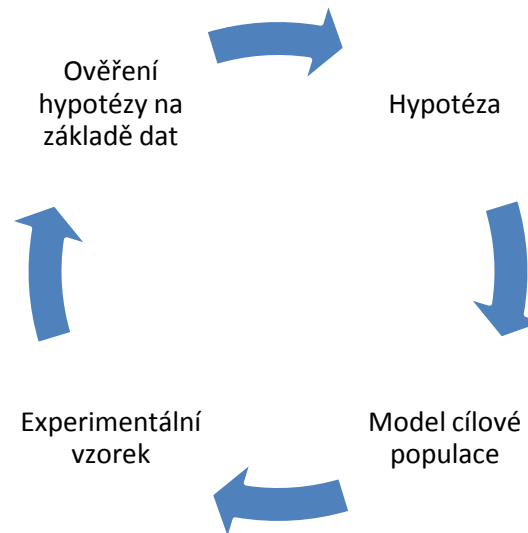
→ Distribuční funkce

→ Hustota – spojité náhodné veličiny

→ Pravděpodobnostní funkce – diskrétní náhodné veličiny

Opět vztah populace × vzorek

- Rozdělení pravděpodobnosti představuje model cílové populace.
- Pomocí vzorku (naměřených pozorování) se ptáme, jestli byl model správný – snažíme se z dat usuzovat na vlastnosti tohoto rozdělení pravděpodobnosti.



Popis rozdělení pravděpodobnosti

- Distribuční funkce popisuje rozdělení pravděpodobnosti kumulativním způsobem.
- Hustota a pravděpodobnostní funkce popisují rozdělení pravděpodobnosti pro jednotlivé „body“ (respektive intervaly) na reálné ose.
- Distribuční funkce a hustota, respektive pravděpodobnostní funkce, jsou navzájem ekvivalentní, tedy známe-li jednu nepotřebujeme druhou.

Distribuční funkce

→ Vyjadřuje pravděpodobnost, že náhodná veličina X nepřekročí dané x na reálné ose.

$$F(x) = P(X \leq x) = P(\omega_i \in \Omega : X(\omega_i) \leq x)$$

→ Vlastnosti distribuční funkce?

Distribuční funkce

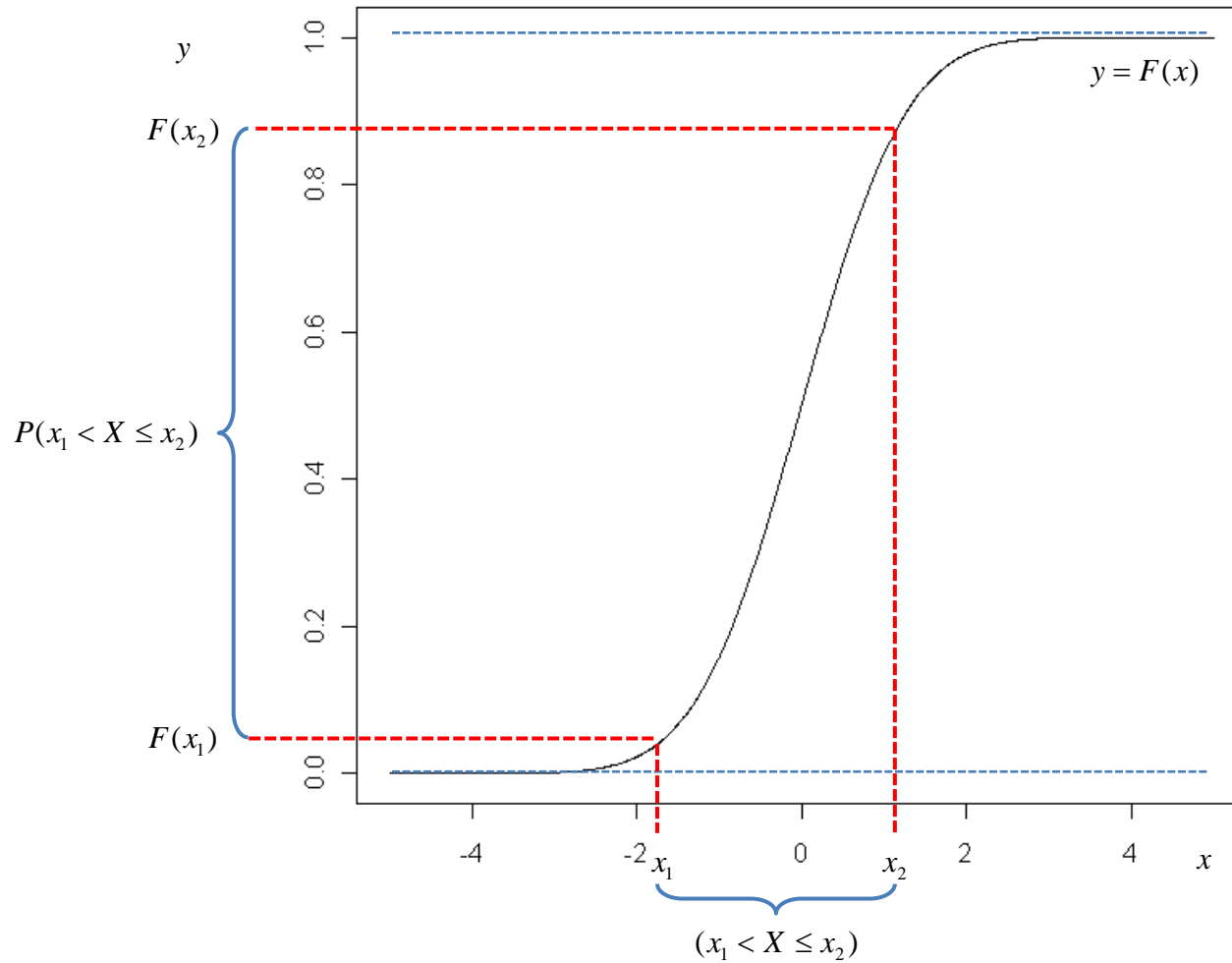
→ Vyjadřuje pravděpodobnost, že náhodná veličina X nepřekročí dané x na reálné ose.

$$F(x) = P(X \leq x) = P(\omega_i \in \Omega : X(\omega_i) \leq x)$$

→ Vlastnosti distribuční funkce:

1. Neklesající
2. Zprava spojitá
3. $0 \leq F(x) \leq 1$
4. $F(x) \rightarrow 0$ pro $x \rightarrow -\infty$
5. $F(x) \rightarrow 1$ pro $x \rightarrow \infty$

Distribuční funkce



Distribuční funkce – příklad

- Uvažujme 5 hodů mincí. Náhodná veličina X představuje počet líců.
- Jak vypadá distribuční funkce X ?

