Pravděpodobnost a Statistika

Alena Černíková

alena.cernikova@ujep.cz

9. listopadu 2026

Podmínky zápočtu a zkoušky

- **Zápočet** – písemka psaná v lednu
  vzorový test bude vyvěšen na mých stránkách
  bude obsahovat
  - otázky na interpretaci popisných statistik
  - otázky na výpočet a interpretaci základních inferenčních metod
  - otázky na výpočrt pravděpodobností
- **Zkouška** – ústní
  dvě až tři losované otázky
  bude se kombinovat schopnost počítat a interpretovat a teoretické znalosti
  zkouška bude u počítačů bez přístupu k internetu
  k dispozici budou **vybrané vzorce** a příkazy z R-ka

Obsah kurzu

1. Definice pravděpodobnosti
2. Pravděpodobnostní rozdělení
3. Úvod do statistiky
4. Popisné statistiky
5. Odhady populačních charakteristik
6. Testování hypotéz
7. Statistické modely

Základní pojmy

- **Náhodný pokus** – pokus konaný za přesně daných podmínek, o němž není dopředu známo jak dopadne
  Př. hod kostkou, měření výšky lidí, výsledek studenta u zkoušky
- **Náhodný jev** – možný výsledek náhodného pokusu
  Př. na kosce padne sudé číslo, výška člověka bude větší než 170 cm, student zkoušku udělá
- **Elementární jev** – nejmenší možné náhodné jevy, které nemohou nastat současně, ale musí nastat vždy alespoň jeden z nich
  Př. na kostce padne 1, 2, 3, 4, 5 nebo 6, výška člověka bude 160 cm, student u zkoušky dostane známku 1, 2, 3 nebo zkoušku neudělá
- Součet všech elementárních jevů je prostor všech možných výsledků náhodného pokusu. Značí se Ω.

Klasická definice pravděpodobnosti

Mějme prostor elementárních náhodných jevů Ω. Označme $\mathcal{A}$ systém množin - algebru na tomto prostoru. **Pravděpodobností** pak nazveme reálnou funkci $P(A)$ definovanou na algebře $\mathcal{A}$ podmnožin prostoru Ω, jestliže platí

$$\begin{aligned} A \in \mathcal{A} &\Rightarrow P(A) \geq 0 \\ A, B \in \mathcal{A}, A \cap B = \emptyset &\Rightarrow P(A \cup B) = P(A) + P(B) \\ P(\Omega) = 1, &\quad P(\emptyset) = 0 \end{aligned}$$

Trojice $(\Omega, \mathcal{A}, P)$ se nazývá klasický pravděpodobnostní prostor.

Poznámka: Algebra je systém množin pro který platí: jestliže máme dvě množiny A a B z tohoto systému, pak do něj patří i jejich sjednocení, jejich průnik a jejich doplňky.

Klasická definice pravděpodobnosti

Algebra $\mathcal{A}$ je množina všech náhodných jevů, které je možné vytvořit z elementářních jevů.

Příklad. *Házíme 10 krát mincí a zajímá nás kolikrát padne orel. Ω je množina všech kombinací, jak mohou jednotlivé hody dopadnout $\Omega = \{0,1\} \times \{0,1\} \times \cdots \times \{0,1\}$, algebra $\mathcal{A}$ jsou pak všechny možné výsledky pokusu: např. v prvním hodu padne orel, celkem padne orel právě 5 krát, orel padne méně než 3x atd. Funkce P pak přiřadí každému z těchto výsledků hodnotu mezi 0 a 1 –* ***pravděpodobnost****.*

Kolmogorova definice pravděpodobnosti

Nechť Ω je neprázdná množina, nechť $\mathcal{A}$ je σ-algebra náhodných jevů definovaných na Ω. **Pravděpodobností** se nazývá reálná funkce $P(A)$ definovaná na $\mathcal{A}$, která pro všechna $A \in \mathcal{A}, A_1, A_2, \ldots \in \mathcal{A}, A_i \cap A_j = \emptyset$ když $i \neq j$ splňuje

$$\begin{aligned} P(A) &\geq 0 \\ P\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{i=1}^{\infty} P(A_i) \\ P(\Omega) = 1, &\qquad P(\emptyset) = 0 \end{aligned}$$

Jedná se o rozšíření klasické definice na spočetné, potažmo nespočetné (reálné) množiny Ω.

Kolmogorova definice pravděpodobnosti

Příklad. *Házíme mincí až do doby, dokud nepadne první orel. Výsledkem pokusu je počet hodů nutný k dosažení tohoto cíle. První orel může padnout v prvním hodu, ve druhém hodu, nebo třeba nemusí padnout vůbec (v každém hodu je pst 1/2, že orel nepadne, takže skutečně nemusí padnout). Jako množinu Ω pak uvažujeme všechny přirozená čísla $\Omega = \mathbb{N} \cup \infty$.*

Příklad. *Uvažujme běžce, který běží lesem vytyčený okruh. V průběhu běhu ztratil kapesník. Je ochoten okruh opustit na jiném místě, než kde začal běžet, ale pouze u nejbližšího možného východu. Cestou k tomuto východu bude ztracený kapesník hledat. Východ se nachází po třetině délky okruhu. Jaká je pst, že kapesník najde, když ho mohl ztratit na libovolném místě se stejnou pravděpodobností? Množinu Ω teď tvoří všechna reálná čísla na intervalu od 0 do délky okruhu.*

Výpočet pravděpodobnosti

Pravděpodobnost množiny A je dána vztahem

$$P(A) = \frac{|A|}{|\Omega|}$$

tj. velikost množiny A (počet možností příznivých jevu A, délka úsečky A) dělená velikostí celého prostoru elementárních jevů Ω (počet všech možností, velikost celé uvažované úsečky).

Výpočet pravděpodobnosti

Příklad. *Házíme dvěma šestistěnnými kostkami, červenou a modrou. Elementární jevy jsou všechny možné dvojice hodnot (1,1), (1,2), (1,3), . . . , (6,5), (6,6). Celkem jich je 36. Nás zajímají pravděpodobnosti následujících náhodných jevů.*

- *Na červené kostce padne liché číslo*
- *Na modré kostce padne číslo dělitelné třemi*
- *Součet na obou kostkách bude větší nebo rovno 10*

Náhodné jevy

- **Jev jistý** Ω – soubor všech elementárních jevů, tj. celý prostor možných výsledků
  Př. na kostce padne číslo od jedné do šesti
- **Jev nemožný** $\emptyset$ – jev, který neobsahuje ani jeden elementární jev
  Př. na kostce padne mínus jedna
- **Jev opačný** k jevu A, tj. $\overline{A}$ – soubor elementárních jevů, které nastanou právě když nenastane jev A
  Př. na kostce padne sudé číslo, a na kostce padne liché číslo
- **Neslučitelné jevy** – jevy A a B jsou neslučitelné, když mají prázdný průnik
  Př. na kostce padne sudé číslo, a na kostce padne 1
- **Podjev** – jev A je podjevem jevu B, když je jeho částí
  Př. na kostce padne liché číslo a na kostce padne 3

Náhodné jevy

Je-li P pravděpodobnost definovaná na algebře $\mathcal{A}$ a jevy $A, B \in \mathcal{A}, A \cap B = \emptyset, A_i \in \mathcal{A}, 1 \leq i \leq n$ pak platí

$$
\begin{aligned}
0 &\leq P(A) \leq 1 \\
P(\overline{A}) &= 1 - P(A) \\
A \subset B &\Rightarrow P(A) \leq P(B) \\
A \subset B &\Rightarrow P(B - A) = P(B) - P(A) \\
P(A \cup B) &= P(A) + P(B) \\
P(A_i \cup A_j) &= P(A_i) + P(A_j) - P(A_i \cap A_j) \\
P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i \leq j \leq n} P(A_i \cap A_j) \\
&\quad + \sum_{1 \leq i \leq j \leq k} P(A_i \cap A_j \cap A_k) \\
&\quad + \ldots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right)
\end{aligned}
$$

Poslední rovnost lze zobecnit na $n = \infty$.

Nezávislost jevů a podmíněná pravděpodobnost

- **Podmíněná pravděpodobnost** – hledáme pravděpodobnost jevu A za podmínky že víme, že nastal jev B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Předpokládáme $P(B) > 0$.
Př. jaká je pst, že na kostce padne 6, když víme, že na na ní padlo sudé číslo?

- **Nezávislost jevů** – jevy A a B jsou nezávislé, když

$$P(A) = P(A|B)$$

nebo jinak zapsáno

$$P(A)P(B) = P(A \cap B)$$

Př. jsou jevy "na červené kostce padne liché číslo" a "na modré kostce padne číslo dělitelné třemi" nezávislé

Nezávislost jevů a podmíněná pravděpodobnost

- **Vzorec pro celkovou pravděpodobnost** – chceme spočítat pst jevu A, když známe pouze podmíněné psti $P(A|H_i)$, kde H_i jsou neslučitelné jevy, jejichž sjednocení je jev jistý, tj. $H_1 \cup H_2 \cup \ldots \cup H_k = \Omega$ a $H_i \cap H_j = \emptyset$ pro všechna i, j

$$P(A) = \sum_{i=1}^{k} P(A|H_i)P(H_i)$$

Příklad. *Zajímá mě pst, že obyvatel ČR dostane chřipku. Vím, jaká je tato pst v jednotlivých krajích. Kraje ale nejsou stejně velké. Pst pro celou republiku dostanu jako součet pstí v jednotlivých krajích, násobených relativní velikostí kraje.*

Nezávislost jevů a podmíněná pravděpodobnost

- **Bayesův vzorec** – jak vypočítat podmíněnou pravděpodobnost $P(A|B)$ ze znalosti $P(B|A)$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

neboli vzorec v obecné podobě

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_{j=1}^{k} P(A|H_j)P(H_j)}$$

pravděpodobnosti $P(H_i)$ se nazývají *apriorní* a pravděpodobnosti $P(H_i|A)$ *aposteriorní*

V čitateli je pravděpodobnost průniku, ve jmenovateli pravděpodobnost podmínky přepsaná pomocí vzorce pro úplnou pravděpodobnost.

Senzitivita a specificita testu

Charakteristiky popisující kvalitu nejčastěji u medicínských testů

- **Senzitivita testu** – pravděpodobnost, že test vyjde pozitivně, pokud je osoba nemocná
  $P(\text{test je pozitivní}|\text{osoba je nemocná})$
- **Specificita testu** – pravděpodobnost, že test vyjde negativně, pokud je osoba zdravá
  $P(\text{test je negativní}|\text{osoba je zdravá})$

Senzitivita a specificita testu

Příklad. *Výzkumu se zúčastnilo 2000 pacientů, z nichž 50 mělo danou nemoc. Všichni podstoupili test na tuto nemoc. Test vyšel pozitivní pro 45 nemocných pacientů a pro 200 zdravých. Spočtěte senzitivitu a specificitu testu a také pravděpodobnost, že člověk bude skutečně nemocný, pokud mu vyjde pozitivní test.*

		Skutečnost Nemocný	Zdravý	Celkem
Test	Pozitivní	45	200	245
	Negativní	5	1750	1755
	Celkem	50	1950	2000

- **Senzitivita testu** – $P(\text{test je pozitivní}|\text{osoba je nemocná}) = 45/50 = 0.9$
- **Specificita testu** – $P(\text{test je negativní}|\text{osoba je zdravá}) = 1750/1950 = 0.897$
- **Jsem nemocný, když mám pozitivní test?** –
  $P(\text{osoba je nemocná}|\text{test je pozitivní}) = 45/245 = 0.184$
  Pomocí Bayesovy věty

$$P(ON|TP) = \frac{P(ON \cap TP)}{P(TP)} = \frac{P(TP|ON)P(ON)}{P(TP|ON)P(ON) + P(TP|OZ)P(OZ)} =$$

$$= \frac{\text{Senzitivita} * \text{podíl nemocných}}{\text{Senzitivita} * \text{podíl nemocných} + (1 - \text{Specificita}) * \text{podíl zdravých}} = \frac{0.9 * 0.025}{0.9 * 0.025 + 0.102 * 0.975} = 0.184$$

Vybrané klasické pravděpodobnostní modely

- **Uspořádaný výběr s vracením**
  Máme urnu, ve které je M rozlišitelných koulí. Náhodně vytáhneme z urny kouli, zapíšeme si její označení a zase jí vrátíme. Tah provedeme m-krát. Záleží-li na pořadí v jakém byly koule taženy, pak prostor elementárních jevů Ω má velikost M^m.
  Př. Uvažujme 8 šachových partií, každá může skončit buď výhrou bílého, výhrou černého nebo remízou. Kolik je možností, jak celý zápas může skončit? $M^m = 3^8$

- **Neuspořádaný výběr s vracením**
  Opět táhnu m krát kouli z urny, ve které je M rozlišitelných koulí a kouli po každém tahu vrátím. Když mi nezáleží na pořadí, pak prostor elementárních jevů Ω má velikost $\binom{M-1+m}{m}$.
  Př. Když mi v oněch 8 šachových partiích nezáleží na pořadí a počítám jenom celkový počet výher bílého, černého a celkový počet remíz.
  $\binom{M-1+m}{m} = \binom{3-1+8}{8} = \binom{10}{8}$

Vybrané klasické pravděpodobnostní modely

- **Uspořádaný výběr bez vracení**
  Z urny, ve které je M rozlišitelných koulí, tahám náhodně m koulí. Koule tam nevracím, ale zaznamenávám pořadí, v jakém byly taženy. Pak prostor elementárních jevů Ω má velikost $\binom{M}{m} m!$.
  Př. Kolika způsoby mohu vybrat 5 hrnků z 10 (hrnky jsou rozlišitelné) a srovnat je na polici? $\binom{M}{m} m! = \binom{10}{5} 5! = \frac{10!}{5!}$
- **Neuspořádaný výběr bez vracení**
  Opět táhnu m krát kouli z urny, ve které je M rozlišitelných koulí. Koule nevracím a nezáleží mi na pořadí, v jakém je táhnu. prostor elementárních jevů Ω má pak velikost $\binom{M}{m}$.
  Př. Kolika způsoby mohu vybrat 5 žáků ke zkoušení z dvaceti přítomných ve třídě? $\binom{M}{m} = \binom{20}{5}$

Vybrané klasické pravděpodobnostní modely

Náhodná procházka

Uvažujme částici, která se pohybuje po celočíselné přímce $\mathbb{Z}$ a označme jako S_k její polohu v čase $k = 0, 1, 2, \ldots$. Předpokládejme, že na začátku je částice v bodě 0, tj. $S_0 = 0$ a je-li v nějakém časové okamžiku k v bodě a, tak v čase $k + 1$ je v bodě $a - 1$ s pravděpodobností 1/2 a v bodě $a + 1$ také s pravděpodobností 1/2.

$$P(S_{k+1} - S_k = 1) = P(S_{k+1} - S_k = -1) = \frac{1}{2}$$

Rozhodování o pohybu částice v každém bodě je náhodné a nezávislé na předchozích krocích. Určeme rozdělení pravděpodobností, kde se bude částice nacházet v čase n.

Vybrané klasické pravděpodobnostní modely

Náhodná procházka

Prostor elementárních jevů je $\Omega = \{0, 1\}^n$ a jeho velikost je $|\Omega| = 2^n$. Stav v čase n je pak jednoznačně určen počtem pohybů doprava. Označme tento počet kroků doprava jako k (nutně $n - k$ kroků musí být doleva) a skončím v bodě $S_n = k - (n - k)$. Pro pravděpodobnost stavu S_n tedy platí

$$P(S_n = 2k - n) = \frac{\binom{n}{k}}{2^n}$$

Bylo zjištěno, že pokud necháme částici "běhat"nekonečně dlouho, pak s pravděpodobností 1 projde každým bodem $a \in \mathbb{Z}$. Budeme-li zkoumat její návrat do bodu 0, pak k němu dojde s pravděpodobností 1, ale střední doba čekání na tento okamžik bude nekonečná.

Náhodná veličina

Veličina, kterou můžeme opakovaně měřit u různých objektů, se nazývá **náhodná veličina**, případně náhodná proměnná. Přesněji:

- **Náhodná veličina** je funkce, která přiřazuje každému elementárnímu náhodnému jevu nějakou (zpravidla číselnou) hodnotu.

Příklad. *Rychlost, za kterou program vypočte danou úlohu.*
Studijní obor, který si zvolil student na vysoké škole.
atd.

Náhodná veličina

Každá náhodná veličina se řídí nějakým pravděpodobnostním rozdělením. Ta dělíme podle typu na

- **Spojitá** – pro číselné proměnné, které teoreticky mohou nabývat libovolné reálné hodnoty z nějakého intervalu, př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní** – pro kategorické proměnné s jasně oddělitelnými kategoriemi, může být i nekonečně mnoho hodnot
  př. binomické, poissonovo, alternativní, . . .
- **Multinomické** – rozdělení popisující chování kategorické proměnné s neuspořádanými kategoriemi

Určit pravděpodobnostní rozdělení znamená definovat pravděpodobnosti elementárních jevů.

Funkce určující rozdělení

Distribuční funkce

- označme X náhodnou veličinu a t hodnotu, jíž může tato veličina nabývat
- distribuční funkce je pravděpodobnost, že náhodná veličina nabude maximálně hodnoty t
- $F(t) = P(X \leq t), t \in \mathbb{R}$
- neklesající, zprava spojitá
- nabývá hodnot mezi 0 a 1 (včetně)
- distribuční funkci má každá veličina nabývající číselných hodnot (diskrétní i spojitá)

Funkce určující rozdělení

Pravděpodobnostní funkce

- označme X náhodnou veličinu a t hodnotu, jíž může tato veličina nabývat
- pravděpodobnostní funkce je pravděpodobnost, že náhodná veličina nabude hodnoty přesně t
- nespojitá, nenulová jen v hodnotách, kterých může náhodná veličina nabývat
- pro hodnoty, jichž veličina nemůže nabývat, má hodnotu 0
- $p(t) = P(X = t), t \in \mathbb{R}$
- definovaná pro diskrétní a multinomické rozdělení

Funkce určující rozdělení

Hustota

- narozdíl od předchozích to není přímo pravděpodobnost, ale k výpočtu pravděpodobností slouží
- definovaná přes distribuční funkci, musí platit

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

- hustota je tedy derivací distribuční funkce $f(t) = \frac{d}{dt}F(t), t \in \mathbb{R}$
- funkce nezáporná
- definovaná pouze pro spojitá rozdělení
- obdoba (zespojitění) pravděpodobnostní funkce
- pravděpodobnost jedné konkrétní hodnoty u spojitého rozdělení je 0

Funkce určující rozdělení

Kvantilová funkce

- inverzní funkce k funkci distribuční
- vstupem je pravděpodobnost (nebo díl dat) a výstupem číselná hodnota, jíž může náhodná veličina nabývat

$$q(p) = x_0 : P(X \leq x_0) = p$$

- hledáme takové číslo x_0, které oddělí $p\%$ nejmenších hodnot
- pomocí této funkce hledám kvantily, např. medián, kvartily, ...
- používá se častěji u spojitých proměnných

Hod kostkou

Příklad. *Určete pravděpodobnostní a distribuční funkci pro hod šestistěnnou kostkou v následujících bodech. Spočtěte a potom nakreslete.*

Hodnota x_i	$P(X = x_i)$	$P(X \leq x_i)$
0	0	0
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
3.5	0	3/6
3.8	0	3/6
4	1/6	4/6
4.5	0	4/6
5	1/6	5/6
6	1/6	1
6.5	0	1
8	0	1

Střední hodnota

- určení středu rozdělení
- definice střední hodnoty pro diskrétní rozdělení

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

– definuje se pomocí pravděpodobnostní funkce
– mají-li všechny možné hodnoty stejnou pravděpodobnost, pak střední hodnota je jejich aritmetický průměr

- definice střední hodnoty pro spojitá rozdělení

$$EX = \int_{-\infty}^{\infty} xf(x)dx$$

– definuje se pomocí hustoty

Rozptyl

- popisuje, jak jednotlivé hodnoty oscilují kolem střední hodnoty
- definice rozptylu pro diskrétní rozdělení

$$\mathrm{Var}(X) = \sum_{i=1}^{n} (X_i - E(X))^2 p_i$$

– mají-li všechny možné hodnoty stejnou pravděpodobnost, pak je rozptyl průměr z druhých mocnin odchylek od průměru

- definice rozptylu pro spojitá rozdělení

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

– proč potřebuji v rozptylu používat druhou mocninu?

