Statistika popisná a induktivní

Statistika popisná popisuje datový soubor zcela, pracuje s údaji o všech uvažovaných subjektech, základní soubor je konečný.
Příklad: všechny nemocnice v ČR, všichni studenti biologických oborů…

Statistika induktivní nemá údaje o všech subjektech a statistický popis odvozuje (indukuje) z vybrané skupiny subjektů.
Příklad: populace netopýrů v jeskyni, charakteristika Pcháče osetu, …

Dvojice pojmů:

základní soubor (populace) – výběrový soubor

parametr – odhad

např. *populační průměr – výběrový průměr*

pravděpodobnost – relativní četnost

Slovníček

základní soubor – výběrový soubor (výběr)
(statistical) population – (random) sample

parametr – odhad
parameter – estimate, estimation

populační průměr – výběrový průměr
population mean – sample mean

populační rozptyl – výběrový rozptyl
population variability – sample variability

pravděpodobnost – relativní četnost
probability – relative frequency

Jedinec, subjekt, objekt, pozorování, měření, hodnota, …

Výběr ze základního souboru

Potřebujeme **reprezentativní vzorek**, který splňuje určitá pravidla. Potom můžeme použít statistické metody a výsledky budou mít smysl.

Pravidla: všechny subjekty mají **stejnou pravděpodobnost**, že budou vybrány, a výběr jednoho **nezávisí** na tom, který byl vybrán dříve. [iid. = independent and identically distributed]

Je tedy rozdíl mezi výběrem z **konečného a nekonečného** základního souboru! Prakticky: pokud je výběr velmi malou částí konečného základního souboru, řekněme do 5 %, potom lze zpravidla použít metody výběru pro nekonečný základní soubor.

Příklad (výběr „velmi malé části“):

Základní soubor má 100 jedinců/subjektů. Pravděpodobnost výběru:

1. subjekt 1/100 = 0,0**10**0
2. subjekt 1/99 = 0,0**10**1
3. subjekt 1/98 = 0,0**10**2
5. subjekt 1/96 = 0,0**10**4
6. subjekt 1/95 = 0,0**10**5 ≈ 0,0**1**1
11. subjekt 1/90 = 0,0**1**11

Základní soubor má 300 subjektů. 5 % ≈ 15 jedinců. Prst. výběru:

1. subjekt 1/300 = 0,00**3**33 ≈ 0,00**3**
11. subjekt 1/290 = 0,00**34**48 ≈ 0,00**3**
14. subjekt 1/287 = 0,00**34**84 ≈ 0,00**3**
15. subjekt 1/286 = 0,00**34**97 ≈ 0,00**3**
16. subjekt 1/285 = 0,00**35**09 ≈ 0,00**4**

Poznámky ke způsobu výběru:

- ideálně: očíslovat všechny jedince a generátorem náhodných čísel vybrat výběrový soubor (příkaz `sample`, případně `set.seed`)
- ! subjektivní výběr typu „jdu loukou a občas vyberu rostlinu“ není náhodný!
- Prakticky mnoho problémů – je třeba znát biologii sledovaných druhů a konzultovat se školiteli. Data sebraná špatným způsobem nelze dobře interpretovat!!!
- Rostliny v ploše: vytvořím systém pravoúhlých souřadnic, v počítači generuji náhodné středy pokusných ploch.
- Rostliny shlukovitě vs. solitéry: pozor, NEFUNGUJE výběr jedince nejbližšího náhodnému bodu, protože solitéry mají vyšší prst. výběru a vychylují výsledek!
- Hraboši do pastí: zkušenější jedinci budou chybět.
- Netopýři v jeskyni: nedosáhnu všude…
- Celý soubor rozdělím na homogenní podsoubory a z těch vybírám
- Laboratorní zvířata = hypotetický základní soubor: reprezentují skupinu stejně starých, stejně živených, stejně … jedinců.

Náhodnost v experimentech –

dodržet náhodný design !

- Když mám laboratorní či polní pokus, typicky také plánuji nějaká **opakování**. Musím pak věnovat pozornost náhodnému přiřazení jedinců do jednotlivých květináčů/terárií apod. Také přiřazení jedince k typu zásahu/ošetření/ životních podmínek musí být náhodné.

- Čti: Lepš & Šmilauer, str. 23 – 24, str. 189 – 190.
- **Pseudoreplikace**: měření, která ve skutečnosti nejsou nezávislá
- Obr. L&Š str 190.

- **R**: přiřazení jedince k typu zásahu (***k*** typů, ***m*** opakování v jednom typu)

Náhodná veličina

Ω = { … } … množina všech možných výsledků; lze definovat přesně?

$\mathbf{p_i}$ … pravděpodobnost, že vyberu konkrétní hodnotu (výsledek)
odhaduji pomocí **relativní četnosti**

→ **„chování“ náhodné veličiny = rozdělení pravděpodobnosti**

Příklad pastelky

Pravděpodobnost, že vyberu červenou pastelku?

Počet červených pastelek v jednom výběru?

Žádná červená pastelka … (ooooo ooooo) … 1 možnost

Jedna červená … 10 možností

(●oooo ooooo),(o●ooo ooooo),(oo●oo ooooo),(ooo●o ooooo),(oooo● ooooo),
(ooooo ●oooo),(ooooo o●ooo),(ooooo oo●oo),(ooooo ooo●o),(ooooo oooo●)

Kombinační číslo: $\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!} = \frac{n\cdot(n-1)\cdot(n-2)\cdot(n-3)\cdot \ldots \cdot 3\cdot 2\cdot 1}{k\cdot(k-1)\cdot(k-2)\cdot \ldots \cdot 2\cdot 1 \cdot (n-k)\cdot(n-k-1)\cdot \ldots \cdot 2\cdot 1}$

Náhodná veličina: příklad pastelky

Počet červených pastelek v jednom výběru?

Dvě červené … 45 možností

9 (●●ooo ooooo), (●o●oo ooooo), (●oo●o ooooo), (●ooo● ooooo), (●oooo ●oooo), (●oooo o●ooo), (●oooo oo●oo), (●oooo ooo●o), (●oooo oooo●)

8 (o●●oo ooooo), (o●o●o ooooo), (o●oo● ooooo), (o●ooo ●oooo), (o●ooo o●ooo), (o●ooo oo●oo), (o●ooo ooo●o), (o●ooo oooo●)

7 (oo●●o ooooo), (oo●o● ooooo), (oo●oo ●oooo), (oo●oo o●ooo), (oo●oo oo●oo), (oo●oo ooo●o), (oo●oo oooo●)

6 (ooo●● ooooo), (ooo●o ●oooo), (ooo●o o●ooo), (ooo●o oo●oo), (ooo●o ooo●o), (ooo●o oooo●)

5 (oooo● ●oooo), (oooo● o●ooo), (oooo● oo●oo), (oooo● ooo●o), (oooo● oooo●)

4 (ooooo ●●ooo), (ooooo ●o●oo), (ooooo ●oo●o), (ooooo ●ooo●)

3 (ooooo o●●oo), (ooooo o●o●o), (ooooo o●oo●)

2 (ooooo oo●●o), (ooooo oo●o●)

1 (ooooo ooo●●)

Náhodná veličina

Náhodná veličina

Když chceme popsat hodně velký (až nekonečný) základní soubor, pracujeme většinou jen s výběrem, výběrovým souborem.
Z výběrových dat potom spočítáme nějakou charakteristiku, která má reprezentovat vlastnost celého základního souboru.
Takový výběr můžeme mnohokrát opakovat. Zvolená charakteristika pak bude mít pokaždé trochu jinou hodnotu, protože na výsledku se podílí i náhoda – ve výběru subjektů do výběrového souboru.
Proto všechny hodnoty pozorované nebo měřené na náhodně vybraných subjektech (počet listů na rostlině, délka zobáku kosa) nazýváme **náhodná veličina** nebo **náhodná proměnná** [random variable] a konkrétní zjištěnou hodnotu realizace náhodné veličiny.

