Pravděpodobnost a Statistika

Alena Černíková

alena.cernikova@ujep.cz

7. prosince 2024

Podmínky zápočtu a zkoušky

- **Zápočet**
  dva domácí úkoly – opakování příkladů ze cvičení
  seminární práce – krátký ucelený text na zpracování dvou proměnných (od zadání až po závěr)
  důraz bude kladen na interpretaci
  odevzdávat přes moje internetové stránky – kapitola Pravděpodobnost a statistika – kombinačně, odkaz Úkoly
- **Zkouška** – ústní
  tři příklady u počítače
  jedna teoretická otázka

Obsah kurzu

- Definice pravděpodobnosti
- Nezávislost jevů a podmíněná pravděpodobnost
- Vybrané klasické pravděpodobnostní modely
- Pravděpodobnostní rozdělení
- Co je statistika
- Popisné statistiky
- Bodový vs intervalový odhad
- Základy testování
- Jednovýběrový, párový a dvouvýběrový test
- Analýza rozptylu
- Korelace
- Jednoduchá lineární regrese
- Vztah dvou kategorických proměnných

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

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

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

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 $A \in \mathcal{A}, A_1, A_2, \ldots \in \mathcal{A}, A_i \cap A_j = \emptyset$ pro všechna $i \neq j$ splňuje

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

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. Teoreticky se může stát, že orel nepadne 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).

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

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

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 součet bodů na dvou kostkách je větší nebo rovno 10, když víme, že na modré kostce 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

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

- **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*

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?

- **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.

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 srovnat 5 různě barevných hrnků na polici?
- **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?

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.

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné 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, . . .

Funkce určující rozdělení

- **Distribuční funkce** – $F(t) = P(X \leq t), t \in \mathbb{R}$
  – neklesající, zprava spojitá, obor hodnot je mezi 0 a 1
- **Pravděpodobnostní funkce** – $p(t) = P(X = t), t \in \mathbb{R}$
  – definovaná pouze pro diskrétní rozdělení
  – nespojitá, nenulová jen v hodnotách, kterých může náhodná veličina nabývat
- **Hustota** – $f(t) = \frac{d}{dt}F(t)$
  – definovaná pouze pro spojitá rozdělení – obdoba pravděpodobnostní funkce, ale nedefinuje konkrétní pravděpodobnosti
  – derivace funkce distribuční
  – pravděpodobnost jedné konkrétní hodnoty u spojitého rozdělení je 0

Střední hodnota a rozptyl

Další charakteristiky pro diskrétní i spojitá rozdělení

- Střední hodnota

$$E(X) = \sum_{i=1}^{n} x_i p_i, \qquad EX = \int_{-\infty}^{\infty} x f(x) dx$$

- Rozptyl

$$Var(X) = \sum_{i=1}^{n} (x_i - E(X))^2 p_i, \; Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

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
$cov(X, Y) \int_{-\infty}^{\infty} (x - E(X))(y - E(Y))f(x, y)dxdy$