Distribuční funkce – příklad

- Uvažujme 5 hodů mincí. Náhodná veličina X představuje počet líců.
- Jak vypadá distribuční funkce X ?

$$X = \{0, 1, 2, 3, 4, 5\}$$

$$P(0) = 1 / 32$$

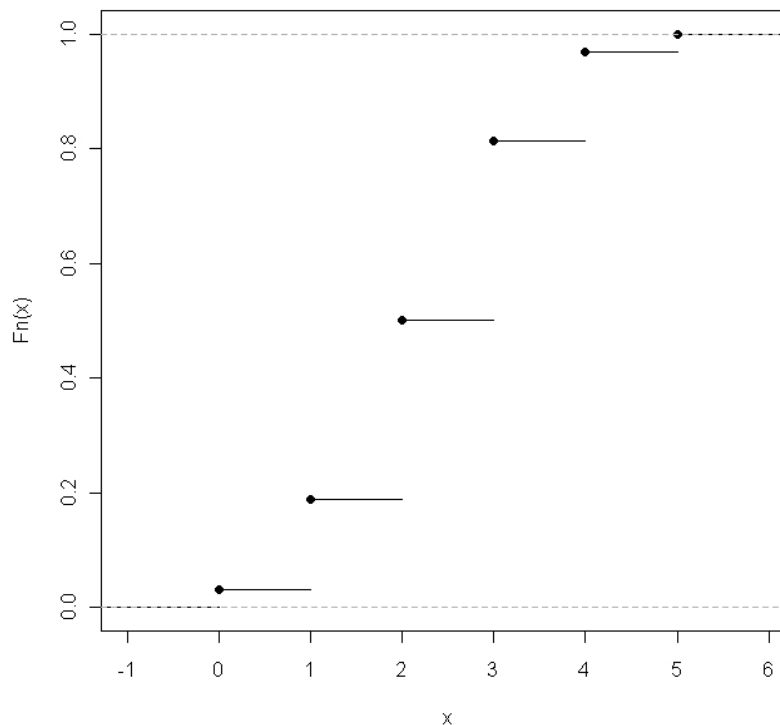
$$P(1) = 5 / 32$$

$$P(2) = 10 / 32$$

$$P(3) = 10 / 32$$

$$P(4) = 5 / 32$$

$$P(5) = 1 / 32$$



Výběrová distribuční funkce

- **Distribuční funkce je teoretická záležitost**, která definuje pravděpodobnostní model pro náhodnou veličinu X . Často neznáme její přesné vyjádření.
- **Výběrová distribuční funkce je charakteristika pozorovaných dat**. Je odhadem teoretické distribuční funkce (je-li vzorek reprezentativní).

→ Vyjádření:

$$F_n(x) = \frac{\#(x_i \leq x)}{n} = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

Výběrová distribuční funkce – příklad

→ Výška studentů 2. ročníku Matematické biologie

Spojité a diskrétní náhodné veličiny

- Náhodné veličiny dělíme dle podstaty na:
 - **Spojité** – mohou nabývat všech hodnot v daném intervalu.
 - **Diskrétní** – mohou nabývat nejvýše spočetně mnoha hodnot.
- Spojitou náhodnou veličinu X s distribuční funkcí $F(x)$ charakterizuje tzv. **hustota pravděpodobnosti**, což je funkce taková, že platí:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

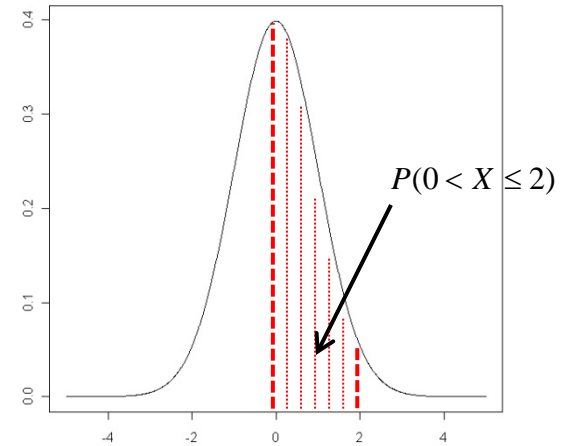
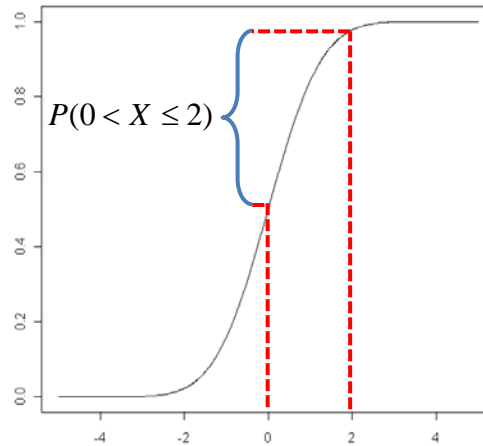
- Diskrétní náhodnou veličinu X s distribuční funkcí $F(x)$ charakterizuje tzv. **pravděpodobnostní funkce**, což je funkce taková, že platí:

$$F_X(x) = \sum_{t \leq x} p_X(t) = \sum_{t \leq x} P(X = t)$$

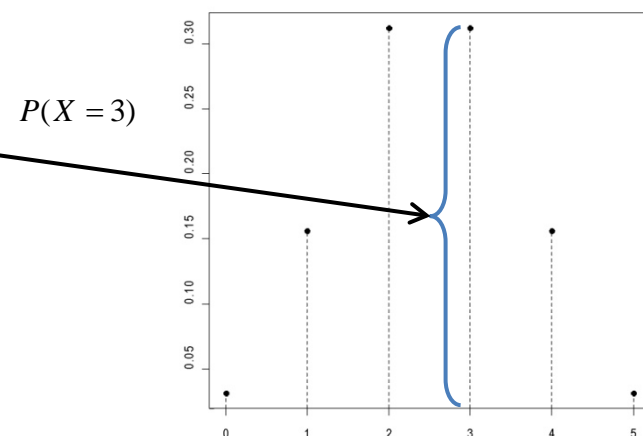
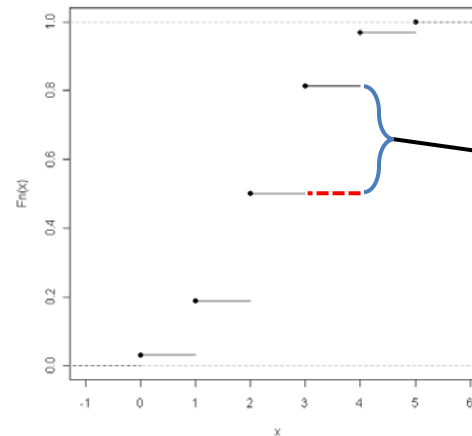


F(x) a f(x) a p(x)

**Spojitá
náhodná
veličina**



**Diskrétní
náhodná
veličina**



Spojité a diskrétní náhodné veličiny - příklady

→ Spojité náhodné veličiny:

- Medicína: výška, váha, krevní tlak, glykémie, čas do sledované události, ...
- Biologie: biomasa na m^2 , listová plocha, pH, koncentrace látek ve vodě, ovzduší, ...

→ Diskrétní náhodné veličiny:

- Medicína: počet krvácivých epizod, počet hospitalizací, počet dní po operaci do odeznění bolesti, ...
- Biologie: počet zvířat na jednotku (plochu, objem), počet kolonií na misku, ...

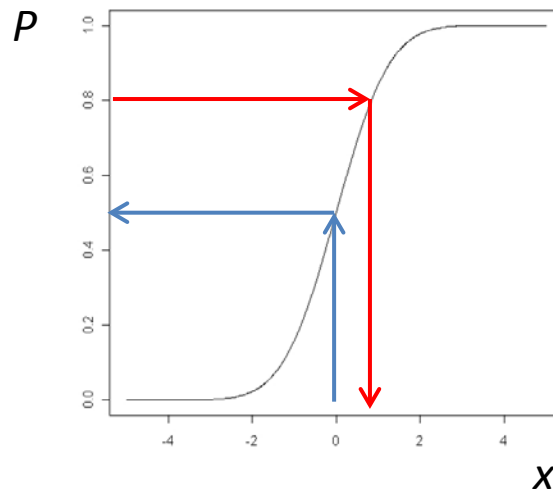
Kvantilová funkce

→ Inverzní funkce k distribuční funkci, výsledkem není pravděpodobnost, ale číslo na reálné ose, které odpovídá určité pravděpodobnosti.