Střední hodnota a rozptyl

- Při hodu kostkou mohou nastat elementární jevy $1, 2, 3, 4, 5, 6$. Každá z těchto možností je stejně pravděpodobná ($p_i = 1/6$ pro všechna $i = 1, \ldots, 6$).
- Střední hodnota určuje "střed" rozdělení.

$$E(X) = \sum_{i=1}^{n} X_i p_i = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

může vyjít hodnota, jíž náhodná veličina nemůže nabývat

- Rozptyl určuje variabilitu rozdělení

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} (X_i - E(X))^2 p_i = ((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + \\ &+ (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2)\frac{1}{6} = 2.92 \end{aligned}$$

Vlastnosti střední hodnoty a rozptylu

- Střední hodnota

$$\begin{aligned} E(a + bX) &= a + bEX \\ E(X + Y) &= E(X) + E(Y) \end{aligned}$$

- Rozptyl

$$\begin{aligned} Var(aX + b) &= a^2 Var(X) \\ Var(X + Y) &= Var(X) + Var(Y) + 2cov(X, Y) \end{aligned}$$

kde $cov(X, Y)$ je kovariance počítaná jako
$cov(X, Y) \sum_{i=1}^{n} (X_i - E(X))(Y_i - E(Y))p_i q_i$ nebo
$cov(X, Y) \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))f(x, y)dxdy$

Alternativní rozdělení

Mějme náhodný pokus, který může skončit jedním ze dvou výsledků: úspěch – neúspěch. Výsledek pokusu má alternativní rozdělení.

Značení $Alt(p)$, kde

- p – pravděpodobnost úspěchu

Hodnoty pravděpodobnostní funkce

$$p(X = k) = p^k(1 - p)^{1-k}, k = 0, 1$$

Střední hodnota a rozptyl

$$\mathrm{E(X)} = \mathrm{p}, \qquad \mathrm{Var(X)} = \mathrm{p(p - 1)}$$

Binomické rozdělení

Mějme náhodný pokus, který může skončit jedním ze dvou výsledků: úspěch – neúspěch. Opakujme tento pokus mnohokrát a počítejme počet úspěchů. Počet úspěchů má binomické rozdělení.

Značení $Bi(n, p)$, kde

- n – počet pokusů,
- p – pravděpodobnost úspěchu

Hodnoty pravděpodobnostní funkce

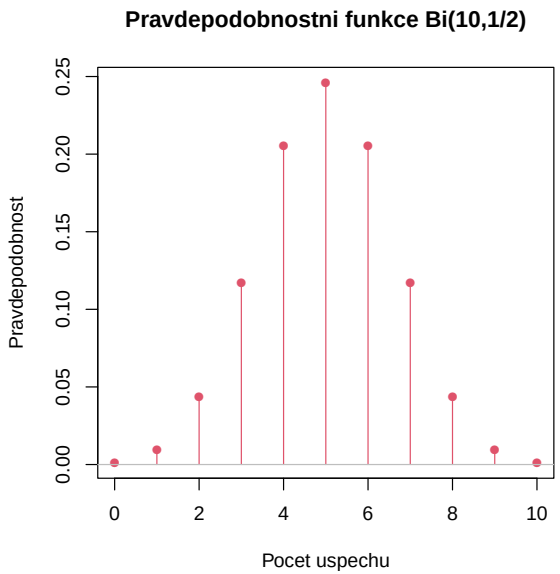
$$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

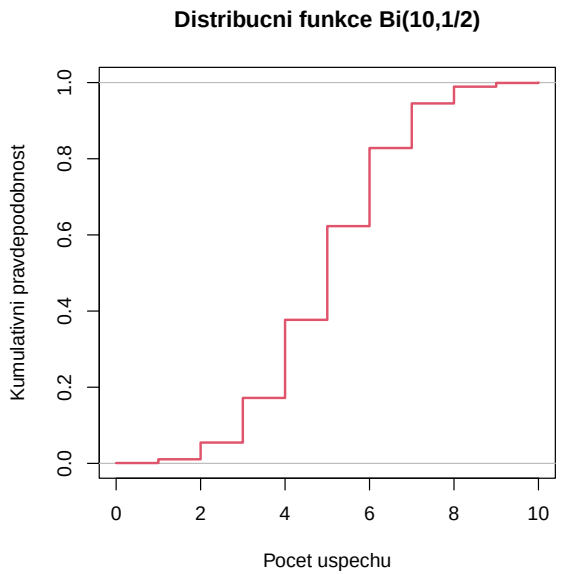
Střední hodnota a rozptyl

$$\mathrm{E(X)} = \mathrm{np}, \qquad \mathrm{Var(X)} = \mathrm{np(p - 1)}$$

Binomické rozdělení

Příklad. *Házíme 10x mincí a počítáme, kolikrát padla panna. Počet pokusů je* $n = 10$*, pravděpodobnost úspěchu* $p = 1/2$*: rozdělení Bi*$(10, 1/2)$*.*
Pravděpodobnostní a distribuční funkce.





Střední hodnota a rozptyl

$$E(X) = np = 10\frac{1}{2} = 5, \qquad Var(X) = np(1-p) = 10\frac{1}{2}\frac{1}{2} = 2.5$$

Poissonovo rozdělení

Příklad. *Sledujeme počet nehod na křižovatce v průběhu jednoho dne. Za normálních okolností nenastane ani jedna nehoda, nebo nastane jedna, maximálně 2 nehody. Ale může se stát, že při náledí jich nastane klidně i 10. Tato veličina má Poissonovo rozdělení.*

Značení $Po(\lambda)$, kde

- λ – parametr rozdělení, intenzita

Hodnoty pravděpodobnostní funkce pro $k = 0, 1, 2, \ldots$

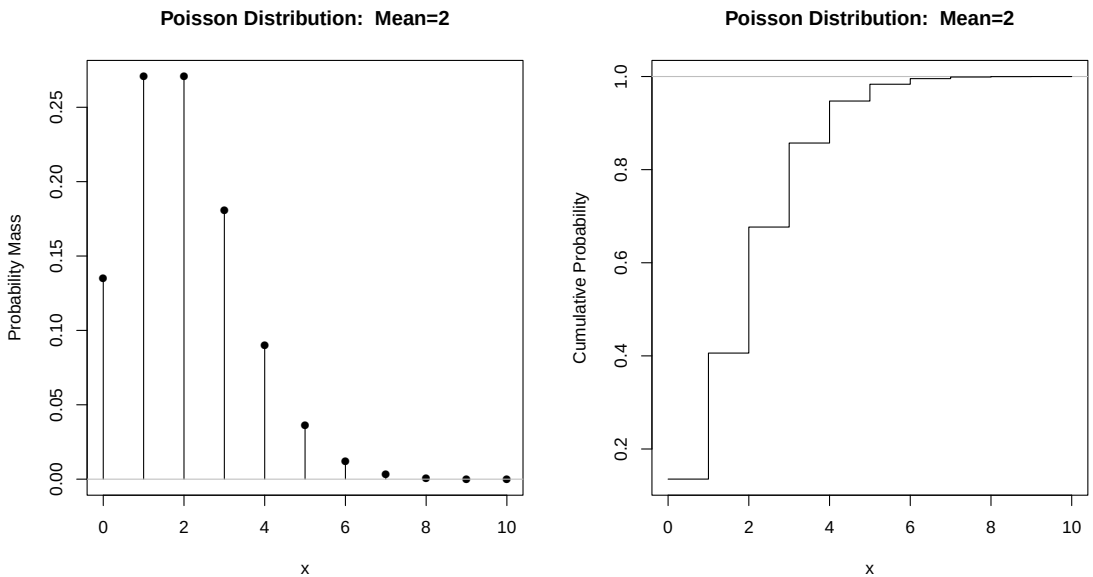
$$p(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Střední hodnota a rozptyl

$$\mathrm{E}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda$$

Poissonovo rozdělení

Pravděpodobnostní a distribuční funkce Poissonova roydělení s parametrem $\lambda = 2$.



Předpokládejme binomická rozdělení $Bi(n, p_n)$, kde $np_n \to \lambda$, pak tato binomická rozdělení konvergují k rozdělení Poissonovu s parametrem λ

Hypergeometrické rozdělení

Př. Uvažujme urnu, ve které máme N koulí, z toho W jich je bílých a zbytek (B) černých. Z urny postupně vytáhneme n koulí bez vracení. Náhodná veličina, která počítá počet bílých koulí mezi vytaženými má hypergeometrické rozdělení.

Značení $Hy(W, B, n)$, kde

- W – počet bílých koulí v urně
- B – počet černých koulí v urně
- n – počet tažených koulí

Hodnoty pravděpodobnostní funkce

$$p(X = k) = \frac{\binom{W}{k}\binom{B}{n-k}}{\binom{N}{n}}$$

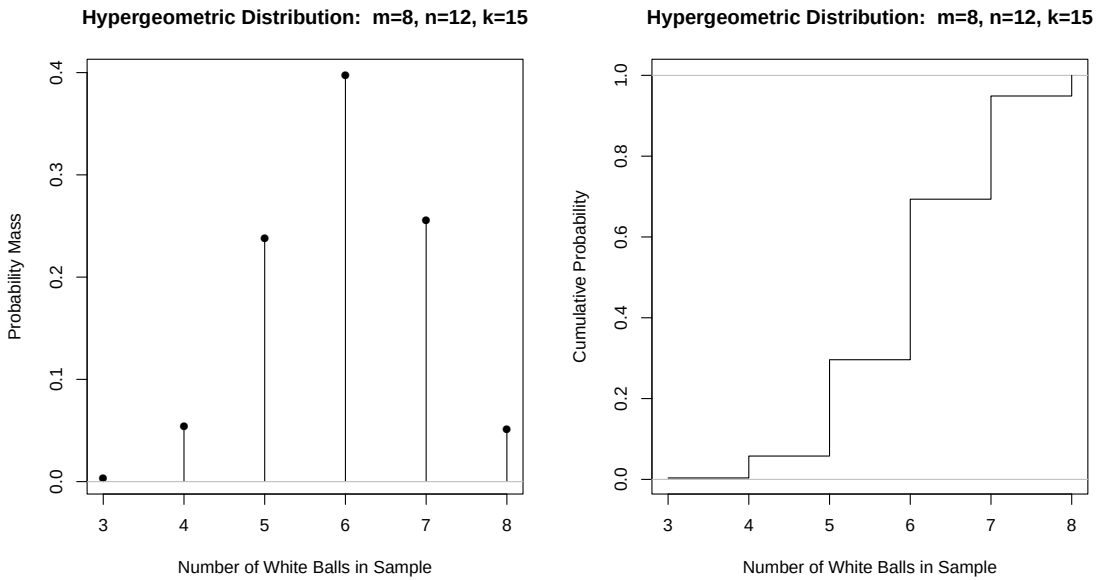
Střední hodnota a rozptyl

$$\mathrm{E(X)} = \frac{\mathrm{nW}}{\mathrm{N}}, \qquad \mathrm{Var(X)} = \mathrm{n}\frac{\mathrm{W}}{\mathrm{N}}\left(1 - \frac{\mathrm{W}}{\mathrm{N}}\right)\left(\frac{\mathrm{N} - \mathrm{n}}{\mathrm{N} - 1}\right)$$

Hypergeometrické rozdělení

Příklad. *Uvažujme 20 koulí v urně, z toho 8 bílých a z urny vytáhneme 15 koulí .*

Pravděpodobnostní a distribuční funkce.



Střední hodnota a rozptyl

$$E(X) = \frac{nW}{N} = 6, \qquad Var(X) = n\frac{W}{N}\left(1 - \frac{W}{N}\right)\left(\frac{N-n}{N-1}\right) = 0.95$$

Normální rozdělení

Jedná se o "hezké"rozdělení, se kterým se dobře pracuje. Toto rozdělení má výška lidí určitého věku, IQ, ... Ve statistice se nejčastěji používá standardní normální rozdělení $N(0, 1)$

Značení $N(\mu, \sigma^2)$, kde

- μ – střední hodnota
- σ^2 – rozptyl

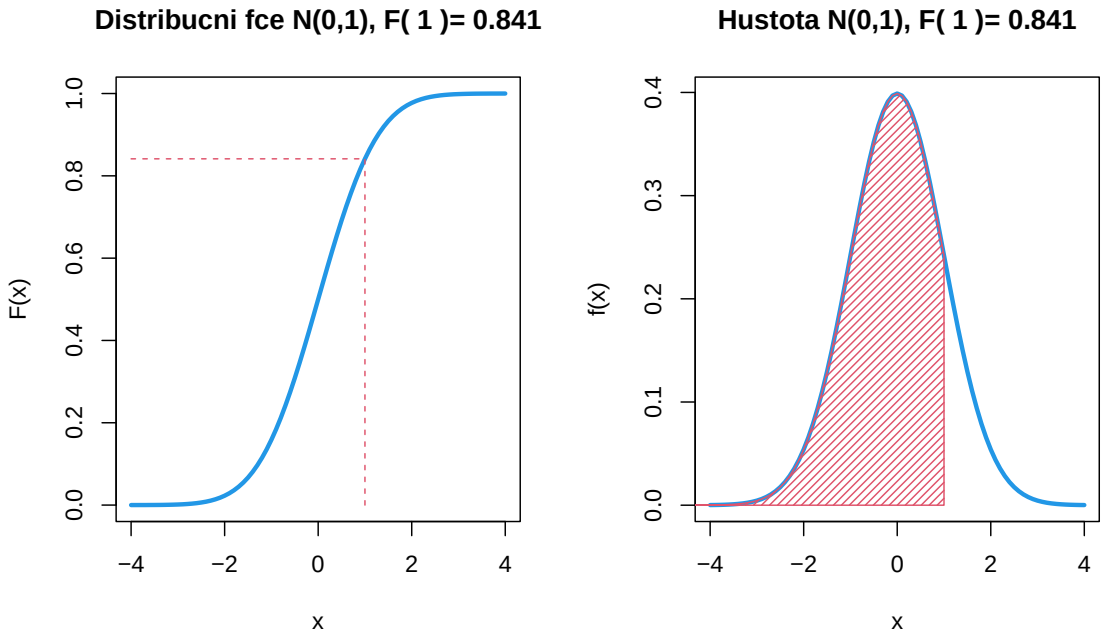
Hustota normálního rozdělení má tvar

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

Je to tak zvaná Gaussova křivka.

Normální rozdělení

Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota. Hustota a distribuční funkce.



Předpokládejme binomické rozdělení $Bi(n,p)$, kde $0.1 \leq p \leq 0.9$, pak pro $n \to \infty$ toto rozdělení konverguje k normálnímu s parametry np, $np(1-p)$.

Normální rozdělení

Nejpoužívanějším normálním roydělením je **standardní normální rozdělení** s nulovou střední hodnotou a jednotkovým rozptylem $N(0,1)$

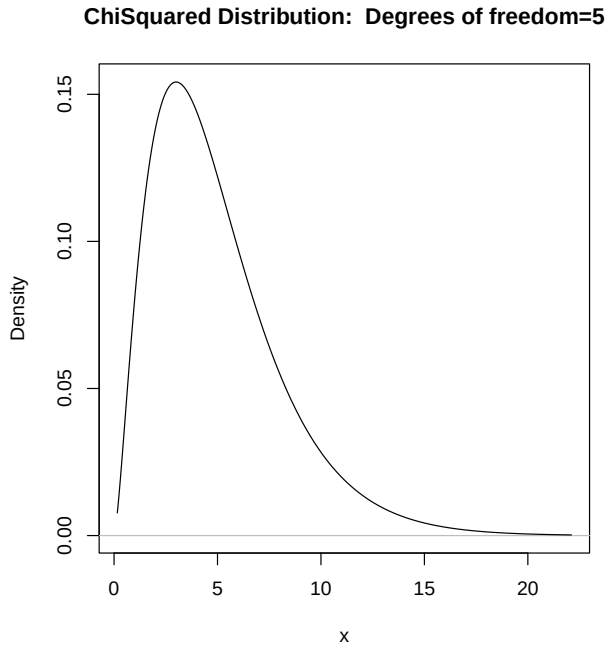
- distribuční funkce standardního normálního rozdělení se značí Φ
- každé obecné normální rozdělení lze na to standardní převést
- označme veličiny $X \sim N(\mu, \sigma^2)$ a $U \sim N(0,1)$

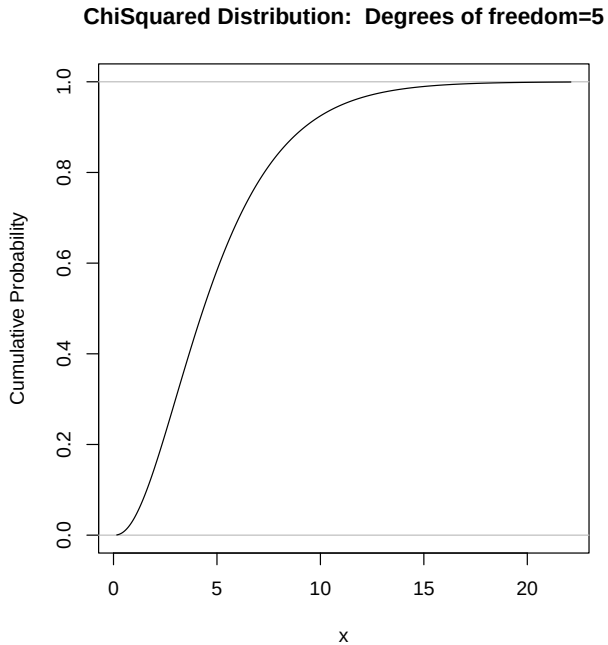
$$U = \frac{X - \mu}{\sigma}$$

- kritické hodnoty jsou tabelované pouze pro $N(0,1)$

χ^2-rozdělení

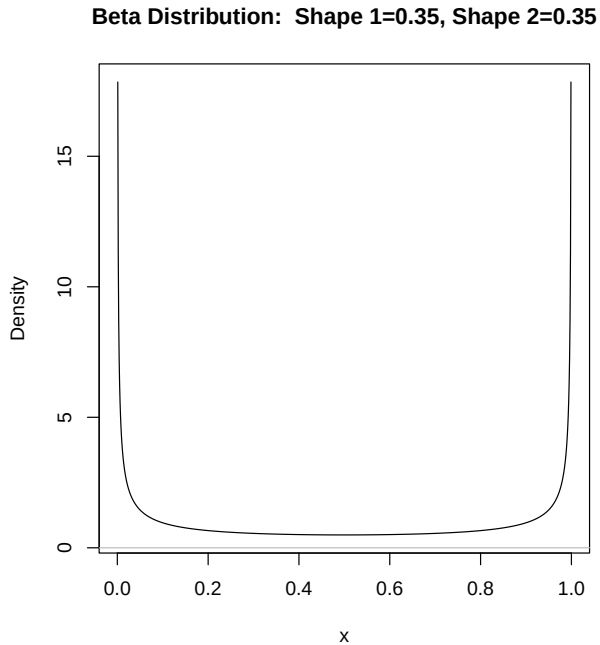
Rozdělení kvadratických forem. Náhodná veličina $Y = X_1^2 + X_2^2 + \cdots + X_n^2$, kde $X_i \sim N(0, 1)$ jsou nezávislé, má χ^2-rozdělení o n stupních volnosti. Dále je to rozdělení některých testových statistik, zejména těch, týkajících se rozptylu. Hustota a distribuční funkce χ^2-rozdělení o 5 stupních volnosti