Někdy mluvíme také o výsledku náhodného procesu. Tím se myslí například měření rychlosti, kterou padá semeno trubicí (proces). Výsledkem je potom ta rychlost.

Náhodná veličina

Matematika vidí náhodnou veličinu jako funkci, která každému subjektu přiřadí hodnotu z množiny možných hodnot. Chování náhodné veličiny potom popisuje jako výčet přípustných hodnot a pravděpodobností, s nimiž mohou subjekty těchto hodnot nabývat.

Příklad: holčičky (●) a kluci (□) v rodině se 3 dětmi:

(●,●,●); (●,●,□); (●,□,●); (□,●,●); (●,□,□); (□,□,●); (□,●,□); (□,□,□).

→ $\mathrm{P}(0\ \text{kluků}) = 1 * \frac{1}{8} = 0{,}125$ $\mathrm{P}(2\ \text{kluci}) = 3 * \frac{1}{8} = 0{,}375$

$\mathrm{P}(1\ \text{kluk}) = 3 * \frac{1}{8} = 0{,}375$ $\mathrm{P}(3\ \text{kluci}) = 1 * \frac{1}{8} = 0{,}125$

[čti: pravděpodobnost, že v rodině bude právě 1 kluk, se rovná…]

Pokud výčet popíše pravděpodobnosti všech možných výsledků, nazýváme ho **rozdělením pravděpodobnosti** [probability distribution], často jenom rozdělení nebo distribuce.

Poznámka: V teorii přísně rozlišujeme náhodné veličiny s diskrétním rozdělením pravděpodobností (počet dětí) a se spojitým rozdělením prstí (výška člověka). V praxi se často použije spojité rozdělení pro diskrétní data tam, kde dostávám alespoň desítky různých hodnot (např. počty krvinek).

Značení

Náhodnou veličinu označujeme **X, Y, Z,** tj. velká písmena z konce abecedy.

Zápis: $P(X = x_i) = p_i, x_i \in \{x_1, x_2, \cdots, x_m\}$

[čti: pravděpodobnost, že náhodná veličina X nabyde hodnoty x_i, je p_i. x_i leží v množině hodnot $x_1, \cdots, x_m$, kterých může nabývat veličina X.]

Diskrétní rozdělení pravděpodobností - tabulkou

- Typické pro data na nominální a ordinální stupnici a počty něčeho
- Hodnoty x_i jsou od sebe jasně odděleny, je jich nejvýše spočetně
- Chování diskrétní náhodné veličiny mohu popsat „tabulkou“ (např. barva pastelky, počet kluků v rodině), ale také vzorečkem.

Vlastnosti:

každé $\boldsymbol{p_i \geq 0}$ (tedy neznáme zápornou pravděpodobnost)

součet $\boldsymbol{\sum_{i=1}^{m} p_i = 1}$ (tedy množina $\{x_1, x_2, \cdots, x_m\}$ popisuje všechny možnosti, hodnoty pro náhodnou veličinu X)

Proto máme v dotazníku políčko „jiné“, popíšeme tak všechny možnosti.

Spojité rozdělení pravděpodobností - funkcí

Příklad: délka pastelky, váha biomasy apod.

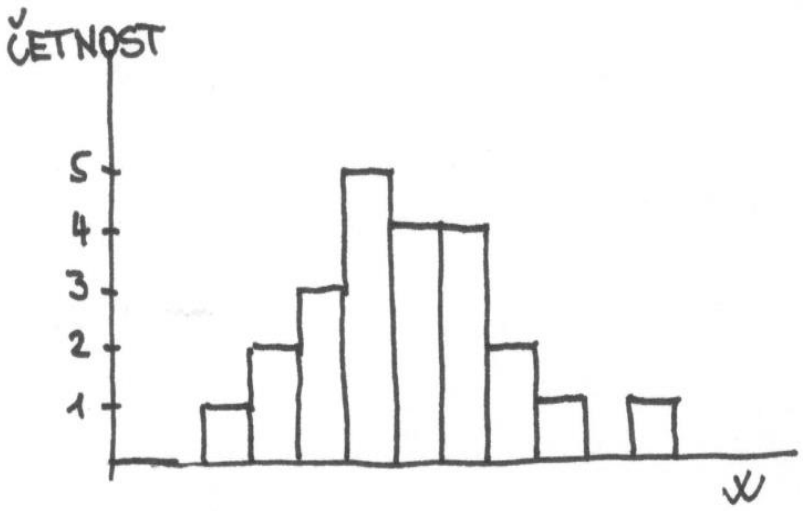
Toto rozdělení „prstí“ nemohu popsat tabulkou, protože veličina X může nabývat nekonečně mnoha hodnot (milimetr a půl, milimetr a $^3/_4$).

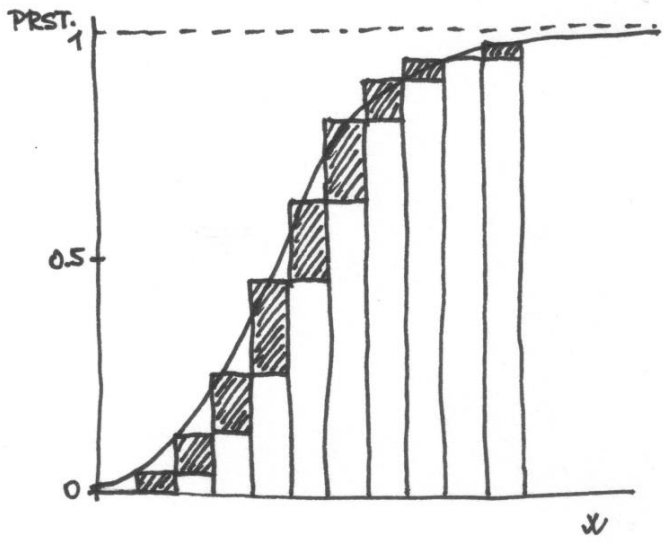
Neptáme se na jednu hodnotu $P(X = 5)$, ale spíše na $P(X \leq x_i)$.

Spojité rozdělení popisujeme funkcí.

Tvar funkce vyčteme z histogramu četností.