→ **Distribuční funkce** $F(x) = P(X \leq x)$

→ **Kvantilová funkce** $x_p = F^{-1}(P(X \leq x)) = F^{-1}(p)$

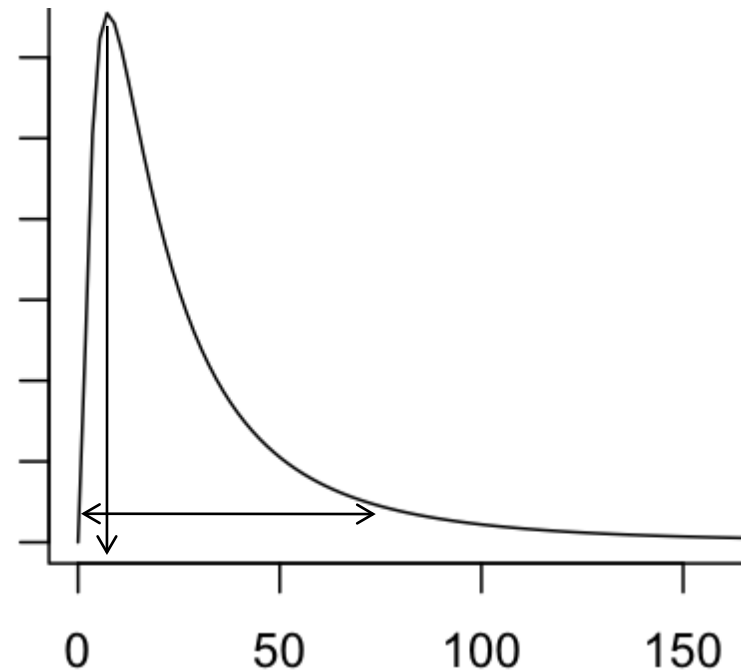
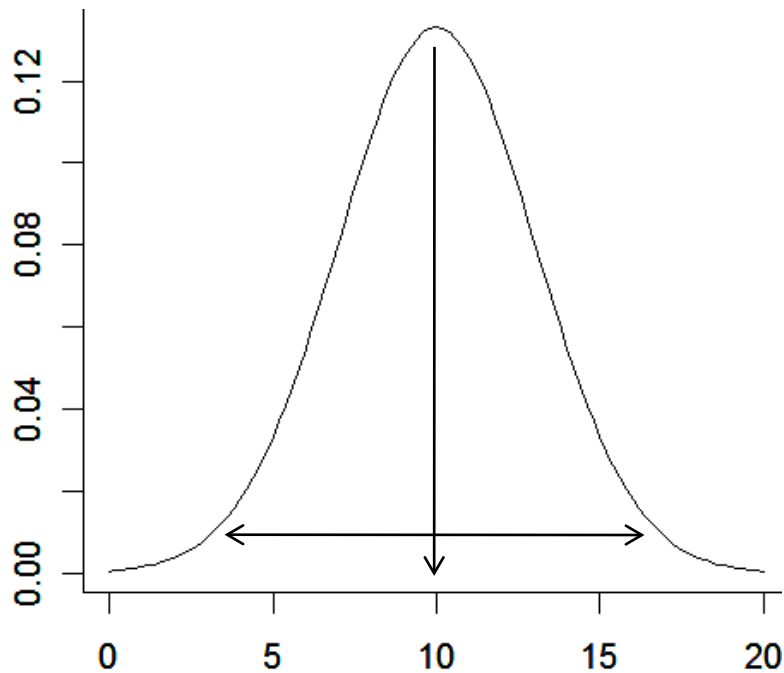
Spojité náhodné
veličina



2. Charakteristiky náhodných veličin

Co chceme u dat popsat?

- **Kvalitativní data** – četnosti (absolutní i relativní) jednotlivých kategorií.
- **Kvantitativní data** – **těžiště a rozsah pozorovaných hodnot.**



Charakteristiky náhodných veličin

- Distribuční funkce, hustota a pravděpodobnostní funkce popisují chování náhodné veličiny sice kompletně, ale trochu neprakticky – složitě.
- Jsou definovány dvě charakteristiky, které odrážejí vlastnosti rozdělení jedním číslem: **střední hodnota** a **rozptyl**.

→ Střední hodnota je definována

→ pro spojitou náhodnou veličinu X s hustotou $f(x)$ jako integrál (pokud existuje):

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

→ pro diskrétní náhodnou veličinu X s pravděpodobnostní funkcí $p(x)$ jako součet:

$$E(X) = \mu = \sum_{x \in R} xp(x)$$



Charakteristiky náhodných veličin

- Rozptyl je definován pro spojitou i diskrétní náhodnou veličinu X jako střední hodnota:

$$D(X) = \sigma^2 = E(X - E(X))^2$$

- Pro výpočet je používán vzorec:

$$\begin{aligned} D(X) &= E(X - E(X))^2 = E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2) - E(X)^2 \end{aligned}$$

- Nevýhoda rozptylu je, že není ve stejných jednotkách jako střední hodnota, proto se používá tzv. směrodatná odchylka – odmocnina z rozptylu.



Charakteristiky náhodných veličin

- To, co nás zajímalo u pozorovaných dat má teoretický ekvivalent (ve smyslu pravděpodobnosti) ve formě charakteristik náhodných veličin:

Těžiště \approx Střední hodnota

Rozsah \approx Rozptyl

- Těmto charakteristikám pak odpovídají parametry rozdělení pravděpodobnosti.
- Charakteristiky však mohou být i lehce zavádějící: **náhodná veličina nemusí nabývat své střední hodnoty**. Příklad: Náhodná veličina X nabývá hodnot -1 a 1 , obou s pravděpodobností $0,5$. Její střední hodnota je 0 !

Význam střední hodnoty

→ Jedná se o formu váženého průměru možných hodnot na základě jejich pravděpodobností.

→ Uvažujme diskrétní náhodnou veličinu

$$\rightarrow X = \{x_1, \dots, x_k\}$$

$$\rightarrow P(X=x_1) = p_1, \dots, P(X=x_k) = p_k$$

→ Pak střední hodnota má tvar:

$$E(X) = \mu = \sum_{i=1}^k x_i p(x_i)$$

Váhu pro jednotlivé hodnoty hraje jejich pravděpodobnost

Jednotlivé možné hodnoty



K čemu všechny ty funkce a čísla vlastně jsou?

- ➔ **Popis vlastností cílové populace** – na základě pozorovaných dat (histogram, box plot, popisné statistiky) jsme schopni usuzovat na charakter rozdělení pravděpodobnosti sledované veličiny. Dokonce jsme schopni otestovat míru shody s teoretickým rozdělením.
- ➔ **Srovnání vlastností cílové populace/populací** – na základě pozorovaných dat a našich předpokladů o teoretickém modelu (hypotéz) jsme schopni pomocí statistických testů srovnávat vlastnosti jedné nebo více cílových populací.
- ➔ **Predikce vlastností cílové populace** – nevyvrátíme-li na základě pozorovaných dat platnost teoretického modelu, jsme schopni se ptát, jak a s jakou pravděpodobností se bude cílová populace v budoucnu chovat.

Příklad – srovnání

→ Pacienti s hypertenzí, léčení ACE-I nebo AIIA.