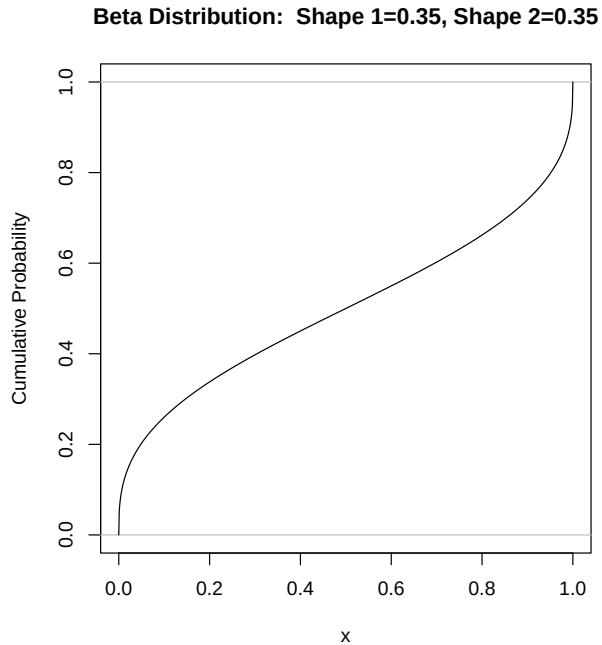




Beta rozdělení

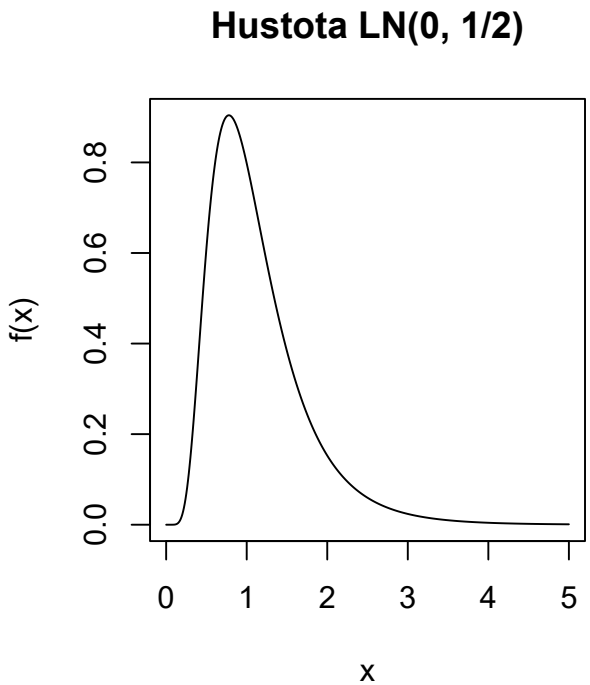
Rozdělení pravděpodobností nějakého jevu. Např. sledujeme pravděpodobnost, že vybryný člověk má nebo nemá nějakou nemoc. Rozdělení má 2 tvarové parametry, které určují, jak vypadají pravděpodobnosti u 0 a 1. Hustota a distribuční funkce Beta rozdělení s parametry 0.35 a 0.35

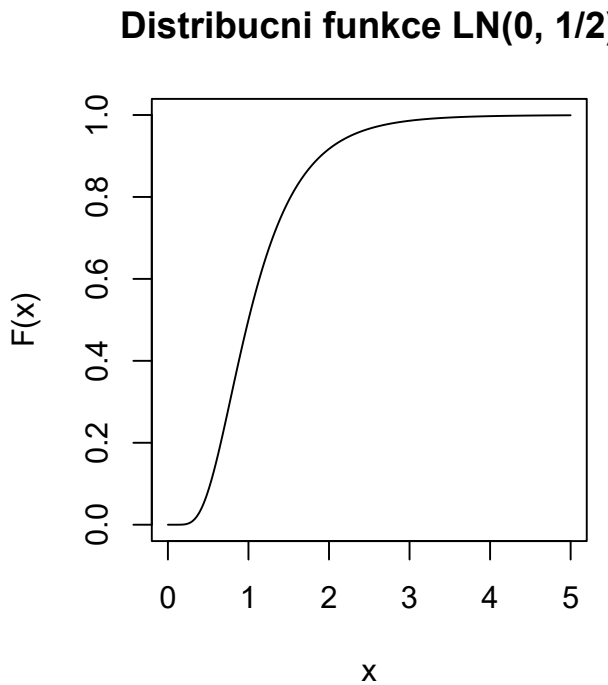




Logaritmické normální rozdělení

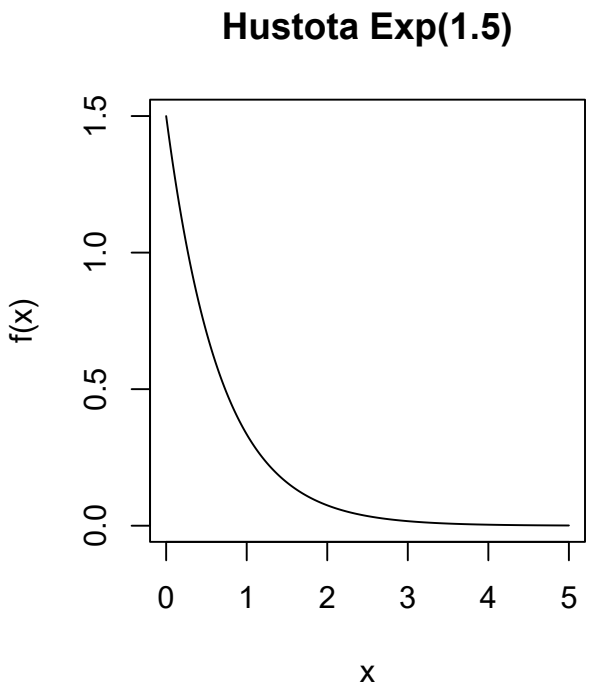
Rozdělení náhodné veličiny, jejíž přirozený logaritmus má normální rozdělení s parametry μ a σ

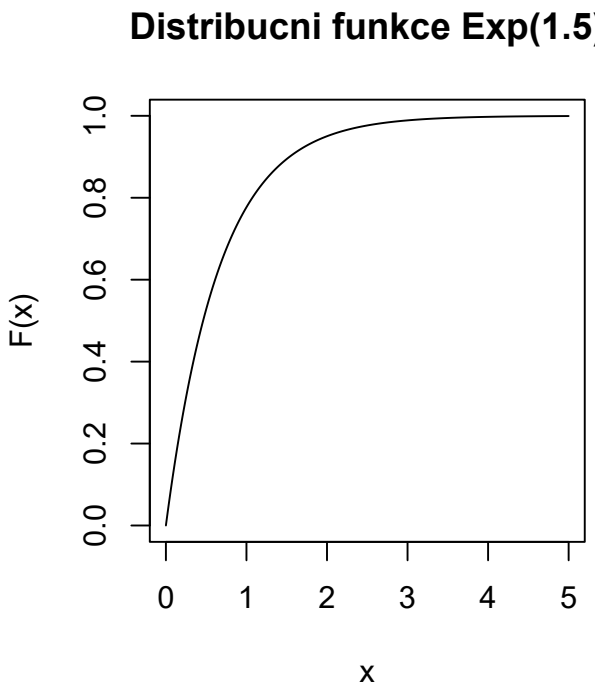




Exponenciální rozdělení

Toto rozdělení popisuje dobu čekání na určitou událost. Velice často se pojí s Poissonovým rozdělením. Když počet událostí za časovou jednotku má Poissonovo rozdělení, tak doba mezi jednotlivými událostmi mívá exponenciální rozdělení s parametrem intenzity λ.





Co je statistika

Statistika je přesná věda o nepřesných číslech.

Zkoumáme náhodnou veličinu na nějaké populaci. Celou populaci změřit neumíme. Uděláme náhodný výběr, na kterém změříme sledovanou veličinu a na základě náhodného výběru děláme závěry pro celou populaci.

Příklad. *Zajímá nás průměrná výška dospělých lidí v celé České republice. Všechny dospělé lidi změřit nemumíme, uděláme náhodný výběr o cca 200 lidech a na základě získaných výsledků se snažíme celkovou průměrnou výšku odhadnout. Průměrná výška pro těchto 200 lidí vyšla 175 cm.*

Co je statistika

- **Nahodná veličina** – jakákoliv veličina, kterou měříme, zde výška
- **Populace** – soubor, pro nejž chceme udělat nějaký závěr, zde všichni dospělí obyvatelé České republiky
- **Náhodný výběr** – v porovnání s populací malý soubor pozorování, jde o nezávislé, stejně rozdělené náhodné veličiny, zde výběr 200 lidí
- **Populační charakteristika** – charakteristika popisující populaci, zde populační průměr
- **Výberová charakteristika** – charakteristika spočítaná na výběru pomocí níž odhadujeme populační ekvivalent, zde výběrový průměr.

Náhodný výběr

Náhodný výběr musí být reprezentativní

- ani sebelepší analýzy nepomohou, pokud nemáte dobrý náhodný výběr
- příklad z prezidentských voleb v USA v roce 1936
  - proti sobě stáli Franklin D. Roosevelt (demokrat) a Alf Landon (republikán)
  - časopis Literary Digest realizoval obří anktetu, s návratností 2.3 milionu odpovědí
  - odhad časopisu byl, že vyhraje Alf Landon (57% hlasů)
  - George Gallup realizoval náhodný výběr o rozsahu 50 tisíc respondentů
  - odhad Gallupa byl, že zvítězí Roosevelt (56% hlasů)
  - reálně zvítězil Roosevelt se ziskem 60.8% hlasů

Náhodný výběr

Příklady reprezentativních výběrů

- **náhodný výběr** ze seznamu
- **systematický výběr** ze seznamu: např. vyberu každého desátého
- **stratifikovaný výběr**: rozdělím populaci podle určitého kritéria na strata (např. podle věku) a v každém stratu dělám náhodný výběr; strata jsou vnitřně homogení
- **vícestupňový výběr**: rozdělím populaci na podobně složené celky, které jsou vnitřně heterogrenní, např. okresy a z nich udělám náhodný výběr, kam půjdu dotazovat; př.: okresy $\rightarrow$ školy $\rightarrow$ třídy
- **kvótní výběr**: dodržuji rozložení hlavních demografií, např. pohlaví, věk vzdělání
  v praxi často používaný, ovšem **nenáhodný**

Máme-li náhodný výběr, je potřeba

- určit, pro jakou populaci je reprezentativní
- vyčistit data, aby byla použitelná pro analýzy
- nasbíraná data **"popsat"**
- pod pojmem popis dat rozumíme vytvořit číselný a grafický přehled, ve kterém shrneme, co jsme naměřili

Problémy v datech

Jaké problémy můžeme potkat a jak je řešit

- **Chybějící pozorování**
  snažíme se, aby jich bylo co nejméně,
  když jich je málo, tak pracujeme bez nich – většina statistických metod implementovaných v různých softwarech si s tím poradí
  je možné je doplnit na základě nějakého modelu (*imputation*)
- **Odlehlé hodnoty**
  kontrola, zda nedošlo k chybě měření
  pokud ne, tak z popisných statistik se většinou nevynechávají, ale je dobré zmínit, že se jedná o odlehlé hodnoty
  pro popis proměnné je pak lépe zvolit ukazatele necitlivé na odlehlé pozorování
  ze složitějších analýz se často vynechávají

Typy proměnných

Abychom správně určili, které charakteristiky máme pro proměnnou počítat, je třeba nejprve určit typ proměnné.

- **Číselné proměnné** – pr. výška, váha, věk, atd.
  - **Spojité** – nabývají všech reálných hodnot z určitého intervalu, př. váha výška
  - **Diskrétní** – nabývají oddělených hodnot, např. počet
- **Kategorické proměnné** – pr. barva, kraj, povolání, nebo taky známka ve škole, číslo, které padne na kostce, atd.
  - **Nominální** – neuspořádané, př. barva, kraj
  - **Ordinální** – uspořádané, př. známka, číslo na kostce

Popisné statistiky

Jak popisujeme jednotlivé typy proměnných

- **Číselné proměnné**
  - popisné statistiky polohy – průměr, medián, vybrané percentily (kvartily, extrémy)
  - popisné statistiky variability – rozptyl, směrodatná odchylka, mezikvartilové rozpětí, koeficient variace
  - popisné statistiky tvaru rozdělení – šikmost, špičatost
  - grafické charakteritiky – krabicový graf, histogram
- **Nominální proměnné**
  - číselné charakteritiky – absolutní a relativní četnosti
  - grafické charakteristiky – sloupcový a koláčový graf
- **Ordinální proměnné**
  - lze použít jak průměr, medián atd.
  - a pro malé počty kategorií i absolutní a relativní četnosti

Popisné statistiky polohy

Potřebujeme stručně popsat, **jakých hodnot veličina nabývá**

- hledáme centrální hodnotu (průměr, medián)
- hledáme interval, na němž hodnoty leží (extrémy)
- hledáme speciální dělící body tohoto intervalu, např. chceme znát 10% největších hodnot (kvartily, percentily)
- hledáme často se opakující hodnoty (modus)

Popisné statistiky polohy

Příklad. *Mějme náhodný výběr 18-ti dospělých lidí a předpokládejme, že jsme u nich naměřili výšky* $176, 184, 167, 193, 174, 182, 181, 179, 187, 165, 168, 172, 184, 178, 160, 168, 171, 159$.
Spočtěme průměr, medián, kvartily a extrémy. Pokusme se určit i modus.

Popisné statistiky polohy

Označme

- X měřenou náhodnou veličinu
- n počet pozorování (naměřených hodnot)
- $X_1, X_2, X_3, \ldots, X_n$ jednotlivé realizace náhodné veličiny, tj. naměřené hodnoty
- $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$ naměřené hodnoty uspořádané podle velikosti
- $\overline{X}$ výběrový průměr náhodné veličiny X

Popisné statistiky polohy

Výběrový průměr

- sečtu všechny hodnoty a vydělím počtem pozorování

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- výběrový průměr je nestranným odhadem střední hodnoty

$$\mathrm{E}\overline{X} = \mathrm{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}X_i = \mathrm{E}X$$

- rozptyl výběrového průměru

$$\mathrm{Var}\overline{X} = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathrm{Var}X_i}{n^2} = \frac{\mathrm{Var}X}{n}$$

protože náhodné veličiny X_i jsou nezávislé

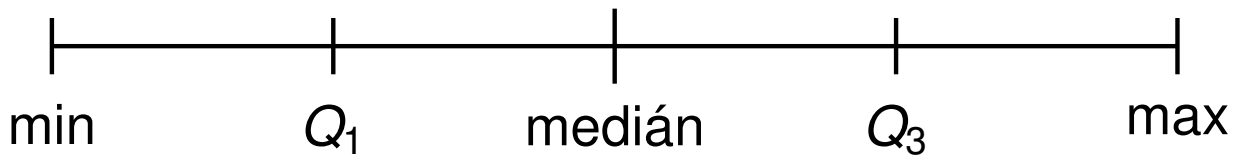
Popisné statistiky polohy

Výběrové kvantily / percentily

- **medián** – hodnota prostřední podle velikosti (při lichém n) nebo průměr prostředních dvou hodnot (při sudém n)
- **kvartily** – dolní kvartil Q_1 odděluje čtvrtinu nejmenších hodnot, horní kvartil Q_3 čtvrtinu největších hodnot
- **extrémy** – minimum a maximum

Popisné statistiky polohy

- Všechny **percentily**, např. medián, kvartily, extrémy, decily, atd., se počítají z uspořádané řady (srovnám naměřené hodnoty podle velikosti).
- Každý percentil odděluje x% nejmenších hodnot (např. minimum $\sim$ 0%, medián $\sim$ 50%).



Popisné statistiky polohy

Výpočet kvartilů podle **R**

Obecný p-tý percentil je vážený průměr dvou sousedních uspořádaných hodnot.
Označme

- p – díl dat, které chcete p-tým percentilem oddělit
  číslo mezi 0 a 1
- q – koeficient, kterým se násobí uspořádané hodnoty do váženého průměru

$$p - ty\ percentil = (1 - q)X_{(k)} + qX_{(k+1)}$$

$$k = \lfloor 1 + (n-1)p \rfloor$$

$$q = 1 + (n-1)p - k$$

Popisné statistiky polohy

Modus má dvě podoby podle typu proměnné

- číselná diskrétní proměnná s málo hodnotami: modus je nejčastěji se opakující hodnota
  takto se definuje modus i u kategorických proměnných
- číselná spojitá proměnná nebo i diskrétní proměnná s neopakujícími se hodnotami: modus je vrchol rozdělení
  - modus se odhaduje přes jádrový odhad hustoty
  - kolem histogramu se nakreslí hladká křivka a hledáme, ve kterém bodě má vrchol
  - křivka by měla být dostatečně hladká, ale ne moc

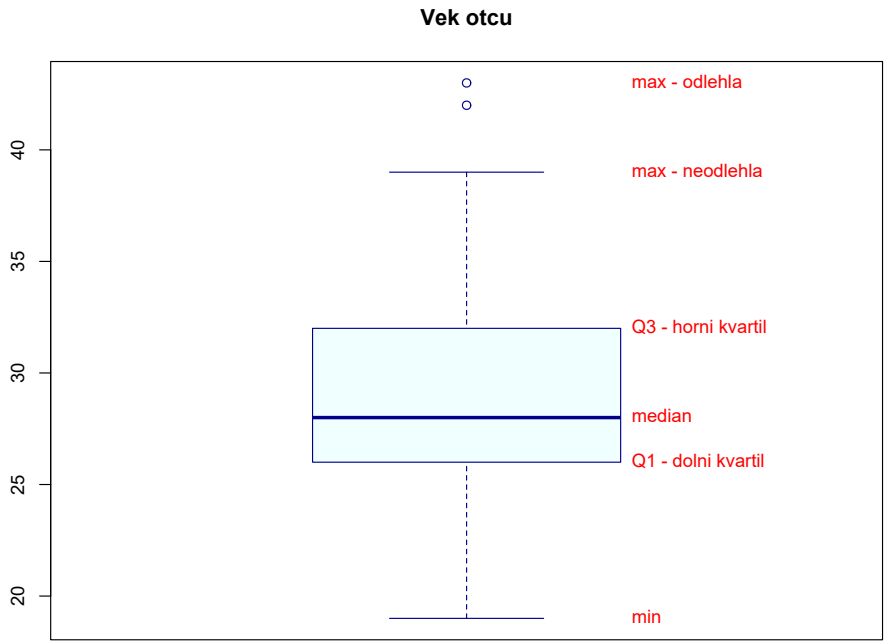
Grafické popisné statistiky

Pro popis číselné proměnné se používají 2 typy grafů

- **Krabicový graf**
  - zobrazeny vybrané percentily: medián (tlustá čára uprostřed krabice) a kvartily (hranice krabice)
  - dobře vidíme, kde se nachází prostřední polovina dat
  - tykadla dosahují k nejvzdálenějšímu neodlehlému pozorování
  - odlehlá pozorování se vyznačují zvlášť
  - *odlehlé pozorování* je takové, které je od bližšího kvartilu (hranice krabice) dál než jeden a půl násobek mezikvartilového rozpětí $1.5 \times (Q_3 - Q_1)$