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

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

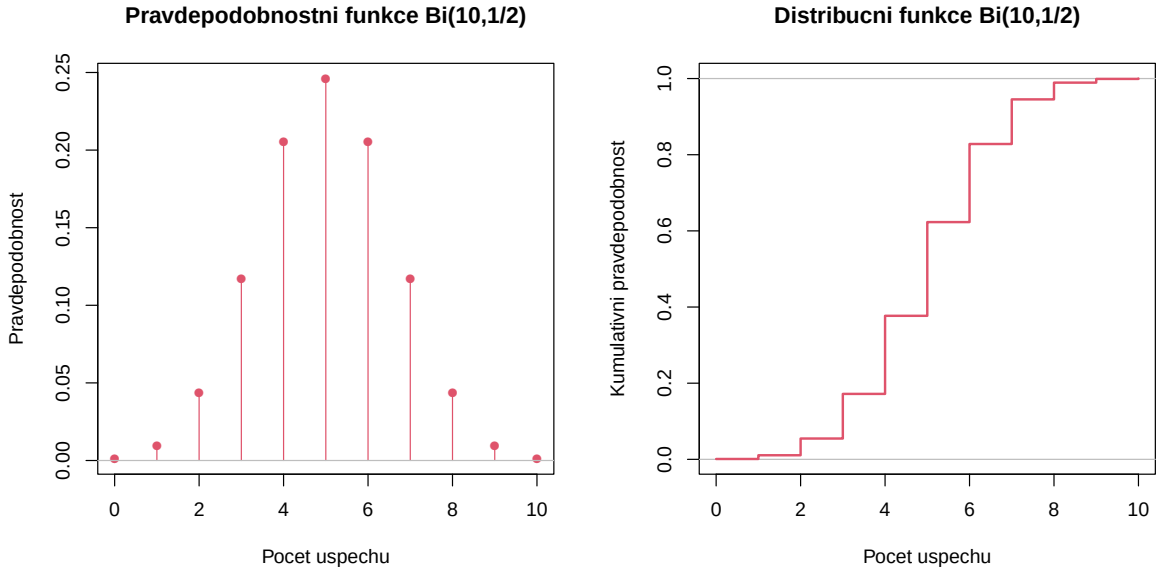
Střední hodnota a rozptyl

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

Binomické rozdělení

Příklad. *Házíme 10x mincí a počítáme, kolikrát padla panna. Počet pokusů je* $n = 10$*, pravděpodobnost úspěchu* $p = 1/2$*. Máme tedy 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

Př. 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 $k = 0, 1, 2, \ldots$

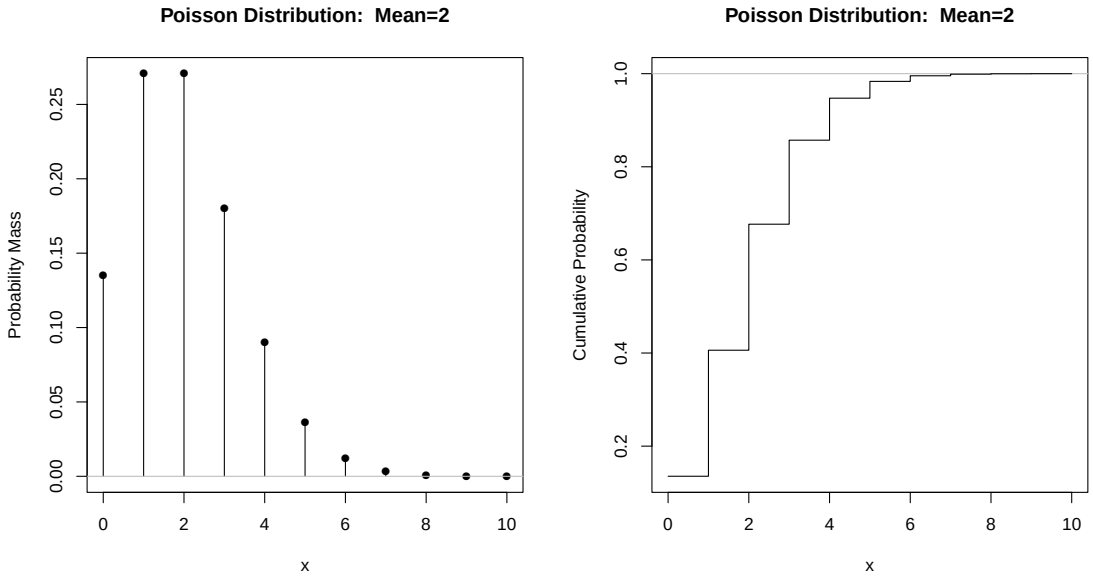
$$p(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

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 A jich je bílých a zbytek černých. Z urny postupně vytáhneme n koulí bez vracení. Náhodná veličina, která počítá počet bílých koulí mezi vytaženými má hypergeometrické rozdělení.