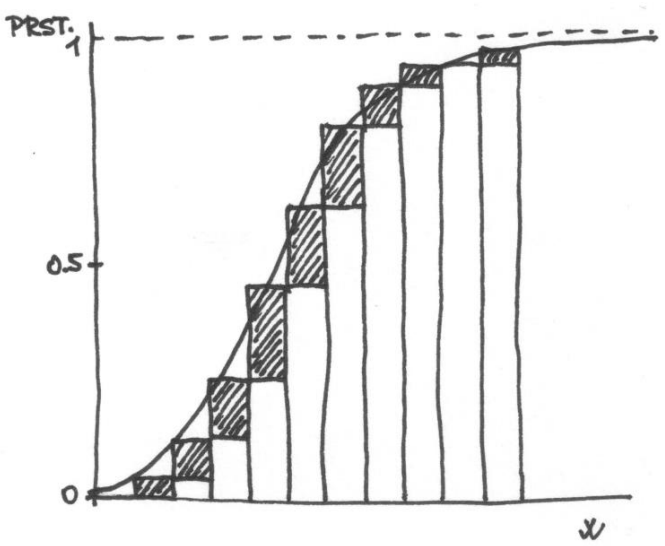
Z něho sestrojíme histogram kumulativních (relativních) četností:





Distribuční funkce náhodné veličiny X

[(cumulative) distribution function]



Dvojí značení:

$$\boldsymbol{F(x) = P(X \leq x)}$$

$$\boldsymbol{\Phi(x) = P(X \leq x) = p}$$

… někteří autoři píší $F(x) = P(X < x)$.
Je to také dobře, protože pro spojitou distribuční funkci platí $\boldsymbol{P(X = x) = 0}$, tedy prst, že veličina nabyde hodnoty právě $\boldsymbol{x}$, je nulová.

Pozor, pro diskrétní náhodnou veličinu musím definovat, kam patří mezní hodnoty, proto je tam rozdíl mezi $P(X \leq x)$ a $P(X < x)$.

Vlastnosti distribuční funkce

- $0 \leq F(x) \leq 1$
- Je neklesající
- $\lim_{x \to -\infty} F(x) = 0; \quad \lim_{x \to +\infty} F(x) = 1$
- $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

Pro X s diskrétním rozdělením píšeme:

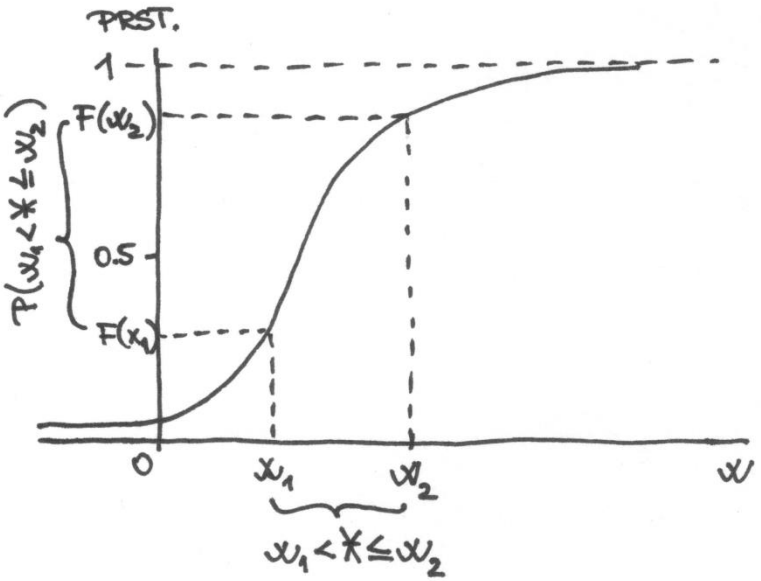
$$F(x) = \sum_{t \leq x} p(t) = \sum_{x_i \leq x} p_i = \sum_{x_i \leq x} P(X = x_i)$$

funkce $p(t)$ se nazývá pravděpodobnostní funkce [probability mass function]

Pro X se spojitým rozdělením lze zapsat:

$$\boldsymbol{F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt}$$

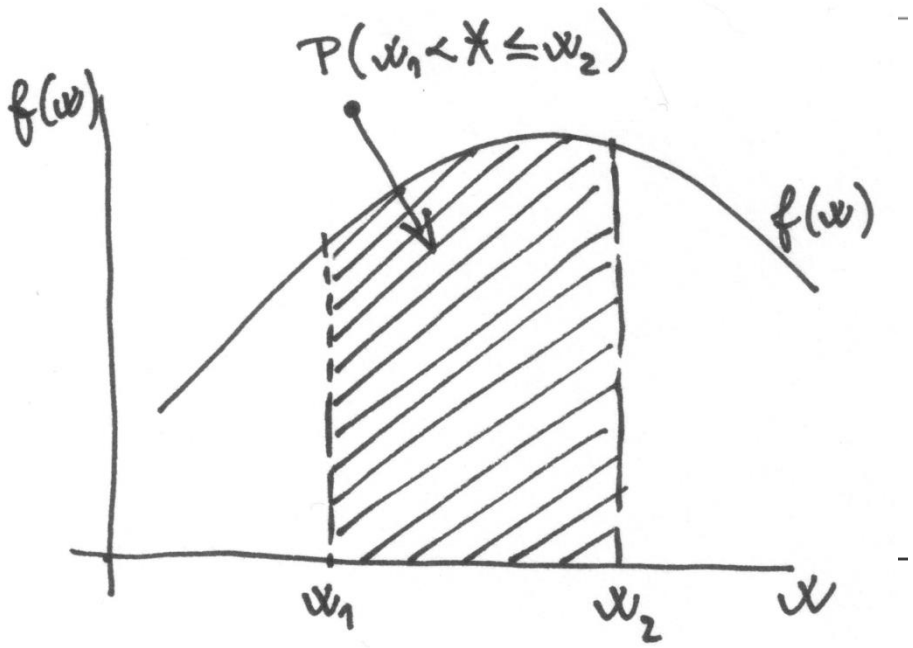
kde funkce *f(x)* je derivací distribuční funkce F(x): $F'(x) = f(x)$



Funkci ***f(x)*** nazýváme **hustotou pravděpodobností** náhodné veličiny X [probability density function].

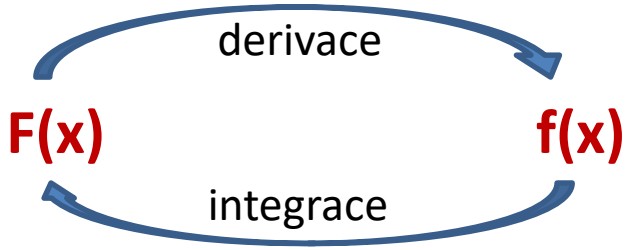
Platí: $\boldsymbol{P(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)dt}$