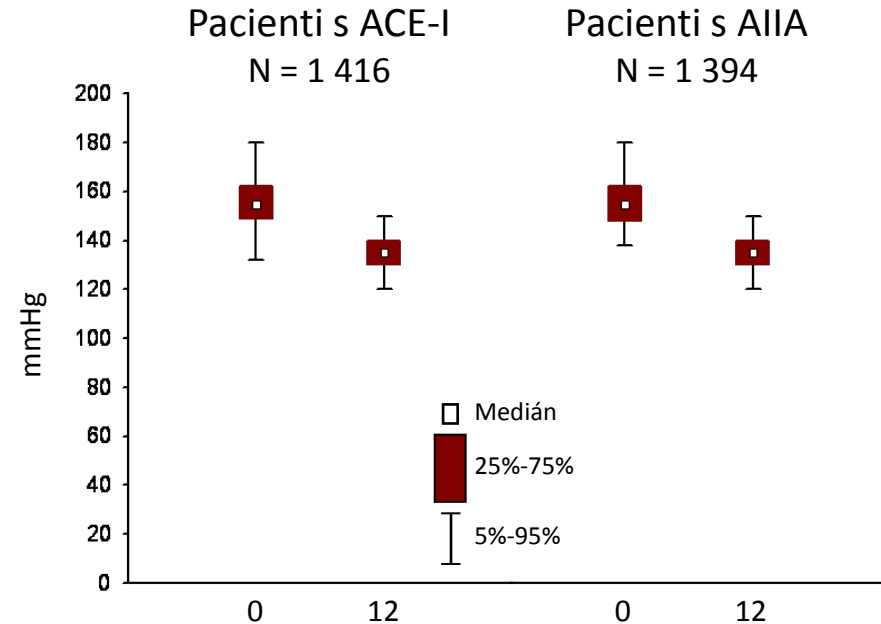
→ **Ted' předbíháme:**

→ *Vizualizace a popis* → *zhodnotíme tvar rozdělení a přítomnost odlehlých hodnot.*

→ *Testem můžeme ověřit normalitu hodnot.*

→ *Testem můžeme ověřit rovnost rozptylů.*

→ *Rozhodneme o aplikovatelnosti jednotlivých testů.*

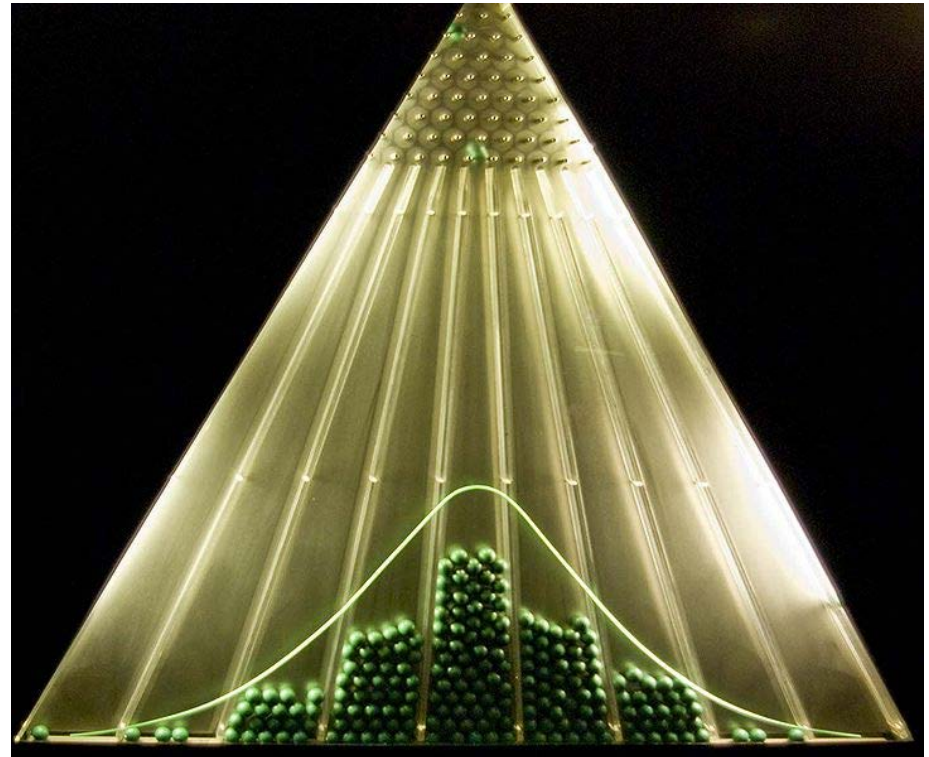


TKs v sedě (mmHg)	B ACE-I	B AIIA	p-hodnota A vs. B
Čas 0 – medián	155	155	0,929
Čas 12 měsíců - medián	135	135	
p-hodnota 0 vs. 12	<0,001	<0,001	

3. Normální rozdělení pravděpodobnosti a rozdělení z něj odvozená

Normální rozdělení pravděpodobnosti

- Klíčové rozdělení pravděpodobnosti. Jak pro teoretickou statistiku, tak pro biostatistiku.
- Označení „normální“ neznámá, že by bylo normálnější než ostatní rozdělení.
- Popisuje proměnné, jejichž hodnoty se symetricky shlukují kolem střední hodnoty. Rozptyl kolem střední hodnoty je dán aditivním vlivem mnoha „slabě působících“ faktorů.
- **Příklad:** výška člověka, krevní tlak



Normální rozdělení pravděpodobnosti

→ Je kompletně popsáno dvěma parametry:

→ μ – střední hodnota, tedy $E(X)$

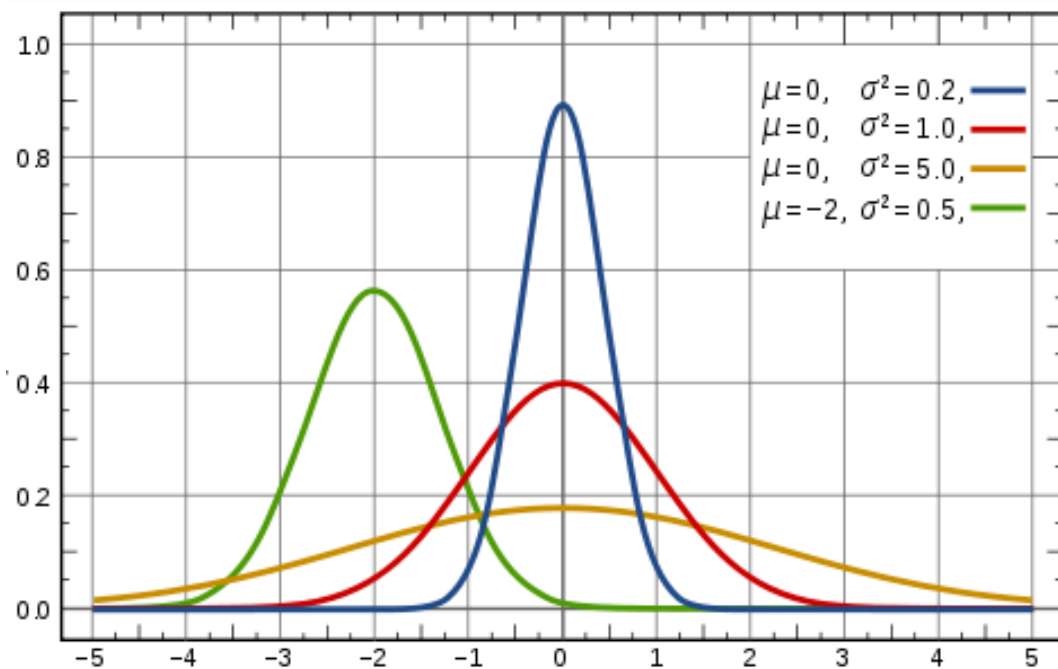
→ σ^2 – rozptyl, tedy $D(X)$

→ Označení: $N(\mu, \sigma^2)$

→ Hustota pravděpodobnosti: $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