Značení $Hy(N, A, n)$, kde

- N – počet koulí v urně
- A – počet označených koulí v urně
- n – počet tažených koulí

Hodnoty pravděpodobnostní funkce

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

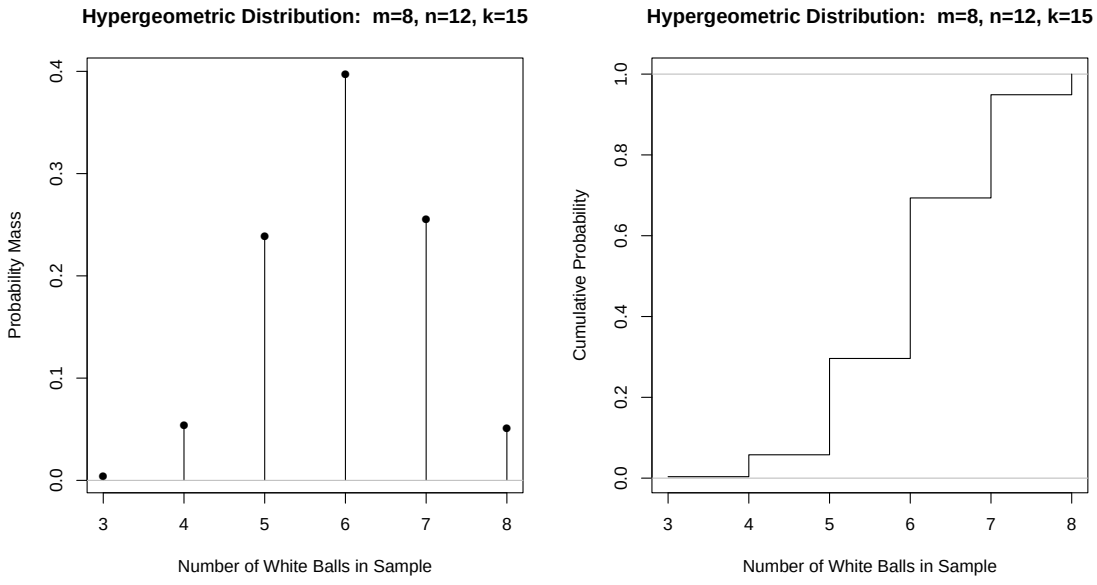
Střední hodnota a rozptyl

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

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

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

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

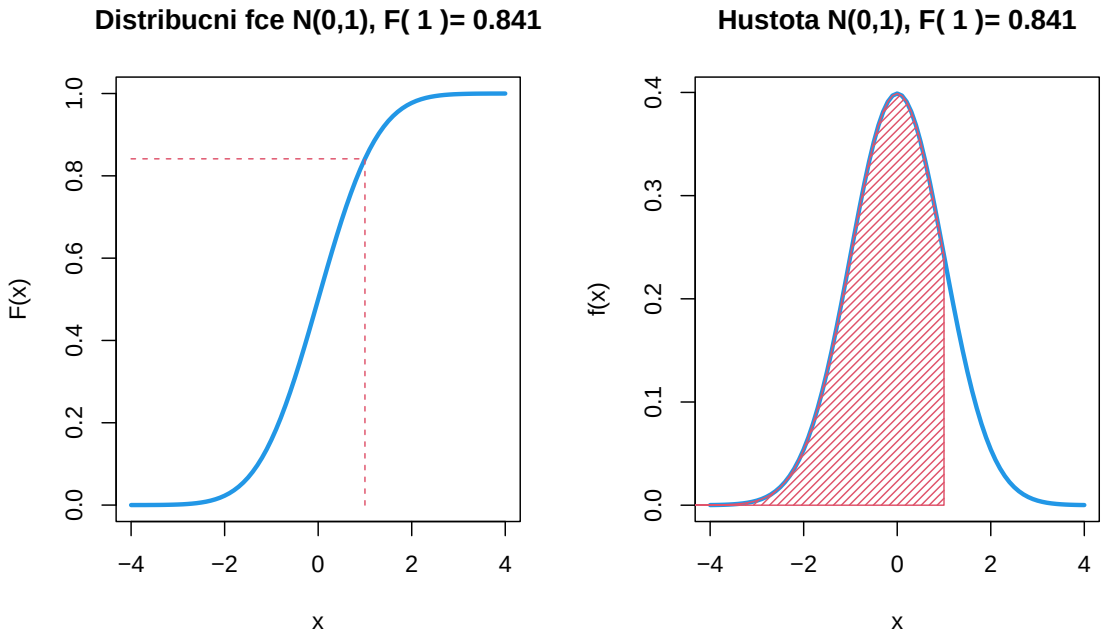
Hustota normálního rozdělení má tvar

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

Je to tak zvaná Gaussova křivka.