Grafické popisné statistiky

Ukázka grafů – boxplot



Grafické popisné statistiky

Pro popis číselné proměnné se používají 2 typy grafů

- **Histogram**
  - jeho tvar závisí na počtu sloupců
  - počet sloupců se určuje na základě počtu pozorování
  - pravidlo určující optimální počet sloupců je *Sturgesovo pravidlo*

    $$k = 1 + 3.32 \log_{10}(n)$$

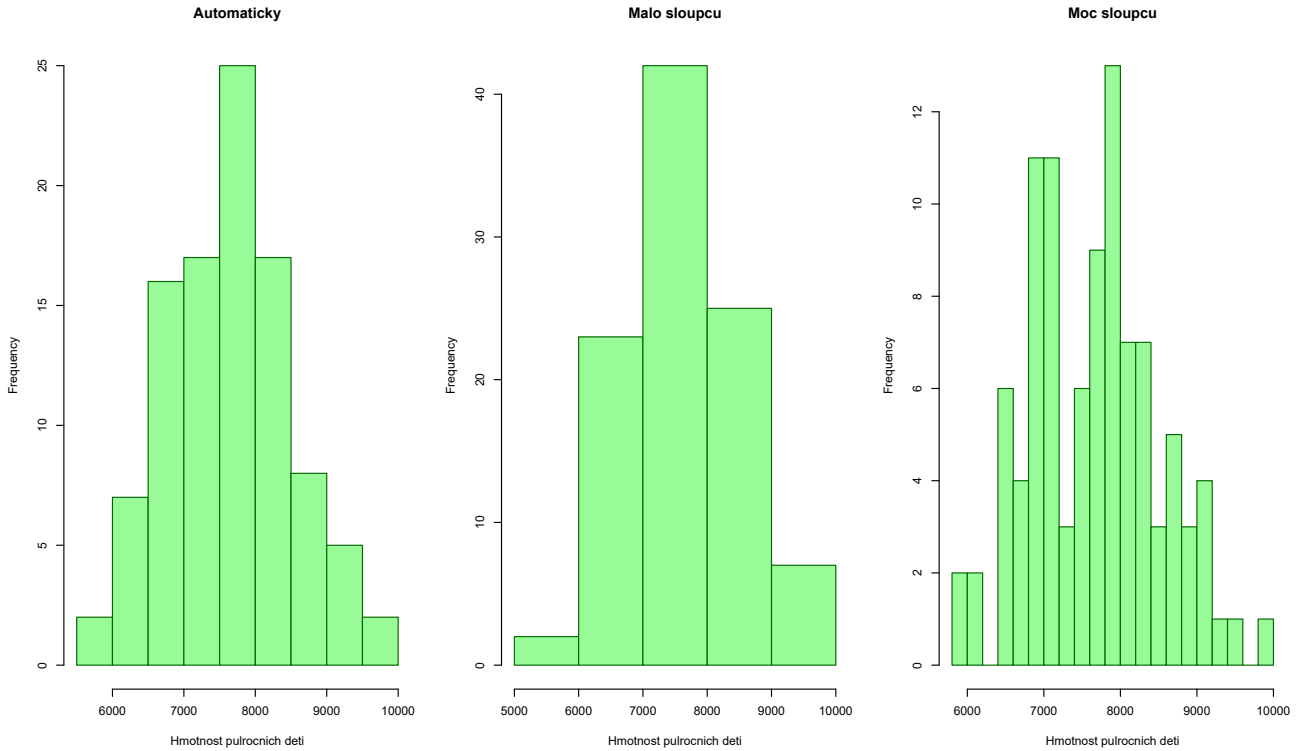
    kde n je počet pozorování
  - pravidlo určující optimální šířku sloupce je *Scottovo pravidlo*

    $$h = \frac{3.49s}{\sqrt[3]{n}}$$

    kde s je výběrová směrodatná odchylka a n je počet pozorování

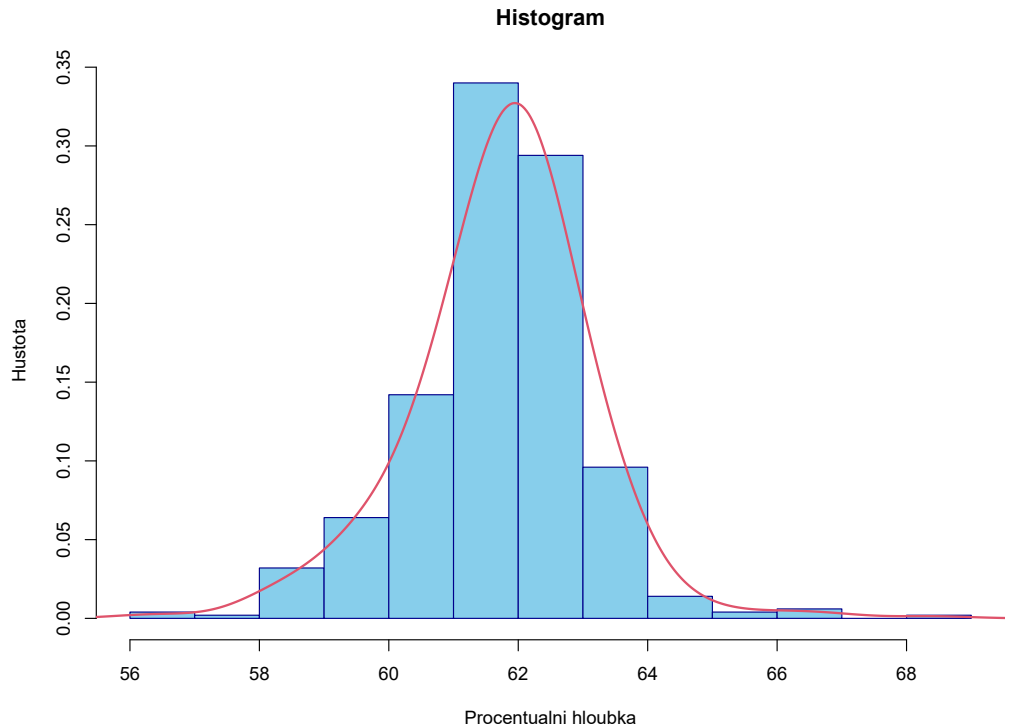
Grafické popisné statistiky

Ukázka grafů – histogram



Grafické popisné statistiky

Ukázka grafů – histogram s jádrovým odhadem hustoty



Grafické popisné statistiky

Jak interpretovat grafy

- **boxplot**
  - Je medián uprostřed boxu?
  - Jak jsou tykadla dlouhá a zda jsou stejně dlouhá?
  - Mám v datech odlehlá pozorování?
  - Mohu poznat (ne)symetrii rozložení hodnot
- **histogram**
  - Který sloupec je nejvyšší?
  - Kde leží nejvyšší sloupec (ve středu grafu)?
  - Kolik vidím na grafu vrcholů?
  - Jak jsou sloupce vysoké relativně jeden k druhému?
  - Mám graf protažen do kraje?

Popisné statistiky variability

- Výběrový rozptyl a **směrodatná odchylka**

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}, \qquad \mathrm{sd}(X) = \sqrt{\mathrm{Var}X}$$

  - základní charakteristika variability
  - nestranný odhad populačního rozptylu
  - směrodatná odchylka je ve stejných jednotkách jako proměnná
  - popisuje, jak moc se pozorování liší od průměru
  - citlivé na odlehlé hodnoty
  - u normálního rozdělení se v intervalu $\overline{X} \pm \mathrm{sd}(X)$ nachází cca $2/3$ pozorování

Popisné statistiky variability

Výběrový rozptyl je **nestranný** odhad populačního rozptylu

- Odhad T parametru θ je **nestranný**, když

$$\mathrm{E}(\mathrm{T}) = \theta$$

- rozdíl $\mathrm{E}(\mathrm{T}) - \theta$ se nazývá vychýlení
- důkaz

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) &= \mathrm{E}\left(\sum_{i=1}^{n}(X_i - \mathrm{E}X - \overline{X} + \mathrm{E}X)^2\right) = \\
&= \mathrm{E}\left(\sum_{i=1}^{n}(X_i - \mathrm{E}X)^2 - 2(\overline{X} - \mathrm{E}X)\sum_{i=1}^{n}(X_i - \mathrm{E}X) + n(\overline{X} - \mathrm{E}X)^2\right) = \\
&= \mathrm{E}\left(\sum_{i=1}^{n}(X_i - \mathrm{E}X)^2 - n(\overline{X} - \mathrm{E}X)^2\right) = \\
&= \sum_{i=1}^{n}\mathrm{E}(X_i - \mathrm{E}X)^2 - n\mathrm{E}(\overline{X} - \mathrm{E}X)^2 = n\mathrm{Var}X - n\frac{\mathrm{Var}X}{n} = (n-1)\mathrm{Var}X
\end{aligned}$$

Popisné statistiky variability

- **Mezikvartilové rozpětí**

$$IQR(X) = Q_3 - Q_1$$

kde Q_3 je třetí kvartil a Q_1 je první kvartil

  - šíře intervalu, kde je prostřední polovina dat
  - výška krabice v krabicovém grafu
  - **robustní** (necitlivé) vůči odlehlým hodnotám

Popisné statistiky variability

- **Variační koeficient**

$$CV(X) = \frac{sd(X)}{\overline{X}}$$

  - porovnává velikost variability (směrodatné odchylky vůči průměru)
  - kolikrát je směrodatná odchylka větší než průměr
  - bezrozměrná veličina
  - často se udává v procentech
  - hodí se pro porovnání variability různých veličin mezi sebou

Popisné statistiky variability

mediánová absolutní odchylka od mediánu

$$MAD = \text{median}\frac{(|X_i - \widetilde{X}|)}{0.6745}$$

- máte-li v datech odlehlé hodnoty, pak počítat s odchylkou od průměru (rozptyl) není dobré
- práce s odchylkou od mediánu je lepší
- konstanta volena tak, aby pro normální rozdělení byla MAD srovnatelná s průměrem

Popisné statistiky tvaru rozdělení

Základem pro výpočet jsou tak zvané **Z-skóry**

$$Z_i = \frac{X_i - \overline{X}}{\mathrm{sd(X)}}$$

- standardizované proměnné
- jejich průměr je 0 a rozptyl 1
- 99% z nich se nachází v intervalu $< -3, 3 >$
- používají se kdekoliv, kde chci srovnat proměnné na stejnou škálu

Popisné statistiky tvaru rozdělení se neurčují u vícevrcholových rozdělení

Popisné statistiky tvaru rozdělení

- **Šikmost** – průměr ze třetích mocnin z-skórů

$$\text{Skew}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\text{sd}(X)}\right)^3 = \frac{\sum_{i=1}^{n} z_i^3}{n}$$

  - hodnotí symetrii dat
    - nula . . . symetrická data
    - záporná hodnota . . . protažení doleva
    - kladná hodnota . . . protažení doprava
  - rozdělení přibližně symetrické má šikmost v rozmezí $< -0.25; 0.25 >$

Popisné statistiky tvaru rozdělení

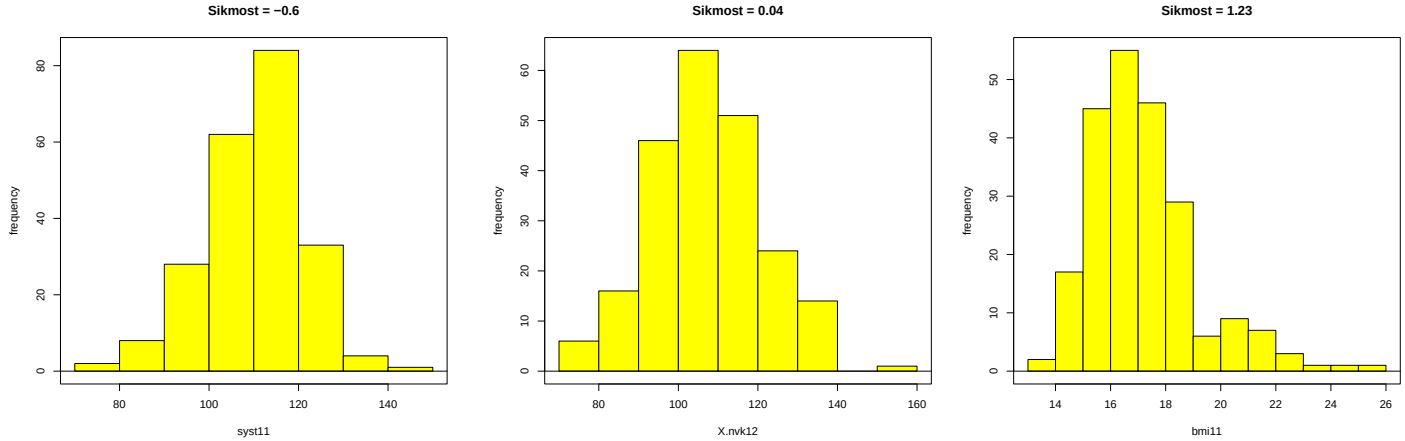
- **Špičatost** – průměr ze čtvrtých mocnin z-skórů mínus 3

$$\text{Kurt}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\text{sd}(X)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

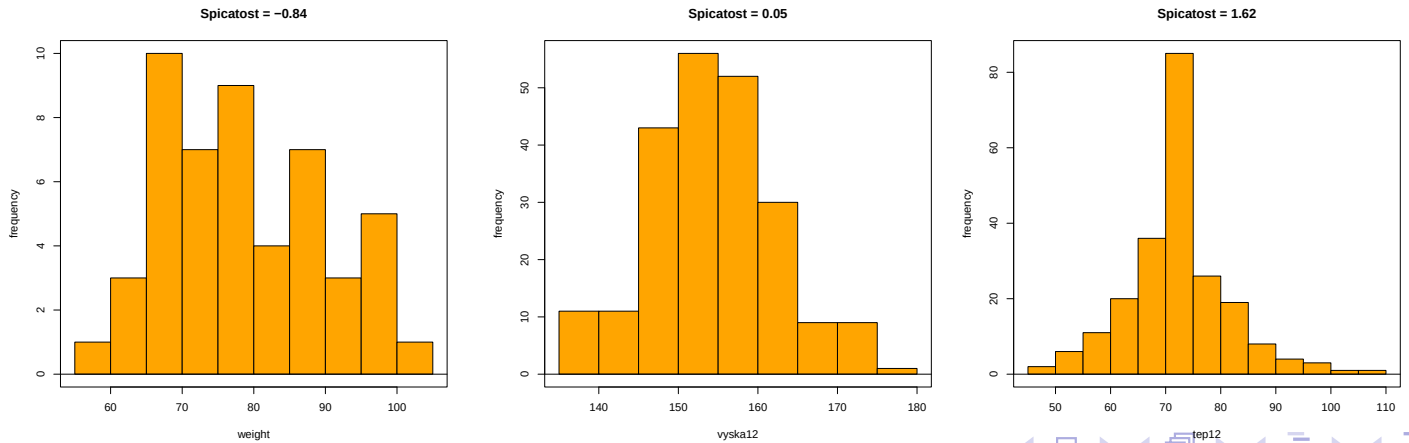
  - porovnává špičatost vůči normálnímu rozdělení
    - nula . . . normální rozdělení
    - záporná hodnota . . . plošší než normální rozdělení
    - kladná hodnota . . . špičatější než normální rozdělení
  - přibližně normální rozdělení má špičatost v intervalu $< -0.5; 0.5 >$
  - špičatost u výrazně sešikmených rozdělení se určuje špatně

Popisné statistiky tvaru rozdělení

Záporná, nulová a kladná šikmost



Záporná, nulová (špičatost normálního rozdělení) a kladná špičatost



Popisné statistiky číselné proměnné

Co si rozmyslet

- jaké popisné statistiky lze vyčíst z histogramu
- jaké popisné statistiky lze vyčíst z krabicového grafu
- umět nakreslit krabicový graf na základě popisných statistik
- umět popsat, jak bude vypadat histogram na základě popisných statistik
- podle sešikmení poznat, zda bude větší průměr nebo medián a naopak
- které charakteristiky variability se hodí pro jaký typ dat
- kdy má smysl hodnotit šikmost a špičatost
- jaké mají jednotlivé popisné statistiky rozměry
- jak popisné statistiky reagují na posunutí (přičtení) a změnu měřítka (přenásobení)

Číselné popisné statistiky nominální proměnné

Příklad. *Mějme náhodný výběr 10-ti dospělých lidí a předpokládejme, že jsme u nich zjišťovali barvu očí. Ve výběru jsme rozlišovali 3 barvy: modrá (M), hnědá (H) a zelená (Z). Zjistili jsme následující barvy M, M, Z, H, H, H, M, Z, M, H. Popišme zjištěné výsledky.*

Tabulka absolutních a relativních četností.

Barva	Značení	Absolutní	Relativní %
Modrá	n_1	4	40%
Hnědá	n_2	4	40%
Zelená	n_3	2	20%
Celkem	n	10	100%

Relativní četnost vypočteme jako $p_j = \frac{n_j}{n}$

Číselné popisné statistiky ordinální proměnné

Příklad. *Respondenti hodnotili určitý výrobek na pětibodové škále s kategoriemi odpovědí: velmi spokojen, spíše spokojen, ani spokojen ani nespokojen, spíše nespokojen, velmi nespokojen. Co všechno mohu k této proměnné měřit?*

Tabulka absolutních, relativních a kumulativních četností.

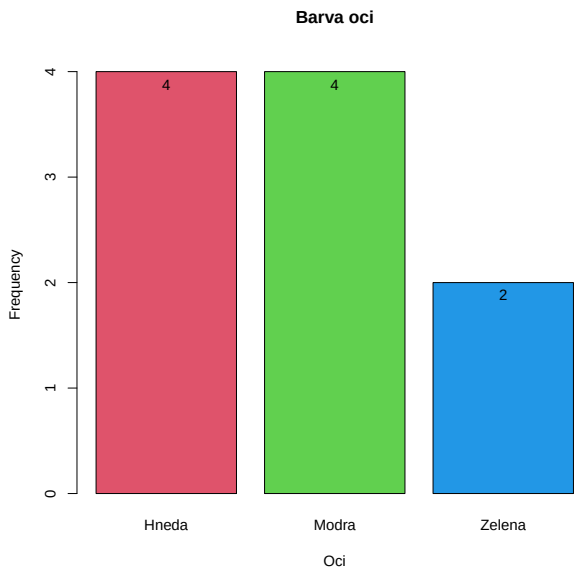
Spokojenost	Značení	Absolutní	Relativní	Kumulativní absolutní	Kumulativní relativní
velmi spokojen	n_1	10	0.2128	10	0.2128
spíše spokojen	n_2	15	0.3191	25	0.5319
ani spokojen ani nespokojen	n_3	8	0.1702	33	0.7021
spíše nespokojen	n_4	6	0.1277	39	0.8298
velmi nespokojen	n_5	8	0.1702	47	1
Celkem	n	10	1		

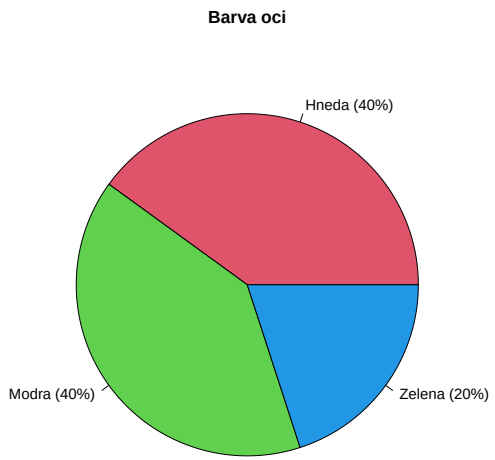
Kumulativní (postupně načítané) četnosti se počítají jen pro ordinální veličinu. Můžeme pro ni spočítat i *průměr.*

Grafické popisné statistiky nominální proměnné

Sloupcový a koláčový graf

- zobrazují se v absolutních počtech, nebo v procentech

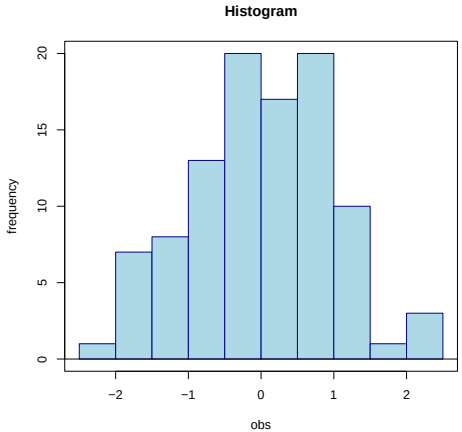


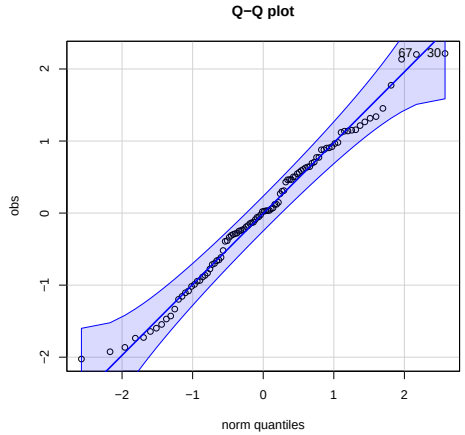


Testy normality

Většina statistických postupů je odvozena pro **normální rozdělení**. Je tedy třeba zjistit, zda ho veličina má nebo nemá.