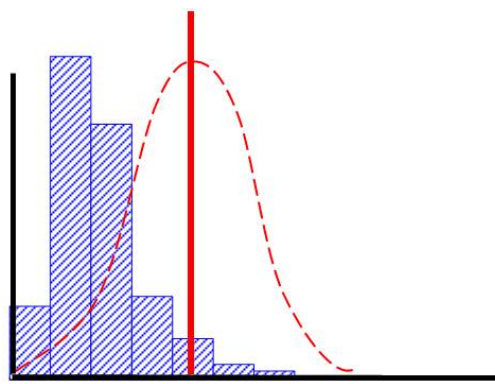
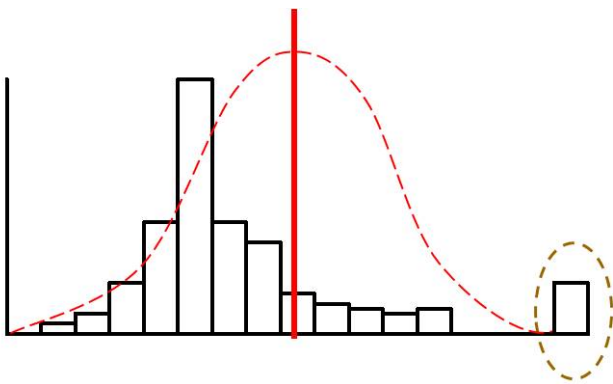
→ Čím bychom mohli jednotlivé parametry normálního rozdělení odhadnout?

Normální rozdělení dle hodnot parametrů μ a σ^2



Normální rozdělení pravděpodobnosti

- Normalita je klíčovým předpokladem řady statistických metod – zejména testů a modelů.
- Není-li splněna podmínka normality hodnot, je špatně celý model se kterým daná metoda pracuje, což vede k neinterpretovatelným závěrům.
- Její ověření je tak stejně důležité jako výběr správného testu.
- Pro ověření normality existuje řada testů a grafických metod.



Standardizované normální rozdělení

- Jakékoliv normální rozdělení může být převedeno (zatím schválně neříkám transformováno) na tzv. standardizované normální rozdělení:

$$X \sim N(\mu, \sigma^2) \rightarrow Y = \frac{X - \mu}{\sqrt{\sigma^2}} \rightarrow Y \sim N(0,1)$$

- Hustota pravděpodobnosti:

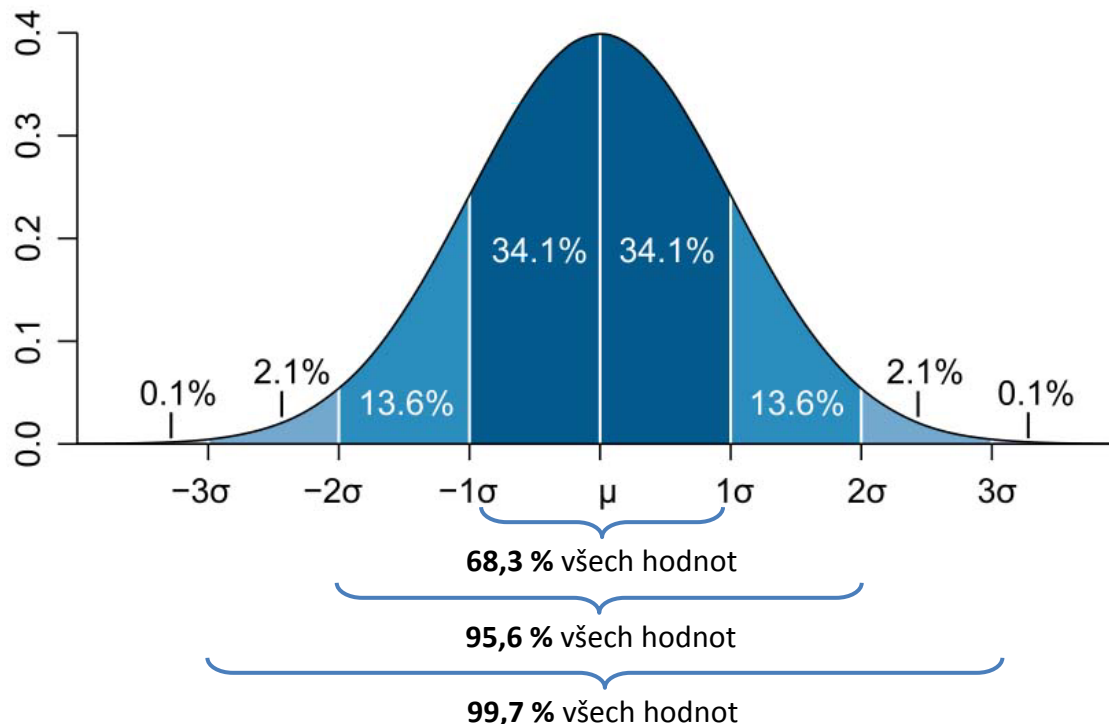
$$f(x;0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- **Klíčové rozdělení řady testů.**
- Výhoda je, že všechny hodnoty distribuční i kvantilové funkce jsou tabelovány a obsaženy ve všech dostupných softwarech.



Pravidlo ± 3 sigma

- U normálního rozdělení lze vyčíslit procento hodnot, které by se měly vyskytovat v rozmezí $\pm x$ násobku směrodatné odchylky od střední hodnoty.
- Lze říci, že v rozmezí $\mu \pm 3\sigma$ by se mělo vyskytovat přes 99,5 % všech hodnot.



Pravidlo ± 3 sigma – k čemu to je?

- Lze ho použít pro jednoduché (ale pouze orientační) **ověření normality** rozdělení pozorovaných dat.
- **Příklad 1:** Hladina sérového albuminu u 216 pacientů s cirhózou jater.
- Sumarizace pozorovaných hodnot:

$$\bar{x} = 34,46 \text{ g/l}$$

$$s = 5,84 \text{ g/l}$$

$$\bar{x} \pm 1s = 28,62 - 40,30 \text{ g/l}$$

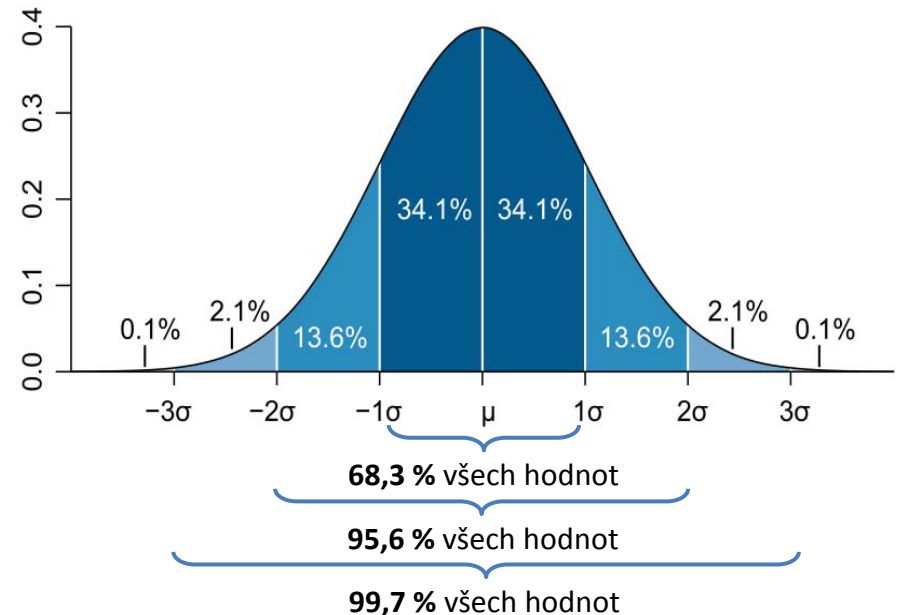
$\approx 73,15$ % hodnot

$$\bar{x} \pm 2s = 22,78 - 46,14 \text{ g/l}$$

$\approx 95,83$ % hodnot

$$\bar{x} \pm 3s = 16,94 - 51,98 \text{ g/l}$$

$\approx 99,07$ % hodnot



Pravidlo ± 3 sigma – k čemu to je?