Normální rozdělení

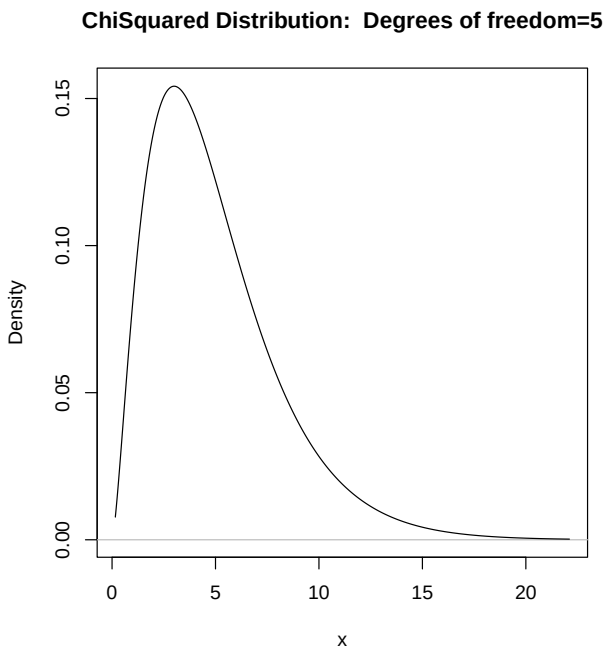
Vztah mezi hustotou a distribuční funkcí u standardního normálního rozdělení $N(0,1)$. Červeně je na obou grafech zobrazena stejná hodnota. Hustota a distribuční funkce.

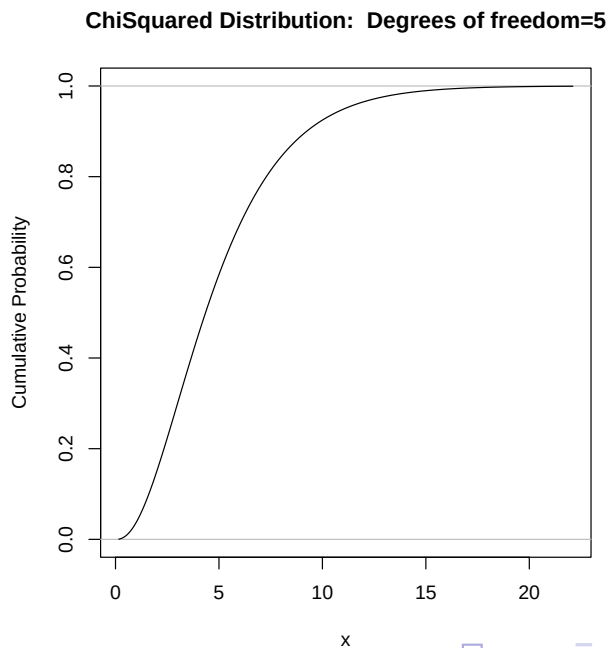


Předpokládejme binomické rozdělení $Bi(n,p)$, kde $0.1 \leq p \leq 0.9$, pak pro $n \to \infty$ toto rozdělení konverguje k normálnímu s parametry np, $np(1-p)$.