- **Grafické testy** – histogram a pravděpodobnostní graf





- **Číselné testy** – např. Shapiro-Wilkův, Andersonův-Darlingův, Kolmogorovův-Smirnovův, Lillieforsův a další

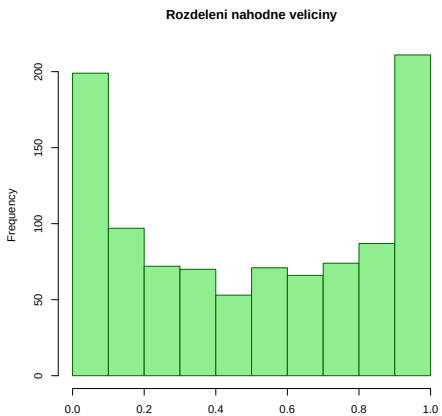
Centrální limitní věta

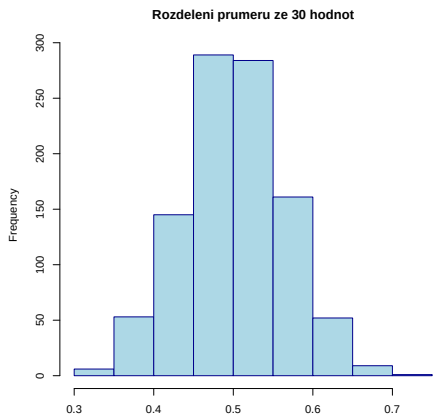
Věta

Rozdělení součtu nezávislých, stejně rozdělených náhodných veličin konverguje k normálnímu pro počet těchto náhodných veličin rostoucí nade všechny meze.

V praxi to znamená, že čím více hodnot sčítáte/průměrujete, tím spíše bude mít průměr normální rozdělení.

Rozdělení průměru 30-ti hodnot z beta rozdělení.



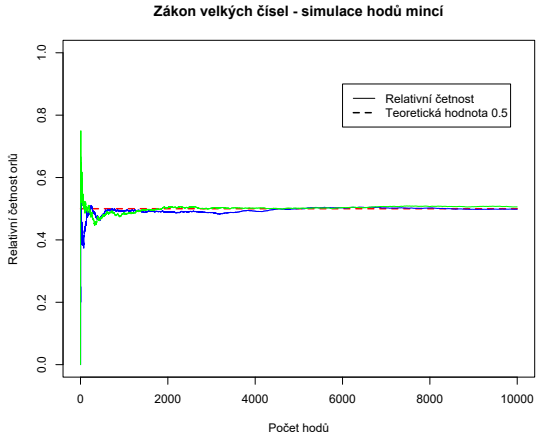


Zákon velkých čísel

Věta

Průměr nezávislých, stejně rozdělených náhodných veličin konverguje ke střední hodnotě jejich rozdělení pro počet těchto náhodných veličin rostoucí nade všechny meze.

V praxi to znamená, že výběrový průměr dobře odhaduje skutečnou střední hodnotu. Čím víc pozorování průměrujeme, tím blíže jsme odhadované hodnotě.



Odhady určujících funkcí

- **pravděpodobnostní funkci** u diskrétních veličin odhadujeme pomocí relativních četností

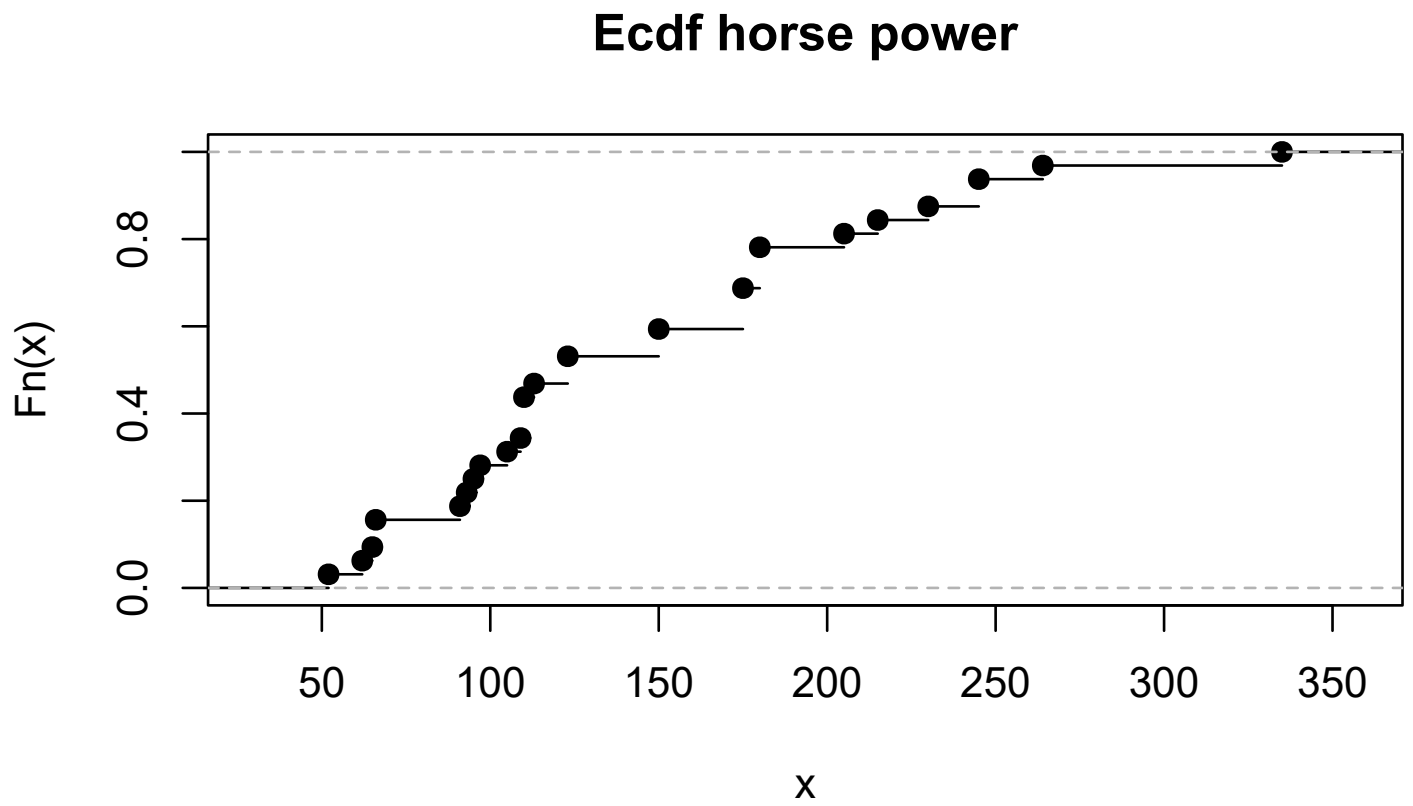
$$p_i = \frac{n_i}{n}$$

- **hustotu** u spojitých rozdělení odhadujeme z histogramu pomocí jádrového odhadu hustoty
- **kvantilovou funkci** odhadujeme pomocí výběrových kvantilů (uspořádaných hodnot)
- **distribuční funkci** odhadujeme pomocí empirické distribuční funkce

$$Ecdf(x) = \frac{\sum_{i=1}^{n} \mathrm{I}(X_i < x)}{n}$$

Empirická distribuční funkce

Příklad grafu empirické distribuční funkce



Bodový odhad střední hodnoty

Příklad. *Mějme situaci, kdy potřebujeme odhadnout průměrnou výšku dospělých lidí v celé České republice. Náhodně jsme vybrali a změřili 500 lidí. Výběrový průměr vyšel 173.12 cm a výběrová směrodatná odchylka 8.9 cm. Odhadněte populační průměr (střední hodnotu) výšky dospělých lidí.*

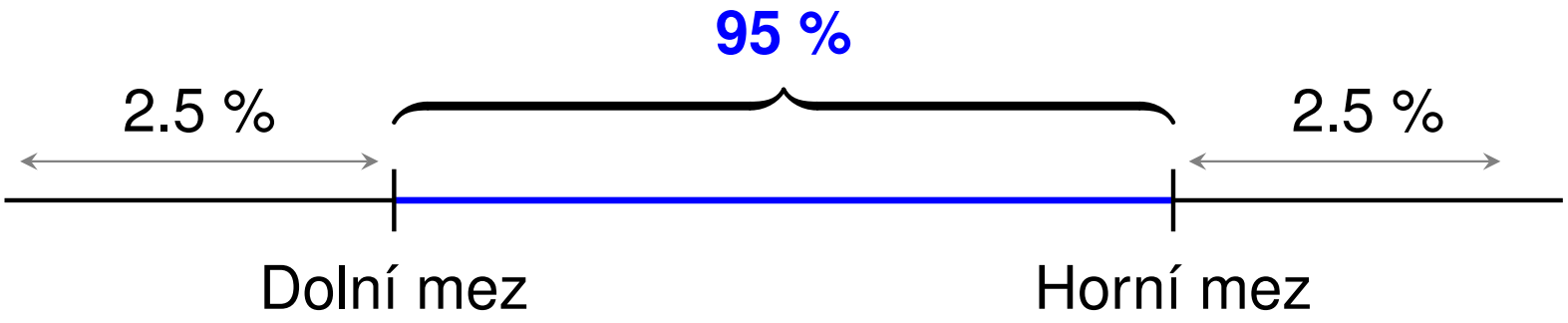
- nejlepší bodový odhad je **výběrový průměr** $\overline{X} = 173.12$
- jaká je pravděpodobnost, že se populační průměr bude rovnat přesně tomuto číslu?
- jaká je chyba tohoto odhadu
- střední chyba odhadu průměru

$$\mathrm{SEM} = \frac{\mathrm{sd(X)}}{\sqrt{\mathrm{n}}}$$

Intervalový odhad

Nyní chceme spojit bodový odhad a jeho chybu do jedné informace

- chceme interval, který by nám dal určitou jistotu, kde odhadovaný parametr hledat
- interval, který hledáme se jmenuje **interval spolehlivosti**, např. 95%-ní interval spolehlivosti
- s 95%-ní spolehlivostí **pokrývá** skutečnou střední hodnotu



- pravděpodobnost chyby označme α:
  $\alpha = 0.025 + 0.025 = 0.05$ (tedy 5%)

Intervalový odhad střední hodnoty

Interval spolehlivosti pro střední hodnotu (průměr): skutečný populační průměr/ skutečná střední hodnota v něm leží s vysokou splehlivostí.
Na čem tento interval závisí a jak?

- Výběrový průměr – leží ve středu intervalu spolehlivosti
- Výběrový rozptyl – čím větší variabilitu výběr má, tím širší bude interval spolehlivosti
- Počet pozorování – čím více pozorování, tím přesnější odhad mám a tím užší bude interval spolehlivosti
- Požadovaná spolehlivost – čím spolehlivější výsledek chci, tj. čím větší pravděpodobnost, že výběrový průměr bude ležet uvnitř intervalu spolehlivosti, tím širší interval dostanu

Intervalový odhad střední hodnoty

Podle centrální limitní věty víme, že výběrový průměr má normální rozdělení

$$\overline{X} \sim N(\mu, \sigma^2/n),$$

kde

- μ je teoretická střední hodnota
- σ je teoretická směrodatná odchylka
- n je počet pozorování

Chceme interval $< LL, UL >$ takový, že
$P(LL \leq \mu \leq UL) = 1 - \alpha$

Intervalový odhad střední hodnoty

Důkaz:

- nejprve převedeme výběrový průměr na veličinu se standardním normálním rozdělením

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1)$$

- o veličině se standardním normálním rozdělením víme

$$P(z(\alpha/2) \leq Z \leq z(1-\alpha/2)) = 1 - \alpha = 0.95$$

kde $z(.)$ značí kvantil standardního normálního rozdělení v daném bodě

- vnitřek závorky potřebuji upravit tak, aby μ bylo mezi dvěma mezema

$$z(\alpha/2) \leq \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \leq z(1-\alpha/2)$$

$$\overline{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

- jelikož $z(\alpha/2) = -z(1-\alpha/2)$ dostaneme

Intervalový odhad střední hodnoty

Interval spolehlivosti pro střední hodnotu

$$(\overline{X} - z(1-\alpha/2)\sigma/\sqrt{n}, \overline{X} + z(1-\alpha/2)\sigma/\sqrt{n})$$

- takto vypadá interval spolehlivosti, když znám skutečný rozptyl σ^2
- nejčastěji používaný kvantil $N(0,1)$ pro $\alpha = 0.05$ (spolehlivost 95%) je $z(0.975) = 1.96$

Intervalový odhad střední hodnoty

Musím-li odhadnout σ^2 pomocí výběrového rozptylu, vycházím ze vztahu

$$T = \frac{\overline{X} - \mu}{\mathrm{sd}(X)}\sqrt{n} \sim t_{n-1}$$

kde t_{n-1} značí t-rozdělení o $n-1$ stupních volnosti

Interval spolehlivosti pak je

$$\left(\overline{X} - t_{n-1}(1-\alpha/2)\mathrm{sd}(X)/\sqrt{n}, \overline{X} + t_{n-1}(1-\alpha/2)\mathrm{sd}(X)/\sqrt{n}\right)$$

kde $t_{n-1}(.)$ značí kvantil t-rozdělení

Intervalový odhad střední hodnoty

Pokračujme v příkladu s výškou lidí.
Jak vychází 95% interval spolehlivosti? Teoretický rozptyl neznám. Vím $\overline{X} = 173.12, \mathrm{sd(X)} = 8.9, \mathrm{n} = 500, \alpha = 0.05$.
Interval spolehlivosti tedy je

$$(\overline{X} - t_{n-1}(1-\alpha/2)\mathrm{sd(X)}/\sqrt{\mathrm{n}}, \overline{\mathrm{X}} + \mathrm{t_{n-1}}(1-\alpha/2)\mathrm{sd(X)}/\sqrt{\mathrm{n}})$$

$$173.12 - 1.96 \times 8.9/\sqrt{500}, 173.12 + 1.96 \times 8.9/\sqrt{500}$$

$$172.34, 173.9$$

Interval 172.34 cm až 173.9 pokrývá skutečnou střední hodnotu se splehlivostí 95%.

Intervalový odhad střední hodnoty

Co znamená slovo pokrývá?

- realizujeme-li mnoho náhodných výběrů a v každém spočítáme interval spolehlivosti,
- pak skutečná hodnota odhadovaného parametru se nachází v 95% z nich
- intervalem spolehlivosti se snažíme trefit skutečnou hodnotu parametru
- ale v praxi si interval spolehlivosti představujeme tak, že Se spolehlivostí 95% v něm leží skutečná hodnota odhadovaného parametru.

Bootstrapový interval spolehvosti pro střední hodnotu

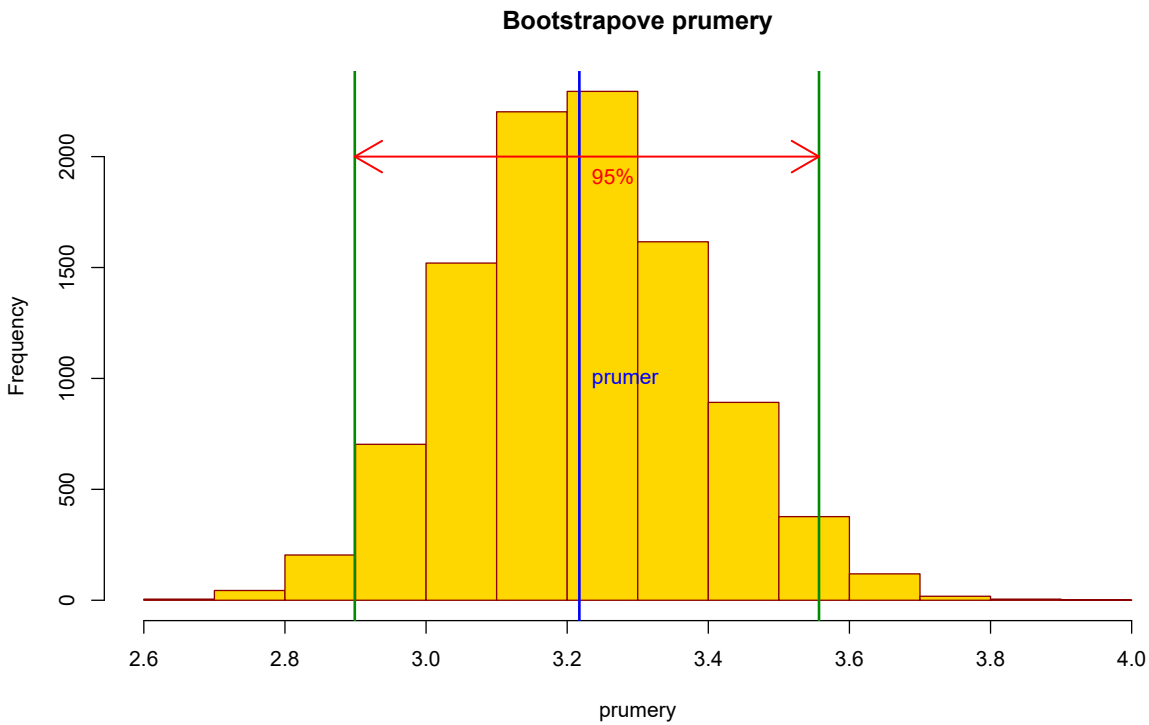
Konstrukce intervalu

- uvažujme dostupný náhodný výběr jako "základnu" dat
- realizujme B bootstrapových výběrů velikosti n na této základně - výběr **s opakováním** (každá naměřená hodnota má pst být vybrána $1/n$)
- z každého výběru spočtěme výběrový průměr
- meze bootstrapového intervalu spolehlivosti jsou $\alpha/2$ a $1-\alpha/2$-tý kvantil z vektoru průměrů.

Počet bootstrapových výběrů má být minimálně $B = 1000$, lépe $B = 10000$.

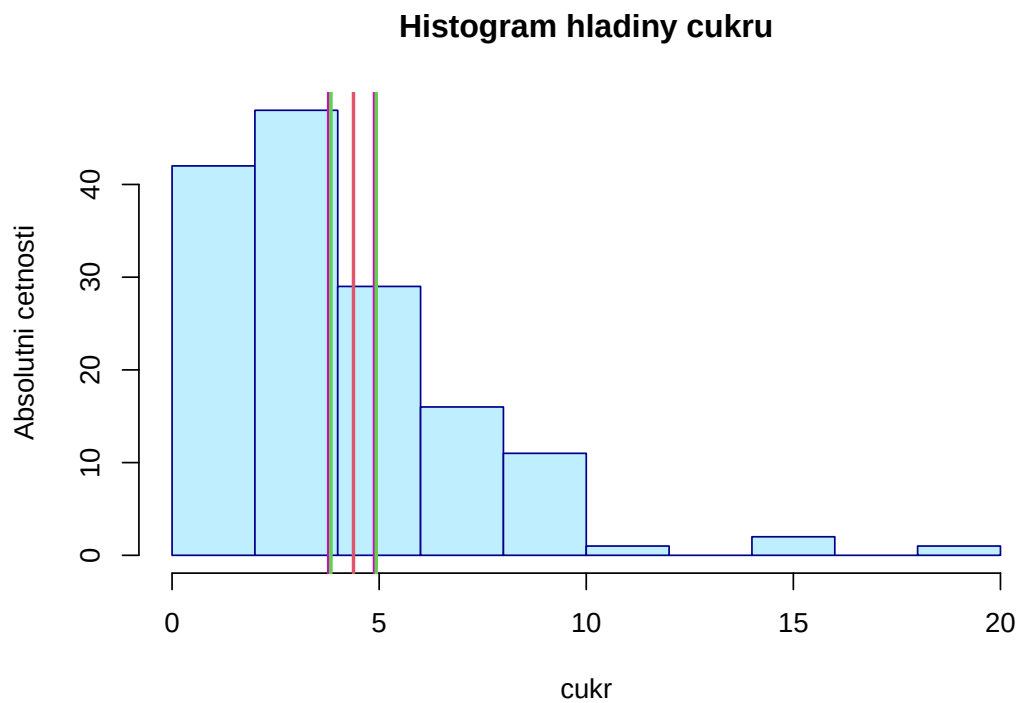
Bootstrapový interval spolehvosti pro střední hodnotu

Konstrukce intervalu



Intervalový odhad střední hodnoty

Příklad. *Uvažujme měření hladiny cukru v krvi. Bylo změřeno 150 mužů s průměrnou hodnotou cukru 4.38 a směrodatnou odchylkou 3.4. Histogram je uveden níže. Klasický interval spolehlivosti vyšel 3.83 – 4.93 (v grafu zeleně). Bootstrapový interval spolehlivosti vyšel 3.78 – 4.89 (v grafu fialově).*



Odhad pravděpodobnosti

Předpokládejme binomické rozdělení s parametrem p, který chci odhadnout z dat.