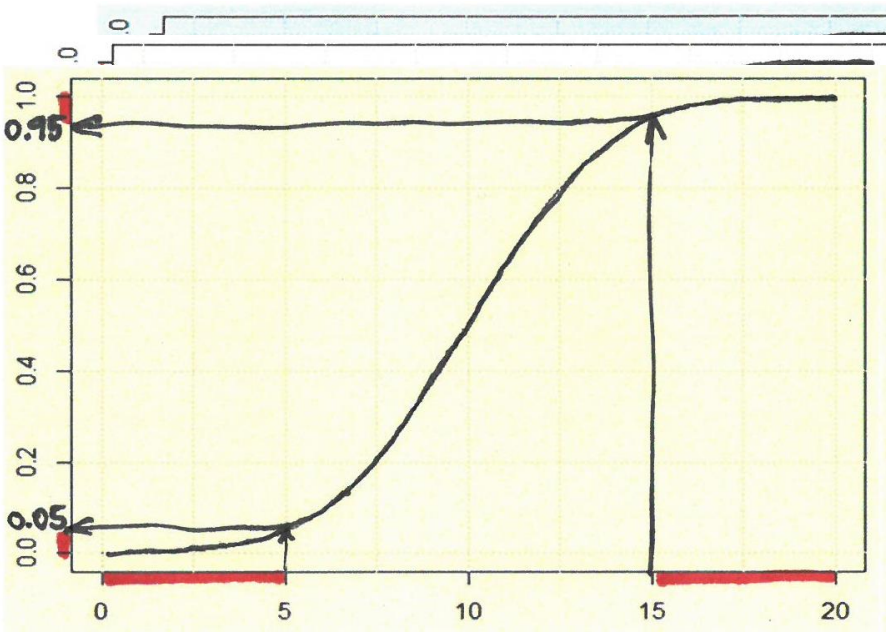
a $\int_{-\infty}^{\infty} \boldsymbol{f(x)dx = 1}$ … hodnota X leží jistě mezi $-\infty$ a $+\infty$.



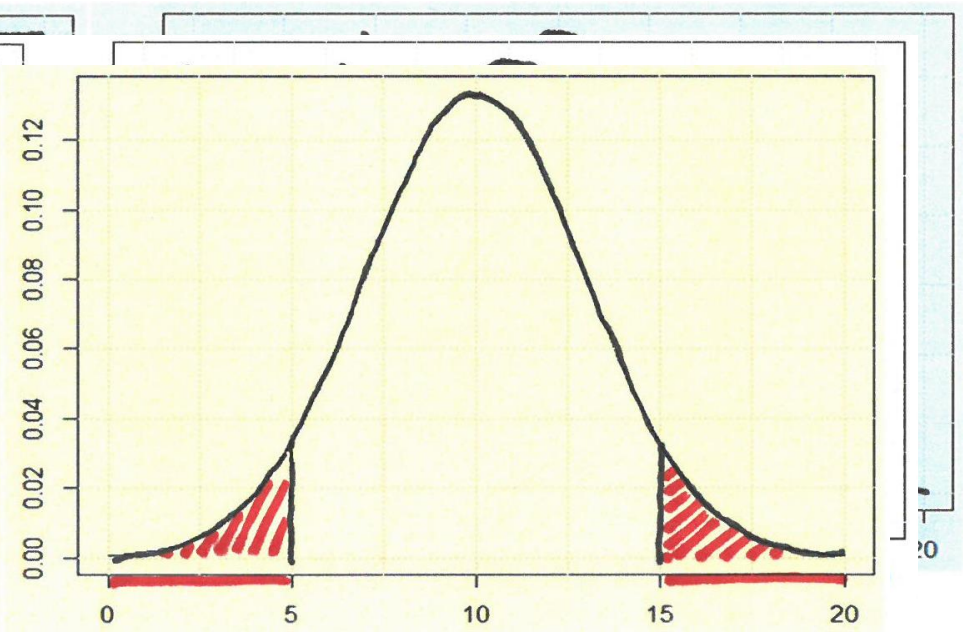
! <u>Pozor na měřítka</u>:
histogram četností ukazuje počty hodnot v intervalu,
histogram <u>relativních</u> četností ukazuje pravděpodobnost, že vyberu hodnotu z daného intervalu.
Funkci hustoty pravděpodobností *f(x)* můžeme chápat jako idealizovaný histogram relativních četností pro nekonečně velký základní soubor.

Vztah mezi distribuční a „hustotní“ funkcí:



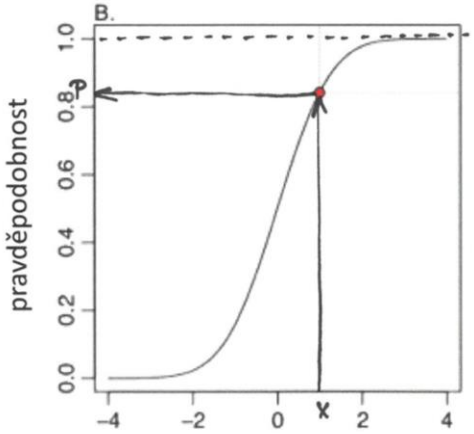


$P(X \leq 5)$ & $P(X > 15)$



$P(X \leq 5)$ & $P(X > 15)$

Kvantil, kvartil [quantile, quartile]



Například $x_{0,95}$ ≈ 95% kvantil [*čti: devadesáti-pěti-procentní kvantil*]

$x_{0,25}$ ≈ 25% kvantil ≈ 1. kvartil

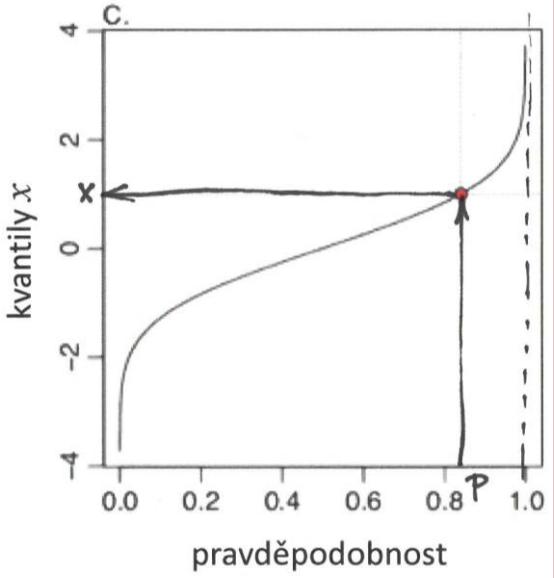
$$\mathbf{P}(\mathbf{X} \leq x_p) = p$$

p-kvantil je taková hodnota x_p, pro kterou $F(x_p) = \mathbf{P}(\mathbf{X} \leq x_p) = p$

$95\%\ kvantil = x_{0,95}: P(X \leq x_{0,95}) = 0,95$

a také $P(X > x_{0,95}) = 0,05$

Tuto vlastnost využíváme při testování hypotéz.