χ^2-rozdělení

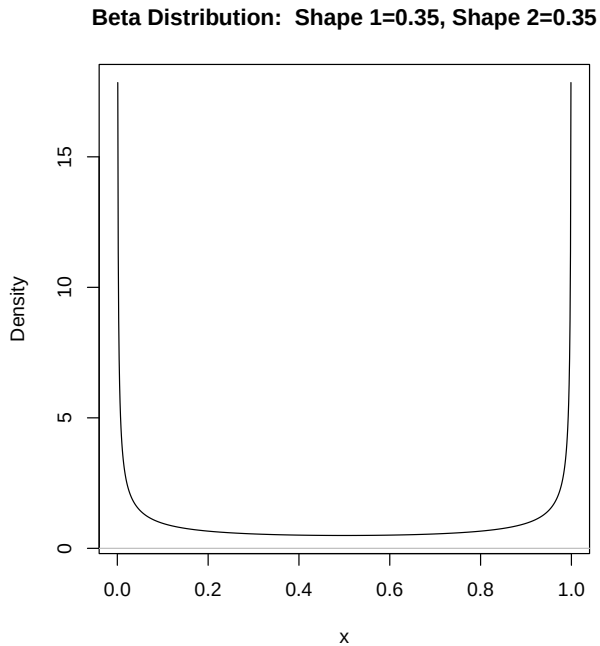
Rozdělení kvadratických forem. Náhodná veličina $Y = X_1^2 + X_2^2 + \cdots + X_n^2$, kde $X_i \sim N(0, 1)$ jsou nezávislé, má χ^2-rozdělení o n stupních volnosti. Dále je to rozdělení některých testových statistik, zejména těch, týkajících se rozptylu. Hustota a distribuční funkce χ^2-rozdělení o 5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

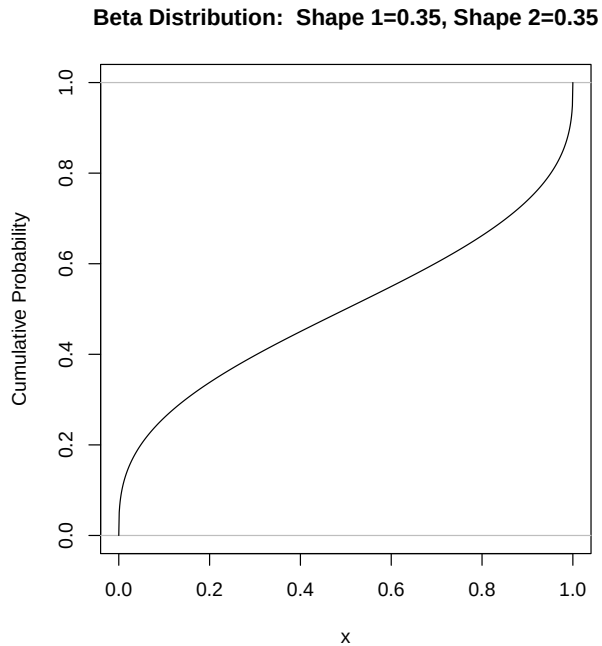




Beta rozdělení

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





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.

Typy proměnných

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

- **Číselné proměnné** – pr. výška, váha, věk, atd.
- **Kategorické proměnné** – pr. barva, kraj, povolání, nebo taky známka ve škole, číslo, které padne na kostce, atd.
- Kategorické proměnné se dále dělí na
  - **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

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

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.*

Jak vypočítat **průměr** z n hodnot značených $X_1, X_2, X_3, \ldots, X_n$

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

Jak vypočítat **medián**

- z uspořádané řady – hodnota prostřední podle velikosti, nebo průměr prostředních dvou

Jak vypočítat **kvartily**

- z uspořádané řady – hodnoty v jedné a ve třech čtvrtinách

Jak vypočítat **extrémy**

- minimum a maximum

Popisné statistiky polohy

Výpočet kvartilů podle **R**

Výpočet pro obecný p-tý percentil – vážený průměr dvou sousedních uspořádaných hodnot.
Označme

- p – číslo mezi 0 a 1, díl dat, které chcete p-tým percentilem oddělit
- $X_{(k)}$ – hodnoty z uspořádané řady, k-tý nejmenší prvek
- q – koeficient, kterým se násobí uspořádané hodnoty do váženého průměru

$$\begin{aligned} p-percentil &= (1-q)X_{(k)} + qX_{(k+1)} \\ k &= \lfloor 1 + (n-1)p \rfloor \\ q &= 1 + (n-1)p - k \end{aligned}$$