→ **Příklad 2:** Simulovaná data, 50 hodnot z $N(0,1)$ + 1 odlehlá hodnota (200).

→ Sumarizace pozorovaných hodnot:

$$\bar{x} = 3,87$$

$$s = 28,02$$

$$\bar{x} \pm 1s = -24,15 - 31,90$$

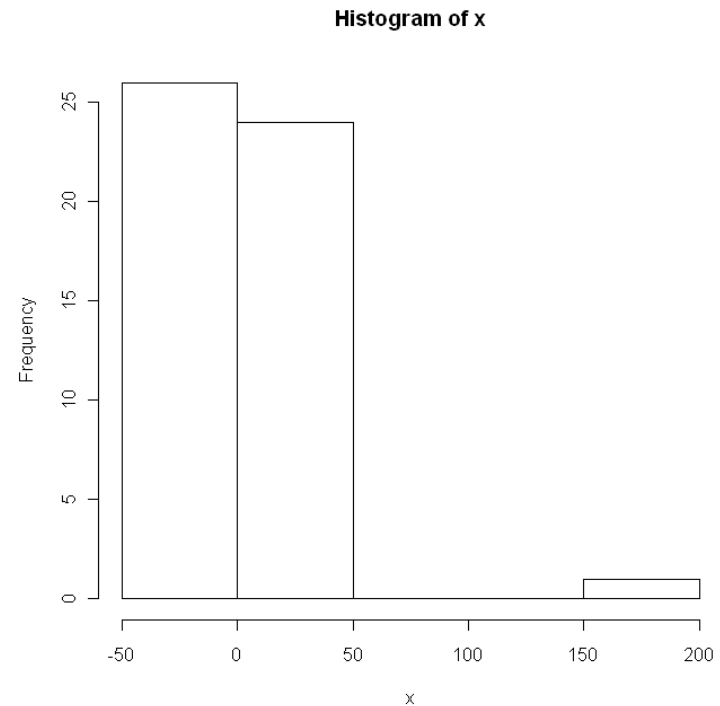
= 98,04 % hodnot \neq 68,3 % hodnot

$$\bar{x} \pm 2s = -52,18 - 59,92$$

= 98,04 % hodnot \neq 95,6 % hodnot

$$\bar{x} \pm 3s = -80,21 - 87,95$$

= 98,04 % hodnot \neq 99,7 % hodnot



Pravidlo ± 3 sigma – k čemu to je?

- Pravidlo 3 sigma můžeme použít pro identifikaci odlehlých hodnot.
- Pravidlo 3 sigma můžeme použít pro orientační ověření normality dat.

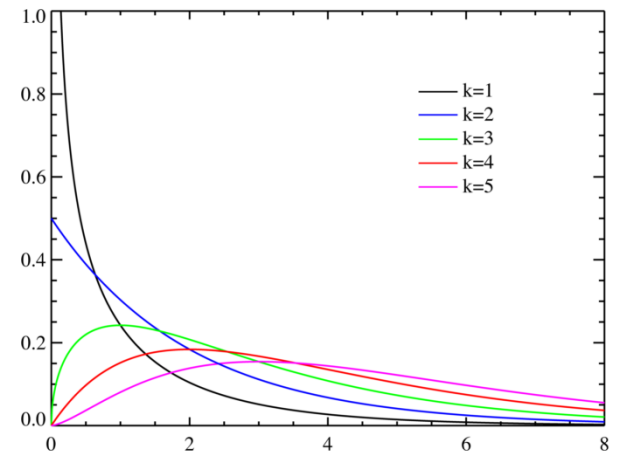
Chí-kvadrát rozdělení

- Vzniká jako součet druhých mocnin k nezávislých náhodných veličin se standardizovaným normálním rozdělením, $N(0,1)$. Konstanta k je nazývána počet stupňů volnosti.

$$X_i \sim N(0,1) \rightarrow Q = \sum_{i=1}^k X_i^2 \rightarrow Q \sim \chi^2(k)$$

- Velký význam v teoretické statistice:

- Výpočet intervalu spolehlivosti pro rozptyl
- Testování hypotéz o nezávislosti kvalitativních dat
- Testy dobré shody

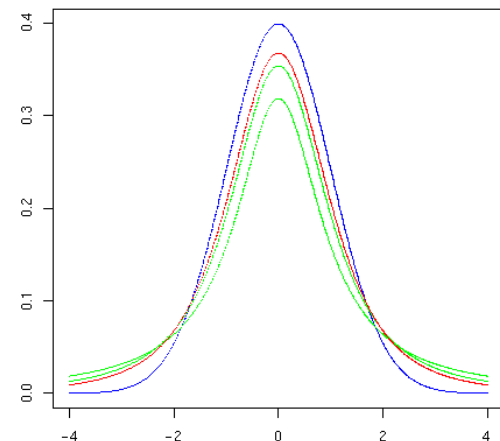


Studentovo t rozdělení

- Charakterizuje rozdělení průměru jako odhadu střední hodnoty veličiny s normálním rozdělením, v případě, že neznáme rozptyl (což je téměř vždy).
- Vzniká jako podíl dvou nezávislých veličin, jedné s rozdělením $N(0,1)$ a druhé s rozdělením $\chi^2(k)$. Parametrem t rozdělení je opět počet stupňů volnosti k .

$$X \sim N(0,1), Q \sim \chi^2(k) \rightarrow T = \frac{X}{\sqrt{Q/k}} \rightarrow T \sim t(k)$$

- Lze ho chápat jako aproximaci normálního rozdělení pro malé vzorky, pro velké velikosti souborů konverguje k normálnímu rozdělení.
- Teoretický základ t testu.



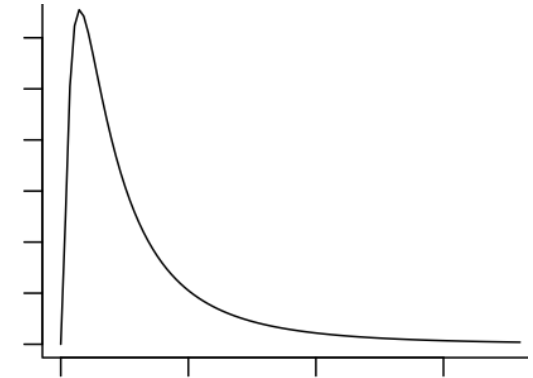
Log-normální rozdělení

→ Náhodná veličina Y má log-normální rozdělení, když $X=\ln(Y)$ má normální rozdělení. A naopak, když X má normální, pak $Y=\exp(X)$ má log-normální.

→ Hustota:
$$f(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\ln x - \mu)^2 / 2\sigma^2}, x > 0$$

→ Normální rozdělení – aditivní efekt faktorů

→ Log-normální rozdělení – multiplikační efekt faktorů



→ Řada jevů v přírodě se řídí log-normálním rozdělením: délka inkubační doby infekčního onemocnění, abundance druhů, řada krevních parametrů (např. sérový bilirubin u pacientů s cirhózou), počet bakteriálních buněk v daném objemu,...

Binomické rozdělení

→ Diskrétní rozdělení, které **popisuje počet výskytů sledované události** (ve formě nastala/nenastala) **v sérii n nezávislých experimentů**, kdy v každém experimentu **je stejná pravděpodobnost výskytu** události a je **$p = \theta$** .