Kvantilová funkce je inverzní funkce k distribuční funkci ***F(x)***

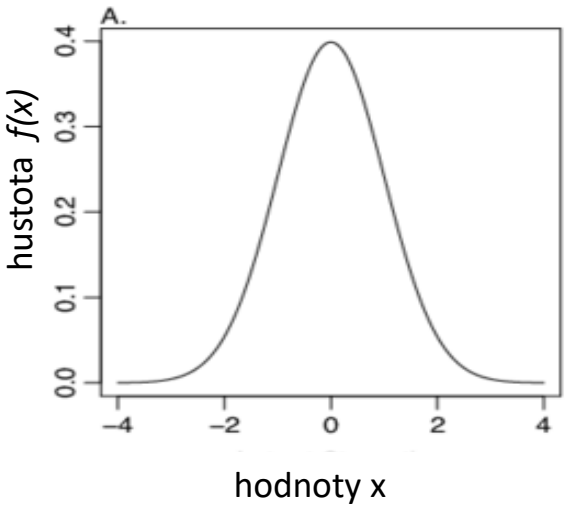
Inverzi značíme $F^{-1}(p)$ a funguje takto: $F(x) = p$

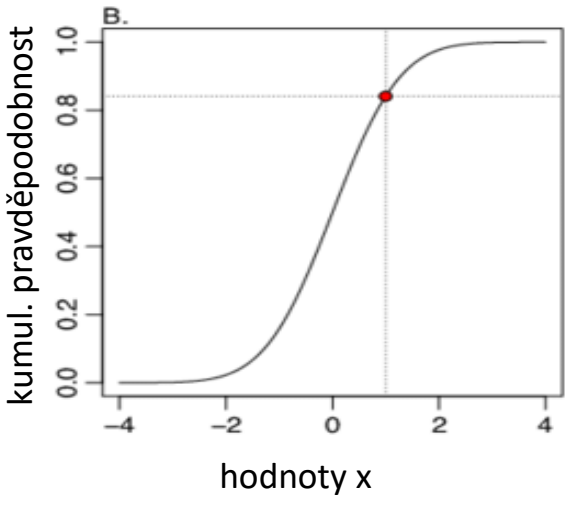
$F^{-1}(p) = x$

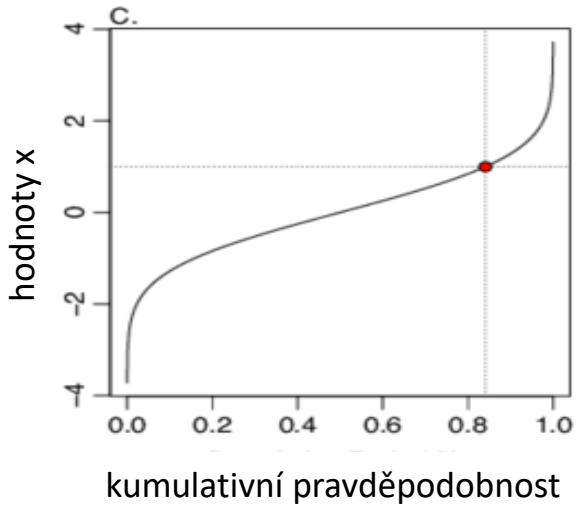
$F^{-1}(F(x)) = x$

Trojice funkcí: **Hustota** – **distribuční funkce** – **kvantilová funkce**

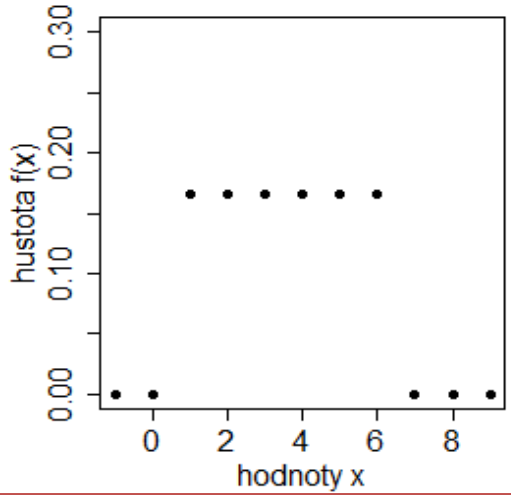
Příklad: normální rozdělení

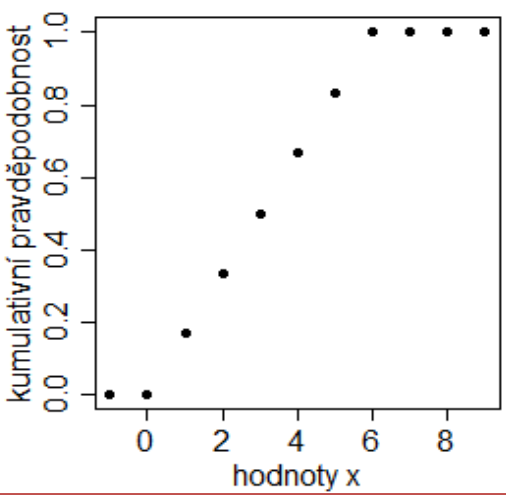


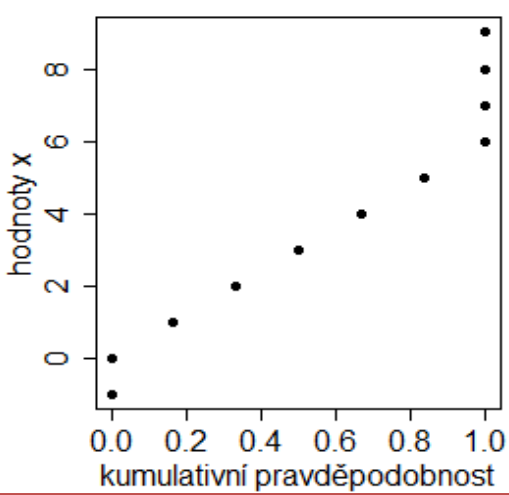




rovnoměrné diskrétní rozdělení: hrací kostka, hodnoty 1 až 6.







Některá rozdělení

Náhodná veličina + náhodný proces = konkrétní hodnota (realizace).
Popis, že určitá hodnota x_i nastane s pravděpodobností p_i, se nazývá rozdělení pravděpodobností náhodné veličiny.
Zde některá rozdělení, která jsou popsaná funkcí a jsou užitečná:

Alternativní rozdělení [alternative distribution] (Bernoulliho rozd.)

$X \sim Alt(p)$ Příklad: Házím korunou. Padne lev?

- Nejjednodušší případ, X nabývá pouze hodnot 0 a 1.
- Data jako „úspěch – neúspěch“, „přítomný – nepřítomný“, „nastal – nenastal“.
- Popis rozdělení: $P(X = 1) = p$ a $P(X = 0) = 1 - p = q$
- Potom $E(X) = p$
  $var(X) = p \cdot (1 - p) = p \cdot q$

Binomické rozdělení [binomial distribution]

Příklad: Házím korunou desetkrát za sebou. Kolikrát mi padne lev?

Vyberu 20 pastelek. Kolik z nich bude červených?

$Y \sim Bi(n, p)$ *[čti: náh. vel. Y má binomické rozdělení s parametry n a p]*

- Zjišťujeme pouze výskyt či nevýskyt jevu ***B*** v pokusu
- Parametr ***n*** udává celkový počet pokusů
- Pokusy jsou na sobě nezávislé
- Prst ***p*** výskytu jevu ***B*** je v každém pokusu stejná

- ***Y*** nabývá jedné z hodnot ***0, 1, 2, 3, …, n*** s pravděpodobností

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k}$$

Binomické rozdělení

Příklad: holčičky (●) a kluci (□) v rodině se 3 dětmi:

(●,●,●); (●,●,□); (●,□,●); (□,●,●); (●,□,□); (□,□,●); (□,●,□); (□,□,□).