Příklad. *Ze 100 hodů šestistěnnou kostkou padla šestka 20 krát. Jak odhadnout pravděpodobnost, že padne 6?*

- nejlepším bodovým odhadem psti je **relativní četnost**
  $\hat{p} = 20/100 = 1/5$
- interval spolehlivosti vychází z faktu, že
  $$p = (\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1)$$
  pro $n\hat{p}(1-\hat{p}) > 9$
  tedy pro velká n má relativní četnost normální rozdělení
- interval spolehlivosti pro pravděpodobnost je
  $$\left(\hat{p} - z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}\right)$$
- Pro výše uvedený hod kostkou vychází 0.135 – 0.265.

Odhad pravděpodobnosti

Vlastnosti intervalu spolehlivosti pro pravděpodobnost

- asymptotický interval využívající aproximaci normálním rozdělením (**Waldův interval**)
- nefunguje pro velmi malá, respektive velmi velká p
- v případě extrémních p se dá využít aproximace Poissonovým rozdělením, nebo přesný interval využívající kvantily binomického rozdělení (**Clopper-Pearsonův interval**)
- R-ko nabízí větší množství vzorců pro výpočet intervalu spolehlivosti pro pravděpodobnost, je možné si vybrat
- pokud výpočet vrací nesmysly (záporné meze, či meze větší než 1), pak je lépe volit jiný vzorec

Odhad rozptylu

- jako bodový odhad populačního rozptylu používáme výběrový rozptyl $\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$
- nestranný odhad
- označme výběrový rozptyl jako s^2 a teoretický rozptyl jako σ^2, pak náhodná veličina $\chi = (n-1)s^2/\sigma^2$ má χ^2 rozdělení o n stupních volnosti
- χ^2 rozdělení není symetrické
- intervalový odhad pro rozptyl je

$$\left(\frac{(n-1)s^2}{\chi^2_n(1-\alpha/2)}, \frac{(n-1)s^2}{\chi^2_n(\alpha/2)} \right)$$

Základy testování hypotéz

Používáme, když potřebujeme **ověřit nějaké tvrzení**, např.

- Nový lék je lepší než ten stávající.
- Průměrná výška lidí se za posledních 50 let zvýšila.
- Výnosy z jednotlivých druhů jabloní se liší.
- Krevní tlak závisí na hmotnosti.

Testované hypotézy

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- **Nulovou hypotézu** – značíme H_0
  – je v ní vždy pouze jedna varianta
  – př. nový lék je stejný jako ten stávající, výnosy druhů jabloní jsou stejné, proměnné spolu nesouvisí.
- **Alternativní hypotézu** – značíme H_1
  – obsahuje více možností (např. interval)
  – je v ní to, co chceme prokázat
  – př. nový lék je lepší než ten stávající, výnosy druhů jabloní se liší, proměnné spolu souvisí.

Výsledek testu

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu
  – tím jsme prokázali platnost alternativy
- Nezamítneme nulovou hypotézu
  – tím jsme neprokázali nic

Při rozhodování můžeme udělat chybu

- chyba prvního druhu – zamítneme H_0, přestože platí
  – značí se α, a jmenuje se **hladina významnosti**
  – závažnější z obou chyb
- chyba druhého druhu – nezamítneme H_0, přestože neplatí
  – značí se β a hodnota $1 - \beta$ se nazývá **síla testu**
  – za dané hladiny významnosti chceme test co nejsilnější

Výsledek testu

	Skutečně platí H_0	Skutečně platí H_1
Zamítáme H_0	Chyba I. druhu $\leq \alpha$	OK síla testu
Nezamítáme H_0	OK	Chyba II. druhu β

Výsledek testu

Presumpce neviny – představte si statistický test jako soudní proces. **Testujeme**

- Nulová hypotéza (H_0): Obžalovaný je nevinný
- Alternativní hypotéza (H_1): Obžalovaný je vinný

Verdikt soudu

- Odsouzení = zamítnutí nulové hypotézy (máme dost důkazů)
- Osvobození = nezamítnutí nulové hypotézy (nedostatek důkazů o vině)

Možné chyby

- Chyba I. druhu: odsoudíme nevinného
- Chyba II. druhu: neodsoudíme viníka
- Síla testu"schopnost odhalit viníka

Výsledek testu

Podle toho, co testujeme, a podle typu dat vybereme vhodný statistický test, kterým budeme o platnosti testovaných hypotéz rozhodovat. Rozhodnutí můžeme udělat buď na základě

- porovnání **testové statistiky** (T) a kritické hodnoty (c, jsou tabelovány)
- porovnání ***p*-hodnoty** a hladiny významnosti (α)

Platí, že

- absolutní hodnota testové statistiky $|T| \geq c$ nebo p-hodnota $\leq \alpha$ potom ZAMÍTÁME H_0
- absolutní hodnota testové statistiky $|T| < c$ nebo p-hodnota $> \alpha$ potom NEZAMÍTÁME H_0

P-hodnota

S testovou statistikou se většinou pracuje při ručním výpočtu.

Statistické softwary vrací jako výsledek testu ***p*-hodnotu**.

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti H_0
  nastal výsledek, jaký nastal,
  nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- definice *p*-hodnoty se týká testové statistiky

(Ne)zamítnout H_0 nestačí, tento výsledek je třeba interpretovat vzhledem k položené otázce.

Jednovýběrový t-test

Nejjednodušším testem je **jednovýběrový test o střední hodnotě**.
Testujeme

- H_0 : střední hodnota $= \mu_0$

Proti jedné ze tří alternativ

- H_1 : střední hodnota $\neq \mu_0$
- H_1 : střední hodnota $< \mu_0$
- H_1 : střední hodnota $> \mu_0$

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$

Jednovýběrový t-test

Testová statistika jednovýběrového t-testu je

$$T = \frac{\overline{X} - \mu_0}{\mathrm{sd}(X)}\sqrt{n}$$

- za platnosti nulové hypotézy má t-rozdělení o $n - 1$ stupních volnosti
- testovou statistiku T porovnáváme s kritickými hodnotami – kvantily t-rozdělení
- nebo na základě ní vypočteme p-hodnotu

Jednovýběrový t-test

Předpokladem jednovýběrového t-testu je, že průměr testované veličiny má **normální rozdělení** (díky CLV většinou splněno).

- máme na platnost CLV dostatek poyorování?
- největším problémem jsou odlehlé hodnoty
- v praxi se normalita ověřuje přímo u **testované proměnné**, nikoliv u jejího průměru
- výsledky jednovýběrového t-testu lze zjistit i z **intervalu spolehlivosti**
  - je-li testovaná hodnota uvnitř intervalu spolehlivosti, pak nulovou hypotézu nezamítáme
  - je-li testovaná hodnota mimo interval spolehlivosti, pak nulovou hypotézu zamítáme
- proč toto spojení funguje?

Jednovýběrový t-test

Příklad. *Bylo změřeno 222 jedenáctiletých dětí. Průměrná výška tohoto výběru je 148.8 cm, a směrodatná odchylka výšky vyšla 7.1. Dá se předpokládat, že průměrná výška všech jedenátiletých dětí v republice je menší než 150 cm?*
Testované hypotézy

- H_0 : průměrná výška $=$ 150 cm
- H_1 : průměrná výška $<$ 150 cm

Testujeme na hladině významnosti $\alpha = 0.05$.

Jednovýběrový t-test

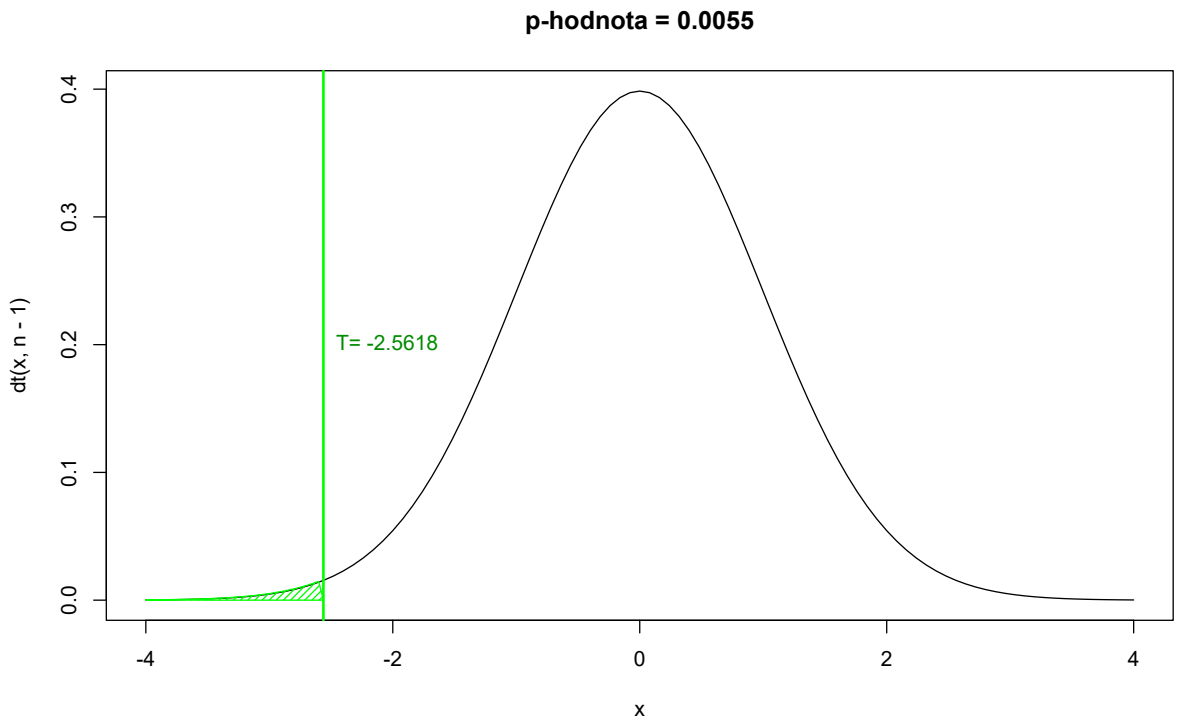
Pokračování příkladu.
Testová statistika vyšla

$$T = \frac{\overline{X} - \mu_0}{\text{sd(X)}}\sqrt{n} = \frac{148.8 - 150}{7.1/\sqrt{222}} = -2.5618$$

Tuto hodnotu porovnám s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$. Jelikož testová statistika je v absolutní hodnotě větší než kritická hodnota, zamítám nulovou hypotézu. P-hodnota vyšla $p = 0.005 < 0.05$, což také vede na zamítnutí nulové hypotézy.
Závěr: Prokázala jsem, že průměrná výška jedenáctiletých dětí je menší než 150 cm.

Jednovýběrový t-test

P-hodnota



Jak by vypadala p-hodnota pro oboustrannou alternativu?

Znaménkový test

Neparametrický test o střední hodnotě

- pro data, která nemají normální rozdělení
- není založen na průměru, ale na znaménkách odchylek od mediánu
- nepotřebuje žádné předpoklady

Testované hypotézy

- H_0 : medián $= m_0$
- H_1 : medián $\neq m_0$ nebo $> m_0$ nebo $< m_0$

Postup

- označme Z počet kladných odchylek od mediánu $X_i - m_0$
- za platnosti H_0 má Z binomické rozdělení $Bi(n, 1/2)$
- pro velká n je možné použít i transformaci

$$U = \frac{2Z - n}{\sqrt{n}}$$

za platnosti H_0 má U normální rozdělení $N(0, 1)$

Znaménkový test

Příklad. *Uvažujme naměřené věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

- H_0 : *medián věku otců je 33 let*
- H_1 : *medián věku otců není 33 let*

Spočtěme rozdíly $X_i - m_0$: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17. Kladných hodnot je mezi nimi $Z = 4$. P-hodnota testu vychází 0.75, což je hodnota větší než $\alpha (= 0.05)$ a H_0 tedy nezamítáme.
Použitím U-transformace dostaneme $U = -0.632$ a p-hodnotu 0.527.
Závěr: *Střední hodnota věku otců může být 33.*

Wilcoxonův jednovýběrový test

Dalším neparametrickým testem je Wilcoxonův test, neboli **Mann-Whitneyův** test.

- používá se, když proměnná nemá normální rozdělení
- předpokládá ale spojité symetrické rozdělení, v takovém případě testuje hodnotu mediánu,
- pokud předpoklad symetrie není splěn, testuje shodu rozdělení
- dobře funguje, když máte v datech odlehlé hodnoty
- je založen na pořadích
- silnější než znaménkový test
- testované hypotézy zůstávají stejné (testuje hodnotu mediánu)

Wilcoxonův jednovýběrový test

Postup testu

- spočítají se rozdíly od testované hodnoty $X_i - m_0$
- určí se jejich znaménko
- určí se pořadí absolutních hodnot rozdílů
- spočítá se součet těchto pořadí patřících kladným rozdílům
- označme S^+ součet pořadí kladných rozdílů a S^- součet pořadí záporných rozdílů, musí platit $S^+ + S^- = n(n+1)/2$.

Pro větší n lze užít transformaci

$$U = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}}$$

která má za platnosti H_0 normální $N(0,1)$ rozdělení.

Wilcoxonův jednovýběrový test

Příklad. *Pokračujme v příkladu s věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a opět testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

- H_0 : *medián věku otců je 33 let*
- H_1 : *medián věku otců není 33 let*

Spočtěme rozdíly $X_i - m_0$: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17 a jejich absolutním hodnotám přiřaďme pořadí 1.5, 6, 1.5, 6, 6, 8, 3.5, 3.5, 9, 10. Sečtěme kladné (modré) pořadí $S^+ = 21$ a záporné (červené) pořadí $S^- = 34$. Testová statistika vychází $U = -0.66$ a p-hodnota 0.51 je větší než $\alpha (= 0.05)$, H_0 tedy nezamítáme.

Závěr: *Střední hodnota věku otců může být 33.*

Párový t-test

Párový test se používá v případě, že porovnáváme střední hodnotu ve dvou **závislých** výběrech.
Např.

- *Jsou otcové v průměru o 10 cm vyšší než matky?*
- *Mají praváci silnější pravou ruku než levou?*
- *Klesl pacientům po podání léku krevní tlak?*

Ať je otázka formulována jakkoliv, tak test porovnává průměrné hodnoty. Vyjde nám tedy odpověď, jak je to "v průměru".

Závislé výběry poznám tak, že data tvoří přirozené páry.

Párový t-test

Při aplikaci testu je důležité udržet párová data u sebe, (abyste neporovnávali Vaší pravou ruku se sousedovou levou).
V prvním kroku jsou pro všechny páry vypočteny **rozdíly**:

$$R_i = X_i - Y_i$$

dále je testována střední hodnota těchto rozdílů, tedy je aplikován jednovýběrový t-test na hodnoty rozdílu.

Párový t-test

Příklad. *Bylo měřeno 222 dětí v jedenáctém a dvanáctém roce věku. Průměrná výška jedenáctiletých vyšla 148.8 cm, u dvanáctiletých pak 154.9 cm. Směrodatná odchylka u jedenáctiletých vyšla 7.1 cm, u dvanáctiletých pak 7.9 cm. Průměrná hodnota rozdílu výšek vyšla 6.1 cm a směrodatná odchylka 2.8 cm. Vyrostly děti mezi jedenáctým a dvanáctým rokem v průměru alespoň o 5 cm?*

Do testové statistiky vkládáme charakteristiky rozdílu (tedy nikoliv rozdíl průměrů, ale průměr rozdílů).

$$T = \frac{\overline{X} - \mu_0}{\text{sd(X)}}\sqrt{n} = \frac{6.1 - 5}{2.8}\sqrt{222} = 5.9$$

Tuto testovou statistiku porovnáváme s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$. Jelikož testová statistika je větší než příslušný kvantil, zamítám nulovou hypotézu. P-hodnota pro tento případ vychází $p = 7.26 \times 10^{-9}$, což je menší než $\alpha = 0.05$.

Závěr: Prokázali jsme, že mezi jedenáctým a dvanáctým rokem děti vyrostly v průměru o více než o 5 cm.

Wilcoxonův párový test

Dva závislé výběry, které nesplňují předpoklad normality, porovnáváme pomocí párového Wilcoxonova testu.

- testované hypotézy zůstávají stejné jako u párového t-testu
- spočítají se rozdíly v rámci párů

$$R_i = X_i - Y_i$$

- otestuje se normalita rozdílů
- pokud rozdíly nemají normální rozdělení, použije se jednovýběrový Wilcoxonův test

Dvouvýběrový test

Porovnává střední hodnotu dvou **nezávislých** výběrů
Testované hypozézy

- H_0 : střední hodnota X – střední hodnota $Y = 0$
- H_1 : střední hodnota X – střední hodnota $Y \neq 0, < 0, > 0$

Kontrolují se zde 2 předpoklady

- normalitu dat
- shodu rozptylů

A vybíráme jeden ze tří testů

- Dvouvýběrový t-test pro normální data a shodné rozptyly
- Welchův dvouvýběrový test pro normální data a různé rozptyly
- Wilcoxonův dvouvýběrový test pro data, která nemají normálí rozdělení

Test shody dvou rozptylů

Test shody rozptylů se vyhodnocuje i u nenormálních dat.
Testované hypotézy

- H_0 : rozptyly se ve výběrech neliší
- H_1 : rozptyly se ve výběrech liší.

Testová statistika testu je

$$F = \frac{\mathrm{Var(X)}}{\mathrm{Var(Y)}} \sim F_{n_1-1,n_2-1}$$

a za platnosti H_0 má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti.

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika tohoto testu má tvar

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

kde

$$S = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \right)$$

a n_1, n_2 je rozsah výběru X, respektive Y. Za platnosti nulové hypotézy má tato statistika t-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.

Welchův test

Testová statistika tohoto testu má tvar

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{\mathrm{Var(X)}}{n_1} + \frac{\mathrm{Var(Y)}}{n_2}}}$$

a za platnosti nulové hypotézy má t-rozdělení o ν stupních volnosti, kde

$$\nu = \frac{(\mathrm{Var(X)}/n_1 + \mathrm{Var(Y)}/n_2)^2}{\frac{(\mathrm{Var(X)}/n_1)^2}{n_1 - 1} + \frac{(\mathrm{Var(Y)}/n_2)^2}{n_2 - 1}}.$$

kritické hodnoty je možno odvodit, přestože ν není celé číslo.