Grafické popisné statistiky

Pro popis číselné proměnné se používají 2 typy grafů

- **Krabicový graf**
  jsou v něm zobrazeny vybrané percentily (medián a kvartily), tykadla dosahují k nejvzdálenějšímu neodlehlému pozorování (odlehlé pozorování se vyznačují zvlášť)
  odlehlé pozorování je takové, které je od bližšího kvartilu dále než jeden a půl násobek mezikvartilového rozpětí $1.5(Q_3 - Q_1)$
- **Histogram**
  počet sloupců je určen vybraným pravidlem
  nejčastěji se používá *Sturgesovo pravidlo*

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

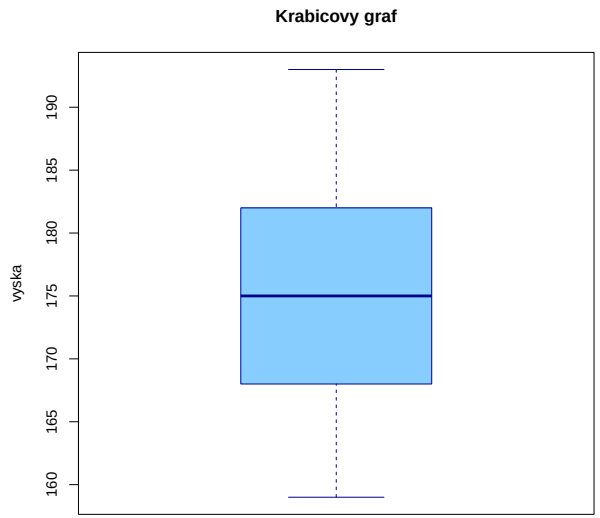
kde n je počet pozorování

Popisné statistiky polohy

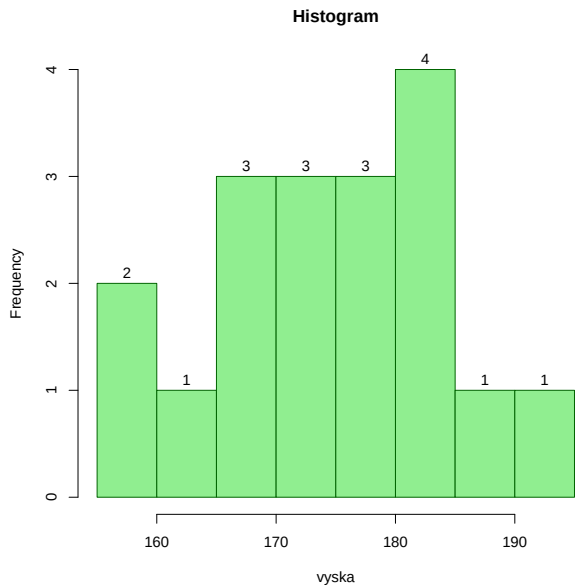
Výsledky pro výšky dospělých mužů

- průměr – 174.89
- medián – 175
- kvartily – 168, 181.75
- extrémy – 159, 193

Grafy



1. Krabicový graf



2. Histogram

Popisné statistiky variability

- Rozptyl a směrodatná odchylka

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

- Mezikvartilové rozpětí