→ $\mathrm{P}(0\ \text{kluků}) = 1 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$ $\mathrm{P}(2\ \text{kluci}) = 3 \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right)\left(\frac{1}{2}\right)$

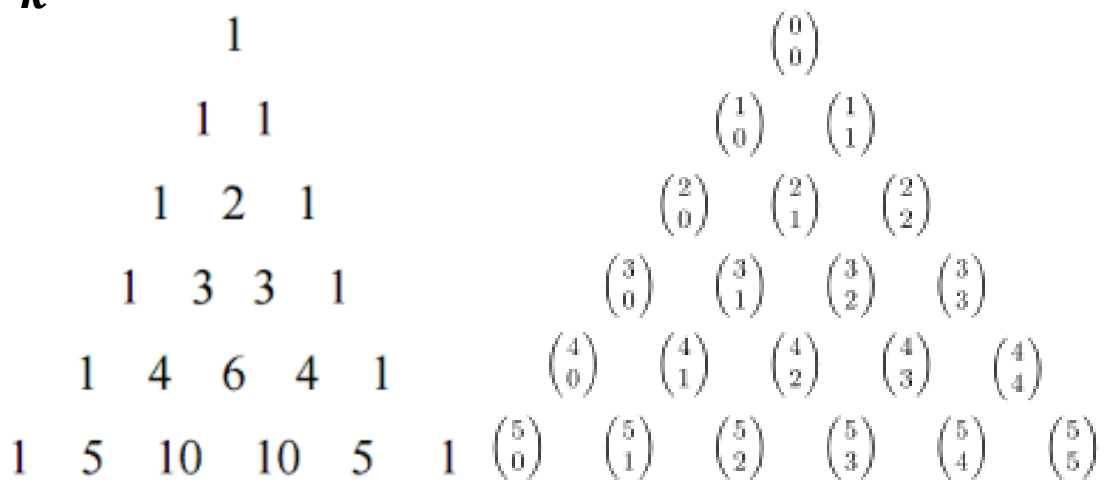
$\mathrm{P}(1\ \text{kluk}) = 3 \cdot \left(\frac{1}{2}\right)\left(\frac{1}{2} \cdot \frac{1}{2}\right)$ $\mathrm{P}(3\ \text{kluci}) = 1 \cdot \left(\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}\right)$

- ***Y*** nabývá jedné z hodnot ***0, 1, 2, 3, …, n*** s pravděpodobností

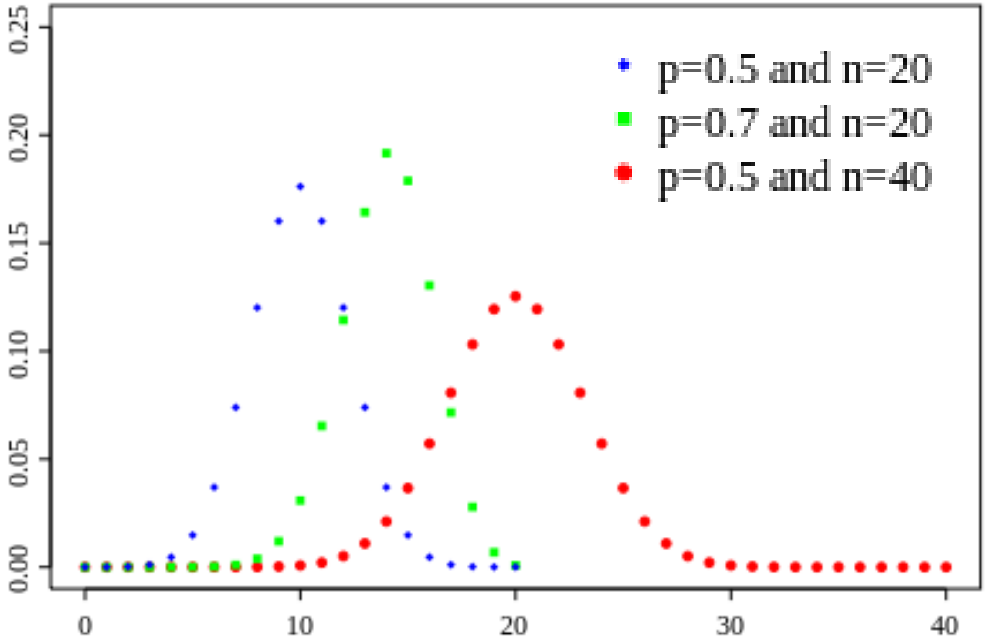
$$\boldsymbol{P(Y = k) = \binom{n}{k} p^k q^{n-k}}$$

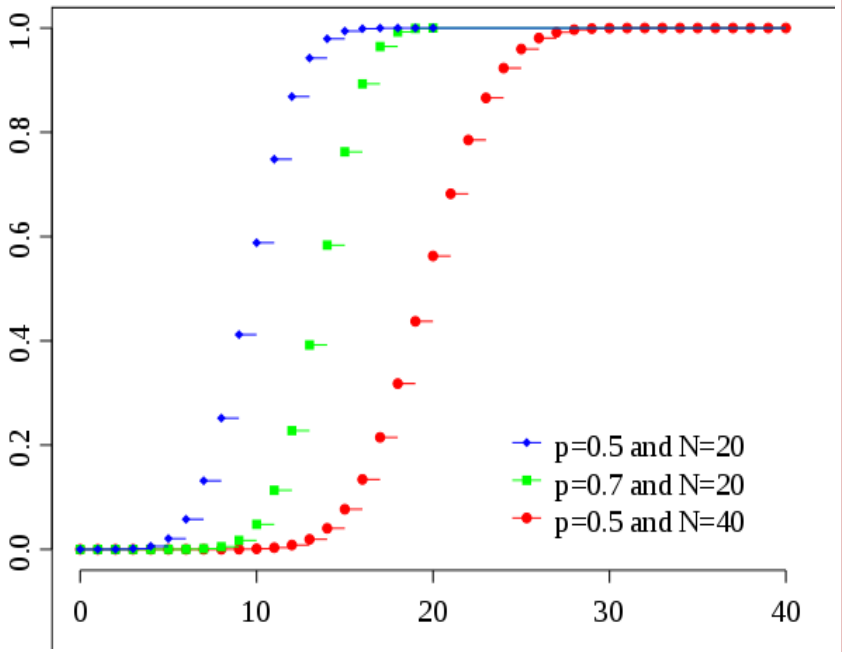
- Pomůcka:

  Pascalův trojúhelník

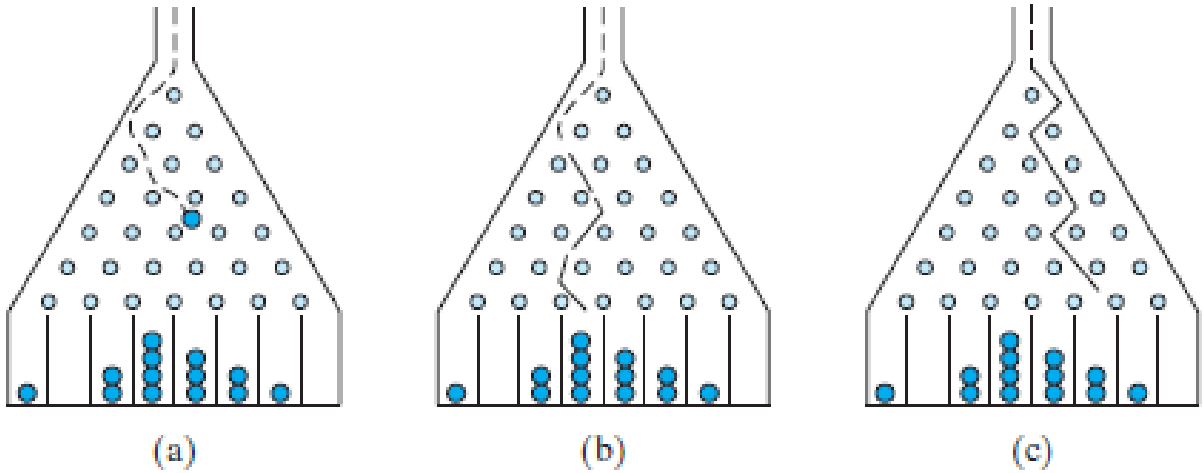


Binomické rozdělení graficky:





Galtonova deska či opilcova procházka



https://www.youtube.com/watch?v=3m4bxse2JEQ

Poissonovo rozdělení [Poisson distribution]

Příklad: Sleduju, kolik trolejbusů projede zastávkou za jednotku času.

$X \sim Po(\lambda)$ *[čti: X má Poasonovo rozdělení s parametrem lambda]*

X nabývá hodnot **k = 0, 1, 2, 3, …** (bez omezení shora) s prstí

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$\mu_X = \lambda$ … střední hodnota

$\sigma^2{}_X = \lambda$ … rozptyl

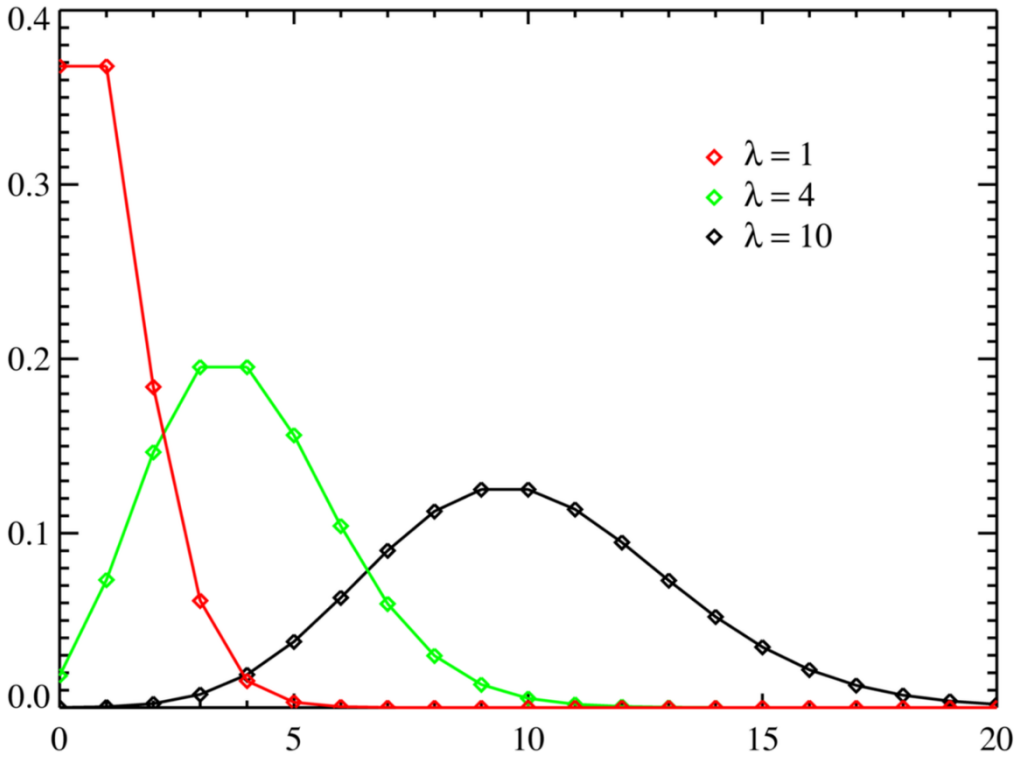
$\lambda > 0$ … kladné reálné číslo (0.25; 1.8)

- Popisuje počet náhodných, vzájemně nezávislých jevů v jednotce času nebo prostoru.
- Jevy v čase nastávající zřídka – „zákon vzácných jevů“.
- Jevy v prostoru mají být o počtech subjektů v malých objemech nebo v řídké suspenzi.
- Užití: pomocí Poissonova rozdělení testujeme otázky o náhodnosti rozmístění jedinců v ploše/času; také zda výskyt jednoho jedince ovlivňuje výskyt dalších jedinců, nebo zda žijí na sobě nezávisle.

Vlastnosti Poissonova rozdělení

Typický tvar dostává Poissonovo rozdělení pro malá λ (< 2): výrazně pozitivně šikmé.

(To vadí při regresní analýze i při analýze rozptylu. Řešíme to odmocninovou transformací nebo lépe užitím GLM /zobecněných lineárních modelů/)



- Platí, že když nezávislé $\boldsymbol{X \sim Po(\lambda_1)}$ a $\boldsymbol{Y \sim Po(\lambda_2)}$, tak $\boldsymbol{X + Y \sim Po(\lambda_3)}$.
- Pro vyšší hodnoty λ (> 10) lze data aproximovat normálním rozd.

Další příklady Poissonova rozdělení

V čase:

- Počet nezávislých kolonizací vzdáleného ostrova za jednotku času;

V prostoru:

- Počet bakterií v jednotce objemu vodní suspenze, pokud se bakterie nevyskytují ve shlucích (mikrobiologie);
- Rozmístění klíšťat v srsti myši (parazitologie);
- Počet jedinců kruštíku bahenního ve 100 pokusných čtverců (ekologie);

Rozmístění rovnoměrné – náhodné – shlukovité

Mám-li data o počtu jedinců na pokusnou plochu, dostávám pro

- rovnoměrné rozmístění: $\mu_X > \sigma^2{}_X$ (Binomické rozdělení)
- náhodné rozmístění: $\mu_X = \sigma^2{}_X$ (Poissonovo rozdělení)
- shlukovité rozmístění: $\mu_X < \sigma^2{}_X$ (negativně–binomické rozd či Neymanovo rozdělení)