→ Pravděpodobnostní funkce:

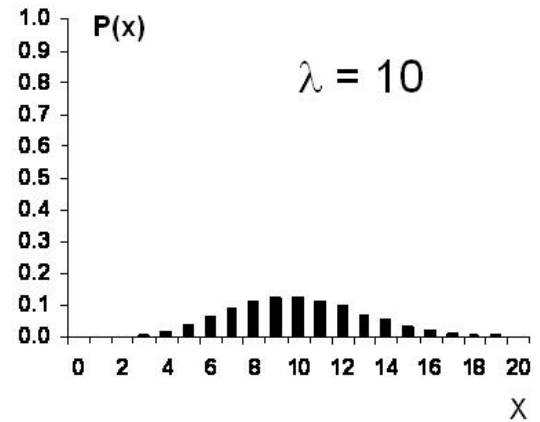
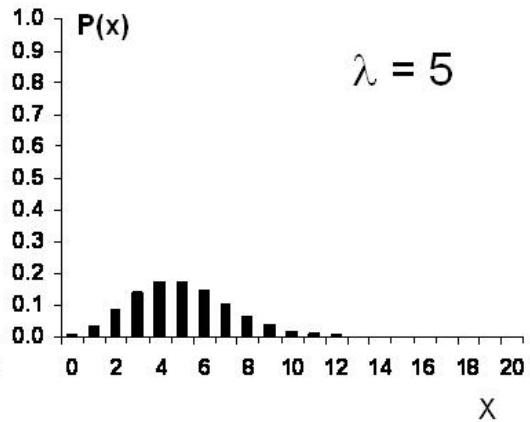
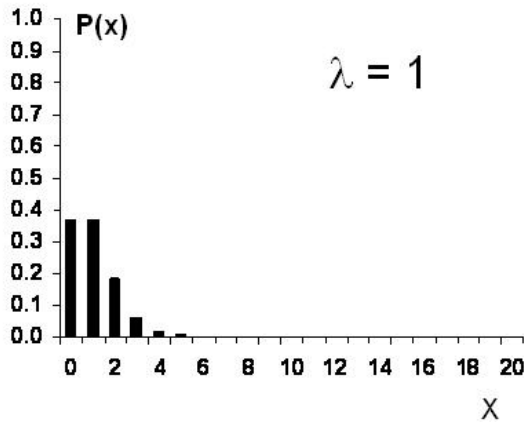
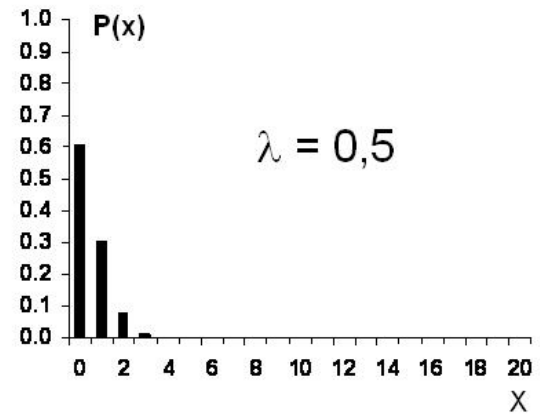
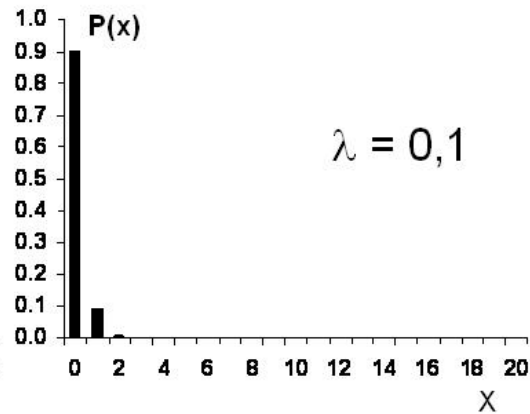
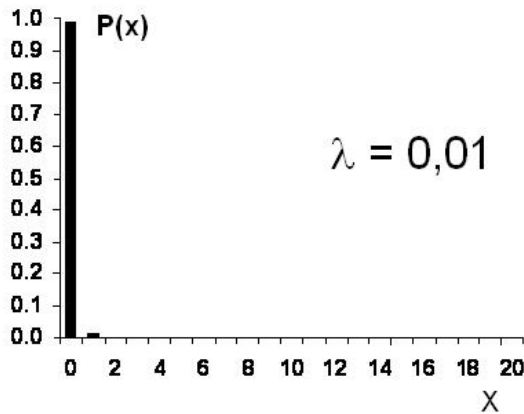
$$P(X = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

→ Základ binomických testů pro srovnávání výskytu sledovaných událostí v populaci nebo mezi populacemi.

Poissonovo rozdělení

- Diskrétní rozdělení, které **popisuje počet výskytů sledované události na danou jednotku** (času, plochy, objemu), když se tyto události vyskytují vzájemně nezávisle s konstantní intenzitou (parametr λ).
- Jedná se o zobecnění binomického rozdělení pro $n \rightarrow \infty$ a $p \rightarrow 0$.
- Pravděpodobnostní funkce: $P(X = x) = p_X(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x \geq 0$
- Střední hodnota, rozptyl: $EX = \lambda, DX = \lambda$
- **Příklady:** průměrný výskyt mutací bakterií na 1 Petriho miskou, počet krvinek v poli mikroskopu, počet žížal vyskytujících se na 1 m², počet pooperačních komplikací během určitého časového intervalu po výkonu.

Poissonovo rozdělení – vliv λ



Exponenciální rozdělení

→ Spojité rozdělení, které popisuje délky časových intervalů mezi jednotlivými událostmi Poissonova procesu. Popisuje tedy časový interval mezi událostmi, když se tyto události vyskytují vzájemně nezávisle s konstantní intenzitou (parametr λ).

→ Hustota: $f_X(x; \lambda) = \lambda e^{-\lambda x}, x \geq 0$

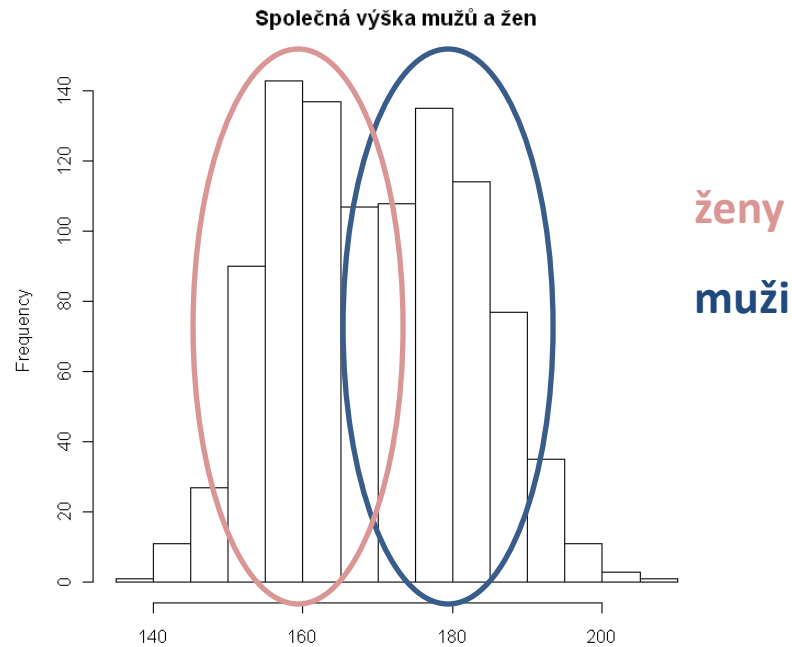
→ Střední hodnota, rozptyl: $EX = 1/\lambda, DX = 1/\lambda^2$

→ Význam v analýze přežití, je to „nejjednodušší“ modelové rozdělení pro délku doby do výskytu sledované události – předpokládá totiž konstantní intenzitu (systém nemá paměť).