Dvouvýběrový t-test

Příklad. *Ve výběru mám 222 jedenáctiletých dětí, z toho 159 hochů a 63 dívek. Průměrná hmotnost hochů vyšla 38.1 kg a u dívek 39.1. Směrodatná odchylka pro hochy vyšla 6.7 kg a pro dívky 7.1.*
Je hmotnost jedenáctiletých dětí v průměru stejná pro hochy jako pro dívky?
Test shody rozptylů

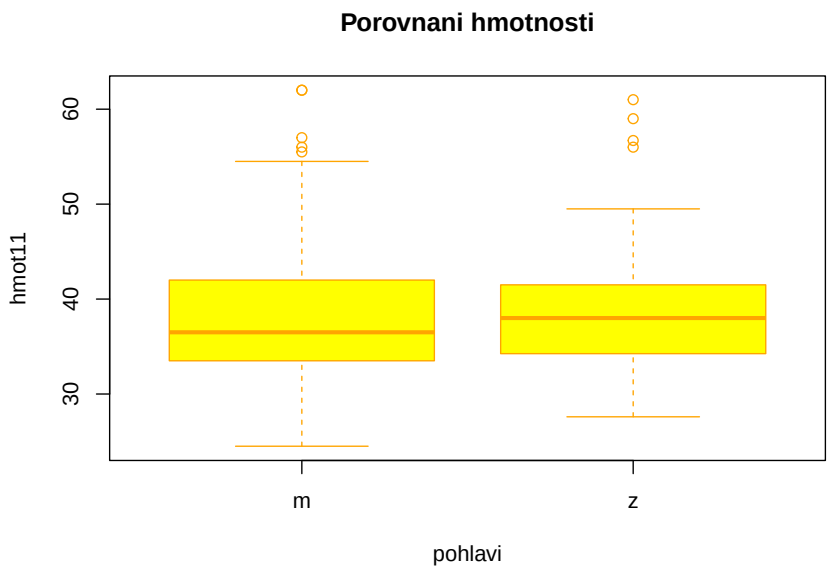
- testová statistika $F = \frac{\text{Var(X)}}{\text{Var(Y)}} = \frac{45.1}{50.6} = 0.89$
- p-hodnota $= 0.56 > \alpha = 0.05$
- nulovou hypotézu nezamítáme
- rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly

Dvouvýběrový t-test

Testujeme

- H_0 : hmotnost hochů a hmotnost dívek se neliší
  hmotnost hochů – hmotnost dívek = 0
- H_1 : hmotnost hochů a dívek se liší
  hmotnost hochů – hmotnost dívek $\neq 0$

Grafické porovnání



Dvouvýběrový t-test

- testová statistika

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{38.1 - 39.1}{6.83}\sqrt{\frac{159 \times 63}{159 + 63}} = -1.001$$

- porovnáváme s kvantilem t-rozdělení $t_{220}(1 - 0.025) = 1.97$ (kvantil pro oboustrannou alternativu)
- testová statistika je v absolutní hodnotě menší než tento kvantil, tak nulovou hypotézu nezamítám.
- p-hodnota $= 0.3151 > \alpha = 0.05$
- **Závěr:** Na hladině významnosti 5% jsem neprokázala, že by se hmotnost jedenáctiletých hochů a dívek lišila.

Wilcoxonův dvouvýběrový test

Používá se pro porovnání dvou nezávislých výběrů, které nesplňují předpoklad normality.
Test je založen na pořadích hodnot sdruženého výběru.

Postup

- oba výběry se spojí do jednoho sdruženého
- sdružený výběr se uspořádá podle velikosti a každé pozorování dostane své pořadí
- pro oba výběry se vypočte součet pořadí a následně i průměrné pořadí
- pokud jsou si průměrná pořadí podobná, výběry se mezi sebou významně neliší

Wilcoxonův dvouvýběrový test

Technický výpočet: označme T_1, T_2 součet pořadí v prvním, respektive druhém výběru. Dále vypočteme

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1, U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2,$$

kde n_1, n_2 jsou rozsahy jednotlivých výběrů. Přesný test porovnává hodnotu $min(U_1, U_2)$ s kritickou hodnotou. Asymptoticky platí, že

$$U_0 = \frac{U_1 - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}$$

má za platnosti H_0 $N(0,1)$ rozdělení.

Wilcoxonův dvouvýběrový test

Příklad. *Chceme porovnat výsledky testů studentů v Ústí nad Labem a v Liberci. Studenti v Ústí dostali bodová ohodnocení 45, 79, 81, 56, 53, 77. Studenti v Liberci získali ohodnocení 76, 62, 84, 80, 41, 79, 66. Testujeme*

- H_0 : *Studenti v Ústí a v Liberci jsou stejní*
- H_1 : *Studenti v Ústí a v Liberci se liší.*
- *V prvním kroku srovnám všechny hodnoty do řady* 41, 45, 53, 56, 62, 66, 76, 77, 79, 79, 80, 81, 84
- *následně jim přiřadím pořadí* 1, 2, 3, 4, 5, 6, 7, 8, 9.5, 9.5, 11, 12, 13
- *pak vypočtu* $T_1 = 38.5$, $T_2 = 52.5$, $U_1 = 24.5$, $U_2 = 17.5$, $U_0 = 0.5$, $p = 0.6678$

P-hodnota $> \alpha$ *a tedy nezamítám nulovou hypotézu, neprokázal se rozdíl mezi studenty v Ústí a v Liberci.*

Analýza rozptylu – ANOVA

Střední hodnotu ve více než ve dvou nezávislých výběrech porovnáváme pomocí **analýzy rozptylu**.
Testované hypotézy

- H_0 : všechny střední hodnoty jsou stejné
- H_1 : alespoň jedna střední hodnota se liší

Opět máme 2 předpoklady

- normalitu a shodu rozptylů

a tři testy

- Klasická ANOVA pro normální data a shodné rozptyly
- Welchova ANOVA pro normální data a různé rozptyly
- Kruskal-Wallisova ANOVA pro data, která nemají normální rozdělení

Klasická analýza rozptylu – ANOVA

Klasická ANOVA pro shodné rozptyly porovnává variabilitu mezi výběry s variabilitou v rámci výběrů.
Označme

- X_{ij} i-té pozorování z j-tého výběru
- $\overline{X}_{i.}$ průměr i-tého výběru
- $\overline{X}_{..}$ celkový průměr všech pozorování
- n_i rozsah i-tého výběru a k počet výběrů

Analýza rozptylu rozkládá celkovou variabilitu

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{..})^2$$

Klasická analýza rozptylu – ANOVA

Rozložení celkové variability SST na variabilitu vysvětlenou výběry (mezi výběry) SS_A a variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SS_e.

$$\begin{aligned} SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{..})^2 = \\ &= \sum_{i=1}^{k} n_i(\overline{X}_{i.} - \overline{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i.})^2 = \\ &= SSA + SSe \end{aligned}$$

Analýza rozptylu – ANOVA

Výstupem z analýzy rozptylu je tzv. **tabulka analýzy rozptylu**

	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	p-hodnota
Faktor A	SSA	$dfA = k - 1$	$MSA = \frac{SSA}{dfA}$	$F = MSA/MSe$	p
Chyba e	SSe	$dfe = n - k$	$MSe = \frac{SSe}{dfe}$		
Celkem	SST	$dft = n - 1$			

Za platnosti nulové hypotézy má testová statistika F-rozdělení o $k - 1$ a $n - k$ stupních volnosti.

Bartlettův test

Test shody rozptylů ve více výběrech

Testované hypotézy

- H_0 : rozptyly jsou shodné
- H_1 : rozptyly se liší

Testová statistika je založena na výběrových rozptylech v každém výběru zvlášť. Označme $\mathrm{Var(X)_i}$ výběrový rozptyl v i-tém výběru a

$$S^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\mathrm{Var(X)_i}}{n-k},$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right)$$

Testová statistika

$$B = \frac{1}{C}\left((n-k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln \mathrm{Var(X)_i}\right)$$

má za platnosti nulové hypotézy χ^2-rozdělení o $k-1$ stupních volnosti.

Párové srovnání

Vzájemné srovnání všech dvojic výběrů se dělá pomocí **Tukeyho testu**, případně **Tukeyho HSD test** pro různě velké výběry.
Testované hypotézy pro všechny dvojice i a j

- H_0 : střední hodnoty μ_i a μ_j jsou stejné
- H_1 : střední hodnoty μ_i a μ_j se liší

Testové statistiky mají tvar

$$Q = \frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{s^*}, \text{ kde } s^* = \sqrt{\frac{SSe}{n(n-k)}}$$

Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty.

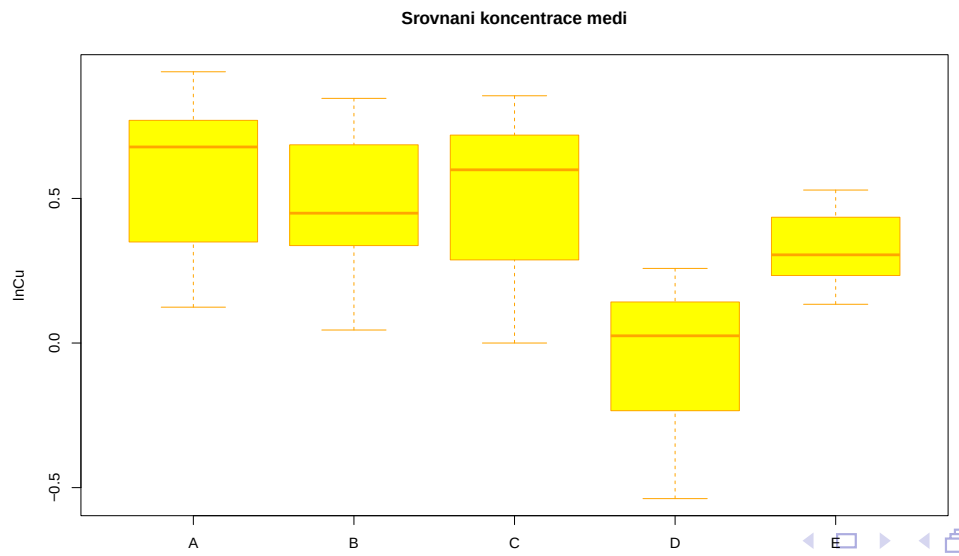
Analýza rozptylu – ANOVA

Příklad. *Byla měřena koncentrace mědi v těle ryb. Porovnáváno bylo 5 rybníků, kde z každého byl vyloven vzorek 7-mi ryb. Výběrové průměry pro jednotlivé rybníky vyšly 0.57, 0.48, 0.50, -0.06 a 0.33. Liší se od sebe tyto rybníky?*

Testované hypotézy

- H_0 : všechny rybníky jsou stejné
- H_1 : alespoň jeden rybník se liší

Grafické porovnání



Analýza rozptylu – ANOVA

Abychom mohli vybrat správnou verzi analýzy rozptylu, otestujme nejprve shodu rozptylů ve všech výběrech. Tyto rozptyly vyšly postupně 0.10, 0.08, 0.10, 0.08 a 0.02. Testované hypotézy

- H_0 : rozptyly jsou shodné
- H_1 : rozptyly se liší

Testová statistika Bartlettova testu vyšla 3.67 při čtyřech stupních volnosti, což dává p-hodnotu 0.45. Jelikož je p-hodnota větší než $\alpha = 0.05$, nulovou hypotézu nezamítáme a můžeme použít klasickou ANOVU pro shodné rozptyly.

Analýza rozptylu – ANOVA

Tabulka analýzy rozptylu vyšla

	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	*p*-hodnota
Rybník	1.796	4	0.4491	5.896	0.00127
Chyba	2.285	30	0.0762		
Celkem	4.081	34			

P-hodnota vyšla menší než $\alpha = 0.05$, což znamená, že nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší.

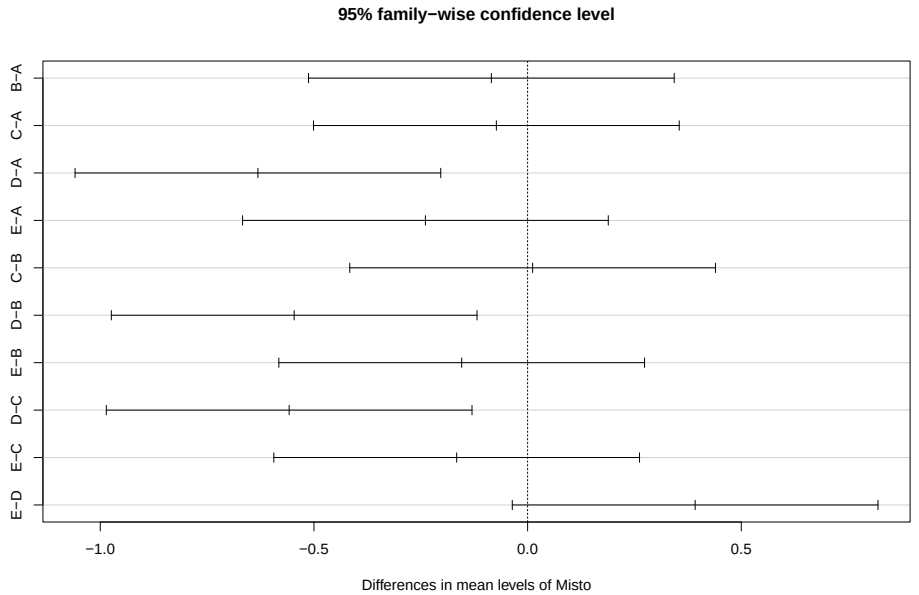
Analýza rozptylu – ANOVA

Párové srovnání vrátí následující tabulku

	rozdíl	dolní mez	horní mez	p-hodnota
B-A	-0.08485714	-0.51274077	0.3430265	0.9777112
C-A	-0.07314286	-0.50102648	0.3547408	0.9871500
D-A	-0.63114286	-1.05902648	-0.2032592	**0.0015454**
E-A	-0.23914286	-0.66702648	0.1887408	0.4960690
C-B	0.01171429	-0.41616934	0.4395979	0.9999904
D-B	-0.54628571	-0.97416934	-0.1184021	**0.0070956**
E-B	-0.15428571	-0.58216934	0.2735979	0.8319549
D-C	-0.55800000	-0.98588362	-0.1301164	**0.0057762**
E-C	-0.16600000	-0.59388362	0.2618836	0.7920009
E-D	0.39200000	-0.03588362	0.8198836	0.0850175

Analýza rozptylu – ANOVA

Graf pro párové srovnání. Pro kterou dvojici rybníků interval spolehlivosti neobsahuje svislou čárkovanou čáru (nulu), pak mezi ní je významný rozdíl.



Závěr: Rybníky se v koncentraci mědi v těle ryb významně liší, konkrétně se liší rybník D od rybníků A, B a C.

Kruskal-Wallisův test

Přímé zobecnění Wilcoxonova dvouvýběrového testu pro více než 2 výběry.

Postup (stejný jako u dvouvýběrového Wilcoxonova testu)

- srovnáme všechny naměřené hodnoty do řady
- určíme jejich pořadí
- sečteme pořadí pro jednotlivé výběry: $T_1, \ldots, n_k$, kde k je počet výběrů

Testová statistika

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i}{n_i} - 3(n+1)$$

má za platnosti H_0 χ^2-rozdělení.

Dunnův test

Párové srovnání pro data, která nemají normální rozdělení
Testová statistika porovnávající i-tý a j-tý výběr je

$$D = \frac{(|\frac{T_i}{n_i} - \frac{T_j}{n_j}|)}{\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

V případě, že v datech jsou shodné hodnoty a je tedy třeba dělit pořadí, používá se statistika

$$D = \frac{(|\frac{T_i}{n_i} - \frac{T_j}{n_j}|)}{\sqrt{\frac{n(n+1) - \sum_{l=1}^{r}(S_l^3 - S_l)/(n-1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

kde S_l je počet l-té shodné hodnoty.

Tato statistika má za platnosti H_0 $N(0,1)$-rozdělení. Pro vícenásobné porovnání se pak použijí upravené p-hodnoty, aby byla udržena celková hladina testu.

ANOVA pro opakovaná měření

Pokud se chystáme porovnávat několik závislých výběrů, používá se **ANOVA pro opakovaná měření**.
Příklady takovéto situace mohou být

- **Ochutnávka jogurtů:** 20 lidí ochutnává a hodnotí každý všech 5 porovnávaných vzorků jogurtu.
- **Měření opakovaná v čase:** chceme hodnotit vývoj pacientova zdravotního stavu v čase. Pro 30 pacientů děláme opakovaná měření játrových testů.

Stále se testují hypotézy

- H_0 : Střední hodnoty výběrů se neliší
- H_1 : Střední hodnoty výběrů se liší

ANOVA pro opakovaná měření

- porovnává se variabilita mezi výběry s variabilitou zbytkovou
- zbytkovou variabilitu získáme tak, že od variability v rámci výběrů odečteme variabilitu mezi jedinci

$$
\begin{aligned}
SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2 = \\
&= \sum_{i=1}^{k} n_i(\overline{X}_{i.}-\overline{X}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2 = \\
&= SSA + SSe \\
SSz &= SSe - SSS = SSe - k\sum_{j=1}^{n_i}(\overline{X}_{.j}-\overline{X}_{..})^2
\end{aligned}
$$

Test je pak založen na porovnání SSA a SSz.

Friedmanův test

Neparamterický test porovnávající závislé výběry.

Postup

- stanoví se pořadí hodnot v rámci každého jedince
- pro každý výběr se spočte součet a průměr pořadí
- označme tyto průměry $\bar{r}_{.i}$

Testová statistika

$$Q = \frac{12n}{k(k+1)} \sum_{i=1}^{k} \left(\bar{r}_{.i} - \frac{k+1}{2} \right)^2$$

za platnosti nulové hypotézy má χ^2-rozdělení o $k-1$ stupních volnosti.

Test dobré shody

Test o pravděpodobnostním rozdělení kategorické proměnné.

- proměnná s **Multinomickým rozdělením** (zobecnění Binomického rozdělení)
- proměnná může nabývat k hodnot/kategorií
- uděláme n pokusů
- počítáme, kolikrát nastala která kategorie
- měříme proměnné $X_1, \ldots, X_k$
- označme p_i pst, že nastane i-tá kategorie

Multinomické rozdělení je dáno pravděpodobnostmi

$$P(X_1 = c_1, \ldots, X_k = c_k) = \frac{n!}{c_1! \cdot \ldots \cdot c_k!} p_1^{c_1} \cdot \ldots \cdot p_k^{c_k}$$

Dále platí, že

$$\mathrm{E}(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1 - p_i)$$

Test dobré shody

Testované hypotézy

- $H_0 : p_1 = \pi_1, \ldots, p_k = \pi_k$
- H_1 : neplatí $p_1 = \pi_1, \ldots, p_k = \pi_k$

Testová statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(c_i - n\pi_i)^2}{np_i}$$

za platnosti H_0 má χ^2-rozdělení o $k - 1$ stupních volnosti.

- předpokladem je, že všechny očekávané četnosti $n\pi_i$, jsou větší než 5
- tímto testem můžeme testovat, zda náhodná veličina má nějaké konkrétní rozdělení

Test dobré shody

Příklad. *Házíme 50 krát šestistěnnou kostkou a počítáme, kolikrát padla která hodnota. Jednička padla 8 krát, dvojka 5 krát, trojka 12 krát, čtyřka 7 krát, pětka 9 krát a šestka také 9 krát. Můžeme o kostce říci, že je spravedlivá?*

Testujeme hypotézy

- $H_0 : p_1 = p_2 = \ldots = p_6 = 1/6$
- H_1 : *alespoň jedna z pravděpodobností* $p_1, \ldots, p_6$ *se nerovná* $1/6$.