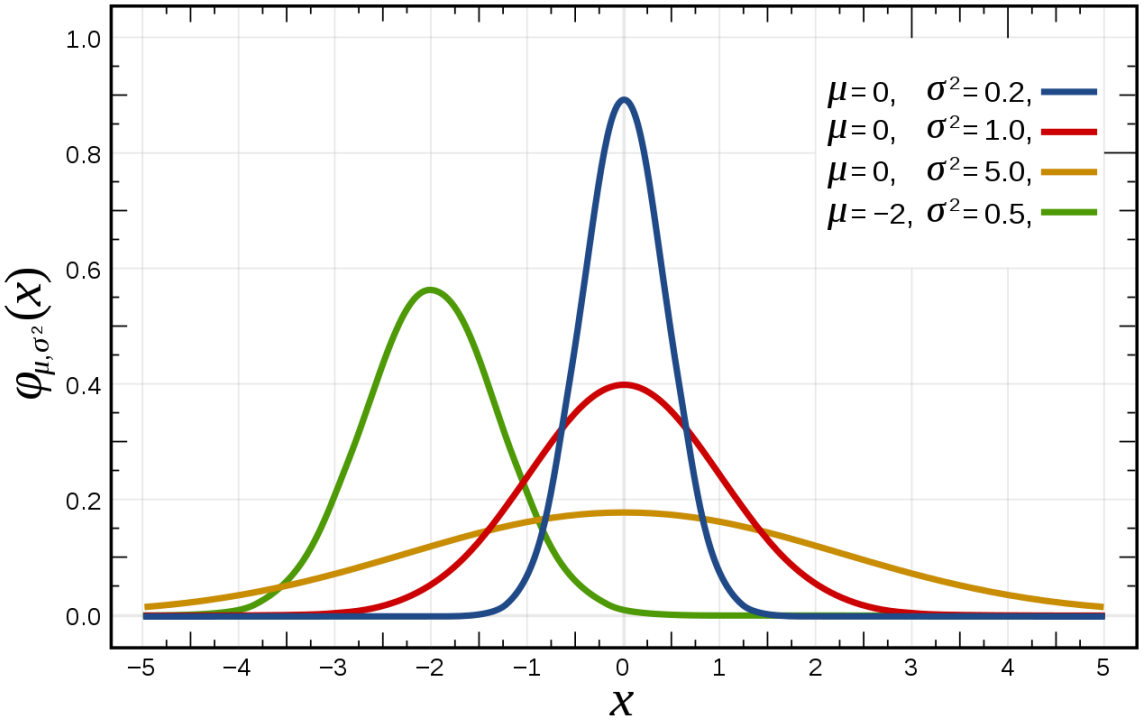
Normální rozdělení [normal distribution, Gaussian distribution]

Příklad: Rozložení výšek studentů ve třídě.

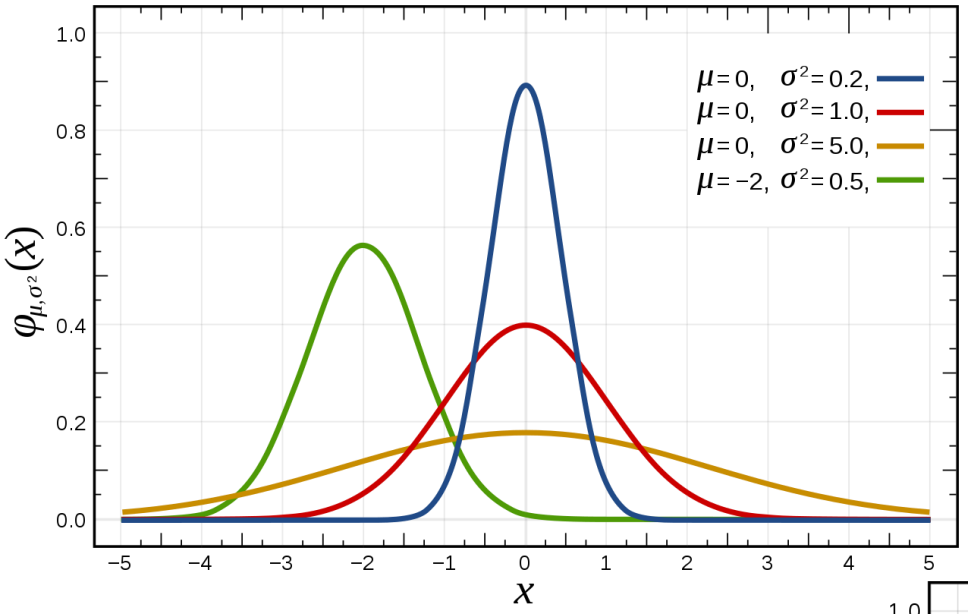
$X \sim N(\mu, \sigma^2)$ *[čti: X má normální rozdělení se střední hodnotou mí a rozptylem sigma na druhou]*

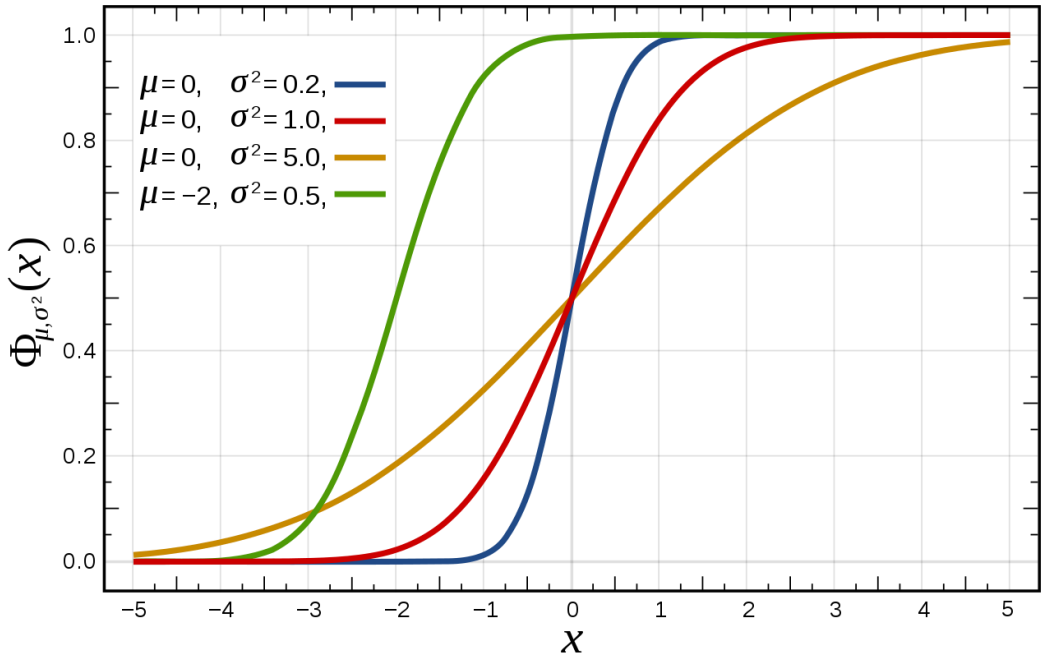
$$f(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} \cdot exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

- Spojitá data
- Symetrické rozdělení
- μ … poloha vrcholu
- σ^2 … šířka zvonu
- Z-rozdělení (STATISTICA)



Normální rozdělení $X \sim N(\mu, \sigma^2)$





Normální rozdělení $X \sim N(\mu, \sigma^2)$

Co modelujeme pomocí normálního rozdělení:

- spojitá data na poměrové stupnici
- spojitá data na intervalové stupnici, pokud je průměr alespoň o několik směrodatných odchylek větší než nula (arbitrární nula!)
- diskrétní data, pokud má X dostatek různých diskrétních hodnot, např. počet semenáčků v rozmezí alespoň 1 – 30