→ Zobecněním jsou další rozdělení: Weibullovo, Gamma.

Bimodální rozdělení

- ➔ Představuje většinou problém, neboť se zřejmě jedná o směs dvou souborů s unimodálním rozdělením.
- ➔ Bimodální rozdělení má např. tento tvar:



Existuje ± 3 sigma i u asymetrických rozdělení?

- ➔ Pro nenormální rozdění existuje pomůcka v podobě obecného pravidla – **Čebyševovy nerovnosti**: Máme-li náhodnou veličinu X se střední hodnotou μ a konečným rozptylem σ^2 , pak pro libovolné reálné číslo $k > 0$ platí:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

4. Transformace náhodných veličin

Transformace náhodné veličiny

- Transformací náhodné veličiny X rozumíme aplikaci matematické funkce g tak, že vzniká nová náhodná veličina (tzv. transformovaná) $Y = g(X)$.
- Nová veličina nabývá nových hodnot → má také jiné rozdělení pravděpodobnosti → je třeba ho najít (hustotu, pravděpodobnostní funkci).
- S transformací se mění škála – mění se i interpretace „vzdáleností“ mezi jednotlivými hodnotami.

Transformace náhodné veličiny

→ **Spojité veličina:** chceme najít hustotu $f_Y(y)$.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), y \in R.$$

$$\text{Pro } g(x) \text{ rostoucí: } f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y), y \in R.$$

$$\text{Pro } g(x) \text{ klesající: } f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} (1 - F_X(g^{-1}(y))) = -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y), y \in R.$$

$$\text{Pro } g(x) \text{ jakoukoliv: } f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, y \in R.$$

→ **Diskrétní veličina:** chceme najít pravděpodobnostní funkci $p_Y(y)$.

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y)) = \sum_{x \in g^{-1}(y)} p_X(x), y \in R.$$

Transformace náhodné veličiny – příklad

- Máme rozdělení náhodné veličiny X dáno tabulkou a chceme najít rozdělení pravděpodobnosti transformované náhodné veličiny $Y = X^2 - 1$.

x	-2	-1	0	1	2
p(x)	0,1	0,25	0,15	0,3	0,2



x	-2	-1	0	1	2
p(x)	0,1	0,25	0,15	0,3	0,2
y	3	0	-1	0	3
p(y)	0,3	0,55	0,15	-	-

Význam transformací pro zpracování dat

- Teoretické vlastnosti transformovaných náhodných veličin nám dávají nástroj pro práci s pozorovanými daty.

- Transformace můžeme použít pro následující cíle:
 1. Normalizaci pozorovaných hodnot
 2. Standardizaci normálních hodnot
 3. Stabilizaci rozptylu pozorovaných hodnot – teď vynecháme
 4. Lepší interpretaci pozorovaných hodnot

1. Normalizace pozorovaných hodnot

- Normalita pozorovaných hodnot je silný předpoklad řady statistických metod, který musí být splněn, aby výsledky byly interpretovatelné!
- Hodnocení normality dat – vizuálně, na základě testu.
- Nenormální data je nutné transformovat nebo použít test bez předpokladu normality.

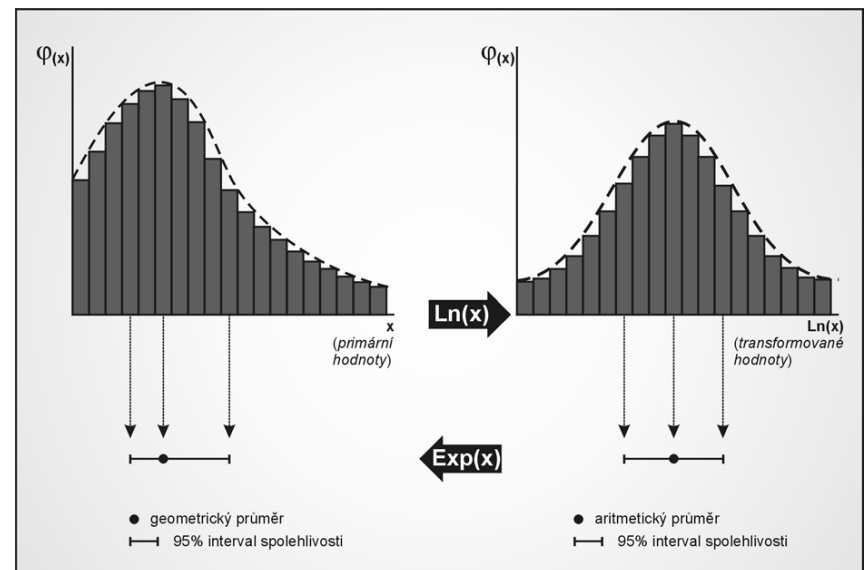
- Logaritmická transformace

$$Y = \ln(X)$$

- Odmocninová transformace

$$Y = \sqrt{x}$$

- Box-Coxova transformace



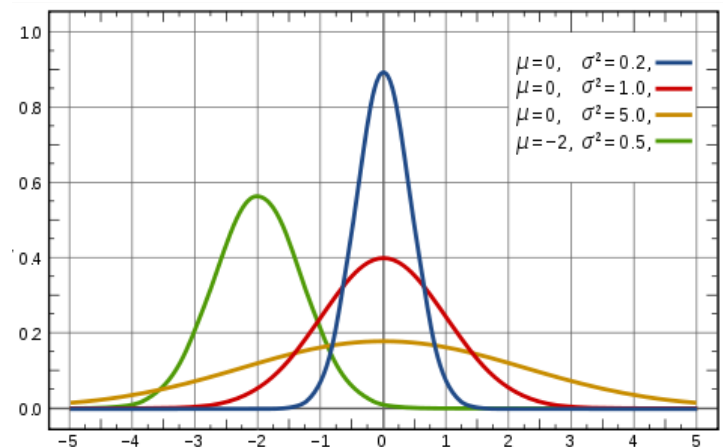
2. Standardizace normálních hodnot

- Standardizace je transformace náhodné veličiny s $N(\mu, \sigma^2)$ na $N(0,1)$.
- Důvod: řada statistických metod byla odvozena pro standardizované normální rozdělení, $N(0,1)$. Děláme to tedy opět kvůli lepší možnosti hodnocení dat.

→ Teoretická standardizace:
$$U = \frac{X - \mu}{\sqrt{\sigma^2}}$$

→ Praktická standardizace:
$$u_i = \frac{x_i - \bar{x}}{\sqrt{s^2}}$$

- Obrázek: standardizace je převod „modré“, „zelené“ a „okrové“ na „červenou“.



4. Lepší interpretace pozorovaných hodnot

- Někdy se nám hodí transformovat pozorovaná data kvůli lepší interpretaci.
- **Příklad:** Microarray experiment se dvěma vzorky, měříme intenzitu genu XY v jedné tkáni (hodnota intenzity A_{XY}) a v druhé tkáni (hodnota intenzity B_{XY}).
- Následně hodnoty převádíme na logaritmus se základem 2 jejich podílu:

$$Z_{XY} = \log_2 \left(\frac{A_{XY}}{B_{XY}} \right)$$

- Jaké to má výhody?

Poděkování...

Rozvoj studijního oboru „Matematická biologie“ PŘF MU Brno je finančně podporován prostředky projektu ESF č. CZ.1.07/2.2.00/07.0318 „Víceoborová inovace studia Matematické biologie“ a státním rozpočtem České republiky



Tomáš Pavlík



Biostatistika