Test dobré shody

Příklad. *Naměřili jsme hodnoty* $c_1 = 8, c_2 = 5, c_3 = 12, c_4 = 7, c_5 = 9, c_6 = 9$. *Teoretická hodnota* $n\pi_i = 50 \times 1/6 = 8.3333$. *Dosadíme do vzorce a dostaneme*

$$\begin{aligned}\chi^2 &= \frac{(8-8.3333)^2}{8.3333} + \frac{(5-8.3333)^2}{8.3333} + \frac{(12-8.3333)^2}{8.3333} + \\ &+ \frac{(7-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} + \frac{(9-8.3333)^2}{8.3333} = 3.28\end{aligned}$$

Kritická hodnota χ^2*- rozdělení o* 5*-ti stupních volnosti je* $\chi^2_5 = 11.07$ *a p-hodnota vyšla* $p = 0.6569$. *Testová statistika je větší než kritická hodnota a p-hodnota menší než* α*, tedy nezamítáme nulovou hypotézu.*

Závěr: *Neprokázali jsme, že by kostka byla falešná.*

χ^2-test nezávislosti

Vztah dvou kategorických proměnných popisujeme **tabulkou absolutních četností**. Označme

- $X_1, \ldots, X_k$ hodnoty jedné kategorické proměnné
- $Y_1, \ldots, Y_l$ hodnoty druhé kategorické proměnné
- $n_{i,j}$ četnost současného výskytu znaků X_i, Y_j
- $n_{i.}$ marginální četnost znaku X_i
- $n_{.j}$ marginální četnost znaku Y_j
- n celkový počet pozorování

χ^2-test nezávislosti

Kontingenční tabulka absolutních četností má tvar

	Y_1	$\dots$	Y_l	
X_1	$n_{1,1}$	$\dots$	$n_{1,l}$	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
X_k	$n_{k,1}$	$\dots$	$n_{k,l}$	$n_{k.}$
	$n_{.1}$	$\dots$	$n_{.l}$	n

χ^2-test nezávislosti

Testované hypotézy

- H_0 : proměnné na sobě nezávisí
- H_1 : proměnné na sobě závisí

Test je založen na porovnání

- pozorovaných četností n_{ij}
- očekávaných četností $n_{i.}n_{.j}/n$
  vychází z definice nezávislosti $P(A \cap B) = P(A)P(B)$

Testová statistika

$$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(pozorovane_{i,j} - ocekavane_{i,j})^2}{ocekavane_{i,j}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{i,j} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

za platnosti H_0 má χ^2-rozdělení o $(k-1)(l-1)$ stupních volnosti.

Fisherův exaktní test

Test nezávislosti pro malá data

- když není splněn předpoklad χ^2-testu, tj. některá očekávaná četnost je menší než 5
- počítá přímo p-hodnotu ke konkrétní tabulce
- známý též jako **Fisherův faktoriálový test**

Pro čtyřpolní tabulku

	Y_1	Y_2	
X_1	n_{11}	n_{12}	$n_{1.}$
X_2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

se p-hodnota vypočítá následujícím způsobem

$$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

Pro větší tabulky je test složitější.

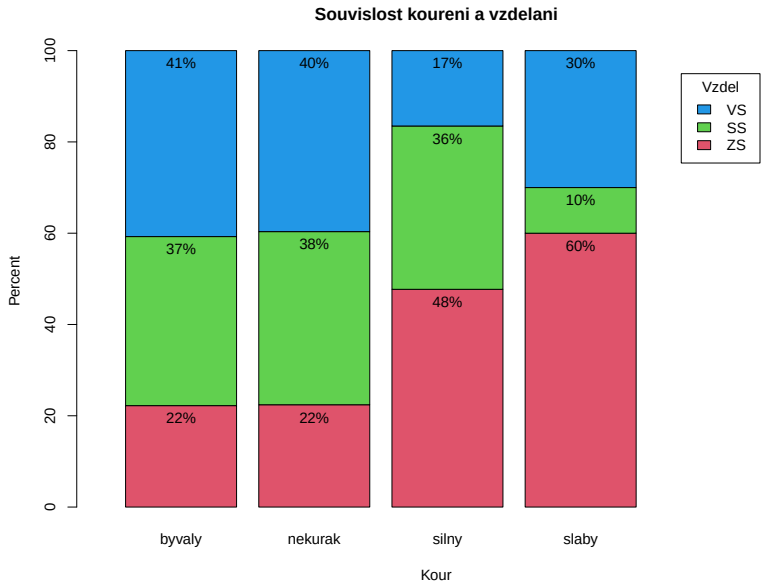
χ^2-test nezávislosti

Příklad. *U 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční bylo zjišťováno vzdělání a kategorie kouření. Výsledky jsou shrnuty v následující tabulce absolutních četností. Souvisí spolu tyto dvě veličiny?*

	ZŠ	SŠ	VŠ
bývalý kuřák	6	10	11
nekuřák	13	22	23
slabý kuřák	52	39	18
silný kuřák	6	1	3

χ^2-test nezávislosti

Vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

χ^2-test nezávislosti

Testované hypotézy

- H_0 : kouření se vzděláním nesouvisí
- H_1 : kouření se vzděláním souvisí

Výsledky testu

- testová statistika χ^2 testu $21.286 > 12.59$, kvantil χ^2-rozdělení s 6 stupni volnosti
- p-hodnota $0.00163 < \alpha = 0.05$
- p-hodnotu Fisherova exaktního testu $0.00084 < \alpha = 0.05$
- některé očekávané četnosti jsou menší než 5 (není splněn předpoklad χ^2 testu)
- na základě Fisherova testu zamítáme nulovou hypotézu

Závěr: Prokázali jsme, že kouření se vzděláním souvisí.

Poměr šancí

Uvažujme dvouhodnotovou veličinu ve dvou populacích.

- např. sledujeme výskyt chřipky ve městě a na venkově

.

	Chřipku má	Chřipku nemá	
Město	n_{11}	n_{12}	$n_{1.}$
Venkov	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Rozdíl mezi populacemi je možné popsat **poměrem šancí**.

- **šance** "mít chřipku proti nemít chřipku"

$$Odds = \frac{P(\text{má chřipku})}{P(\text{nemá chřipku})}$$

- poměr šancí je podíl šancí v obou populacích.

Poměr šancí

Definice **poměru šancí**

$$OR = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

Interpretace tohoto poměru říká, kolikrát je větší šance na chřipku ve městě než na venkově.

Testované hypotézy

- $H_0 : OR = 1$, šance jsou stejné
- $H_1 : OR \neq 1$, šance se v populacích liší

Testová statistika je rovna

$$Z = \frac{\ln(OR)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

a za platnosti nulové hypotézy má $N(0, 1)$ rozdělení.

Poměr šancí

Příklad. *Uvažujme následující čtyřpolní tabulku*

	Chřipku má	*Chřipku nemá*	
Město	58	17	75
Venkov	32	30	62
	90	47	137

- *šance mít chřipku ve městě vychází* $58/17 = 3.41$
- *šance mít chřipku na venkově vychází* $32/30 = 1.07$
- *poměr šancí ve městě vs. na venkově vychází* $3.41/1.07 = 3.2$
  Ve městě je více než třikrát větší šance mít chřipku než na venkově.
- *testová statistika* $3.27 > 1.96$ *kritická hodnota*
- *p-hodnota* $0.001 < \alpha = 0.05$
- *zamítáme nulovou hypotézu*

Závěr: *Ve městě je významně větší šance dostat chřipku než na venkově.*

Korelační koeficient

Je-li cílem výzkumu zjistit, zda spolu lineárně souvisí dvě číselné proměnné, používá se **korelační koeficient**. Rozlišujeme tři základní korelační koeficienty

- **Pearsonův** korelační koeficient – používá se pokud obě proměnné mají přibližně normální rozdělení
- **Spearmanův** korelační koeficient – používá se, pokud máme obě proměnné spojité, ale alespoň jedna z nich nemá normální rozdělení
- **Kendallův** korelační koeficient – používá se, pokud pracujeme s kategorickými uspořádanými (ordinálními) veličinami

Korelační koeficient

Korelační koeficient měří **lineární závislost** dvou proměnných hodnotou z intervalu $< -1, 1 >$:

- $\mathrm{Cor}(X, Y) = -1$ značí absolutní nepřímou závislost,
- $\mathrm{Cor}(X, Y) = 0$ značí lineární nezávislost/ nekorelovanost,
- $\mathrm{Cor}(X, Y) = 1$ značí absolutní přímou závislost.

Hodnota korelačního koeficientu říká, jak těsný je vztah mezi proměnnými.

Korelační koeficient

Pearsonův korelační koeficient vypočteme jako

$$\begin{aligned} \mathrm{Cor}(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \\ &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \end{aligned}$$

Spearmanův korelační koeficient se počítá dle stejného vzorce, jen místo původně naměřených hodnot se do něj vkládají pořadí.

Korelační test

Testované hypotézy

- H_0 : korelační koeficient $= 0$
- H_1 : korelační koeficient $\neq 0$,
  H_1 : korelační koeficient > 0,
  H_1 : korelační koeficient < 0

Za platnosti nulové hypotézy platí, že testová statistika pro

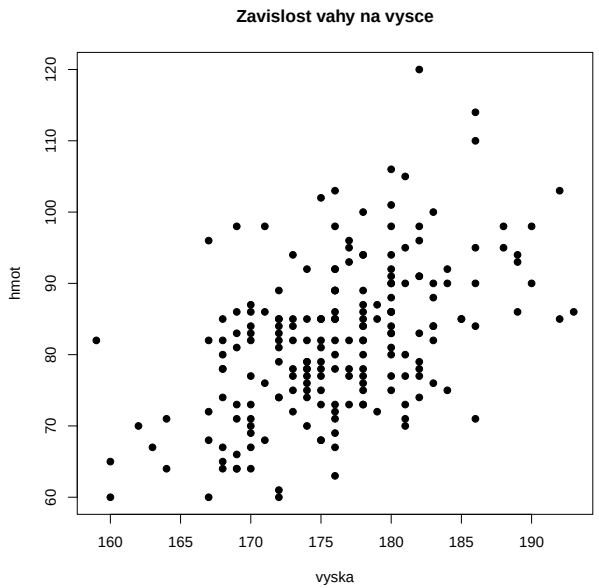
$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

má t-rozdělení o $n - 2$ stupních volnosti.

Korelační koeficient

Příklad. *Do výzkumu bylo zahrnuto 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční. Zjišťujeme, zda spolu souvisí výška a hmotnost těchto mužů?*

Nejprve grafické znázornění



Z grafu je patrná rostoucí závislost mezi oběma proměnnými.

Korelační koeficient

Testované hypotézy

- H_0 : váha a výška spolu nesouvisí, korelační koefcient $= 0$
- H_1 : váha a výška spolu souvisí, korelační koefcient $\neq 0$

Pearsonův korelační koeficient vyšel 0,5 a testová statistika

$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)} = \frac{0.5}{\sqrt{1 - 0.25}} \sqrt{202} = 8.19.$$

Testová statistika je větší než kvantil t-rozdělení $t_{202}(1 - 0.975) = 1.97$. P-hodnota testu vyšla $2.926 * 10^{-14}$, což je menší než $\alpha = 0.05$. Nulovou hypotézu tedy zamítáme.
Závěr: Souvislost mezi váhou a výškou je průkazná.
Spearmanův korelační koeficient vyšel 0.48.

Kendallův korelační koeficient

Jinak funguje **Kendallův korelační koeficient** (Kendallovo τ) pro ordinální veličiny.

- označme porovnávané proměnné X a Y
- dvojici pozorování označme jako souhlasnou (konkordantní), pokud $X_i < X_j \& Y_i < Y_j$ nebo $X_i > X_j \& Y_i > Y_j$
- dvojici pozorování označme jako nesouhlasnou (diskordantní), pokud $X_i < X_j \& Y_i > Y_j$ nebo $X_i > X_j \& Y_i < Y_j$

Kendallovo τ je založeno na rozdílu počtu souhlasných (n_s) a počtu nesouhlasných (n_n) dvojic.

Kendallův korelační koeficient

Kendallovo τ

$$\tau = \frac{n_s - n_n}{n} = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)$$

Rozptyl tohoto koeficientu je

$$\text{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

a testová statistika $\tau/\text{Var}(\tau)$ má za platnosti nulové hypotézy asymptoticky $N(0,1)$ rozdělení.

Kendallovo τ

Obdoba koeficientu, pokud se v datech opakují hodnoty
Pro proměnné se **stejným počtem možných hodnot**

$$\tau_B = \frac{n_s - n_n}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

kde $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$ a t_i jsou počty shodných hodnot u proměnné X, $n_1 = \sum_i u_i(u_i - 1)/2$ a u_i jsou počty shodných hodnot u proměnné Y.
Pro proměnné s **různým počtem možných hodnot**

$$\tau_C = \frac{2(n_s - n_n)}{n^2 \frac{m-1}{m}},$$

kde m je minimální počet hodnot u obou proměnných.
Výpočet rozptylů a následných testových statistik pro τ_B a τ_C je složitý. Přenechme ho tedy softwarům.

Lineární regrese

Příčinnou závislost popisuje **lineární regrese.**
Terminologie

- **nezávisle proměnná** X – příčinu
- **závisle proměnná** Y – důsledek

Odhad lineárního modelu ve tvaru

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

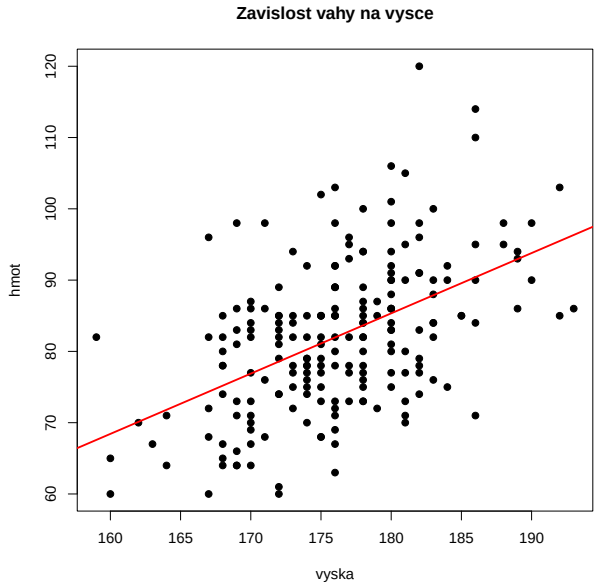
kde

- Y_i jsou hodnoty závisle proměnné
- X_i jsou hodnoty nezávisle proměnné
- β_0 je absolutní člen
- β_1 je lineární člen
- e_i jsou náhodné chyby

Lineární regrese

Graficky popisujeme pomocí bodového grafu, ale není jedno, která proměnná je na které ose

- na x-ovou osu se kreslí nezávisle proměnná
- na y-ovou osu se kreslí závisle proměnná



Regresní model je rovnice přímky proložené daty.

Metoda nejmenších čtverců

Odhad probíhá **metodou nejmenších čtverců**, která minimalizuje součet druhých mocnin residuí

$$\min \sum_{i=1}^{n} R_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

kde

- $\hat{Y}_i$ jsou odhady / predikce
- b_0 a b_1 jsou odhady regresních koeficientů

Pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné

$$\hat{Y}_0 = b_0 + b_1 x_0$$

např. ze známé výšky můžeme predikovat očekávanou hmotnost.

Koeficient determinace

Procento variability závisle proměnné vysvětlené modelem

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \mathrm{cor}(X, Y)^2$$

Na základě modelu lze též zkonstruovat **test nezávislosti**.
Testované hypotézy

- H_0 : Proměnná Y (váha) na proměnné X (výšce) lineárně nezávisí, $\beta_1 = 0$
- H_1 : Proměnná Y (váha) na proměnné X (výšce) lineárně závisí, $\beta_1 \neq 0$

Test je založen na faktu, že $b_1/\mathrm{se}(b_1) \sim N(0, 1)$, kde b_1 je odhad lineárního členu β_1 a $\mathrm{se}(b_1)$ je jeho střední chyba.

Lineární regrese

Příklad. *Pokračujme v příkladu s muži s jedním rizikovým faktorem ischemické choroby srdeční. Popište lineární závislost hmotnosti na výšce.*

- odhadnutý model má tvar

$$Y_i = -66.85 + 0.85 X_i$$

- střední chyba odhadu lineárního členu vyšla 0.1, testová statistika testu nezávislosti 8.19, p-hodnota $< 2.9 \times 10^{-14} < \alpha = 0.05$
- zamítáme nulovou hypotézu nezávislosti
- koeficient determinace vyšel 0.2493

Závěr: Můžeme tedy říci, že u mužů s jedním rizikovým faktorem ischemické choroby srdeční hmotnost na výšce závisí. Závislost je přímá a vysvětlí se jí 25% variability závisle proměnné (hmotnosti).

Předpoklady lineární regrese

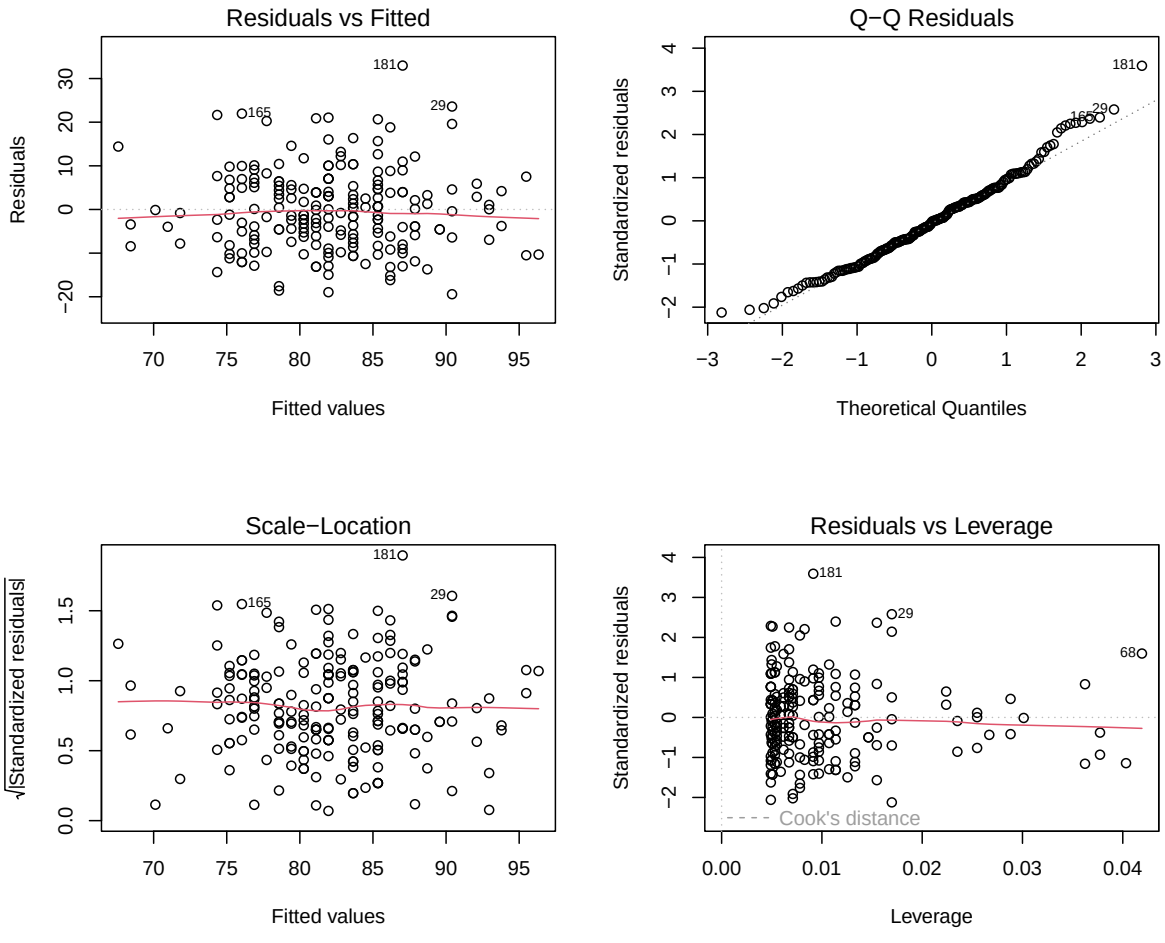
I lineární regrese má své **předpoklady**

- Mezi proměnnými je skutečně lineární vztah
- Residua jsou nezávislá
- Residua mají normální rozdělení
- Stabilita rozptylu
- V datech nejsou vlivná pozorování

Jednotlivé předpoklady můžeme hodnotit buď na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy.

Předpoklady lineární regrese

Ukázka grafických testů předpokladů



Předpoklady lineární regrese

Ukázka grafických testů předpokladů

- **1. graf**: lineární vztah – červená čára nemá mít trend
- **2. graf**: normalita residuí – body mají ležet na přímce
- **3. graf**: stabilita rozptylu – červená čára nemá mít trend
- **4. graf**: body nemají překročit meze (čárkované křivky)