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

kde Q_3 je třetí kvartil a Q_1 je první kvartil

- Variační koeficient

$$\mathrm{cv}(X) = \frac{\mathrm{sd}(X)}{\overline{X}}$$

Popisné statistiky tvaru rozdělení

Počítají se ze standardizovaných proměnných, tak zvaných **Z-skórů**

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

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

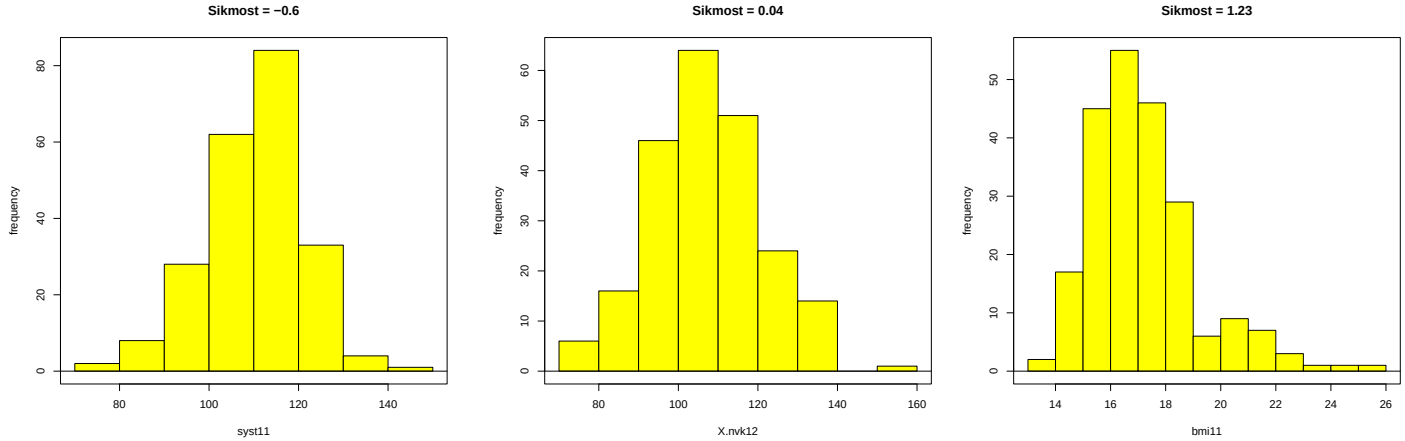
$$\mathrm{Skew}(X) = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\mathrm{sd}(X)}\right)^3 = \frac{\sum_{i=1}^{n} Z_i^3}{n}$$

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

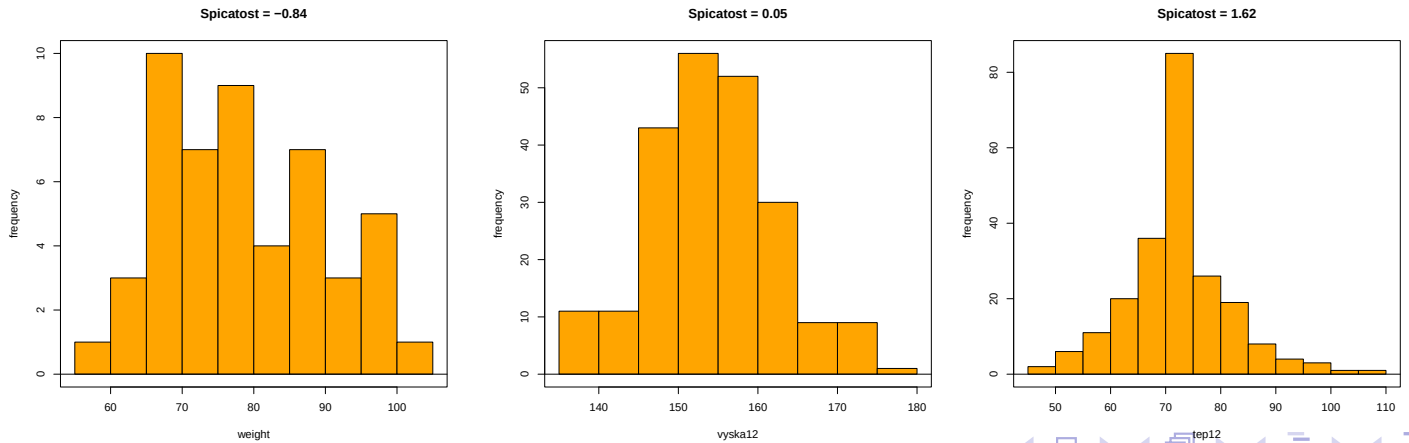
$$\mathrm{Kurt}(X) = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\mathrm{sd}(X)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$$

Popisné statistiky tvaru rozdělení

Záporná, nulová a kladná šikmost



Záporná, nulová (špičatost normálního rozdělení) a kladná špičatost



Popisné statistiky variability a tvaru rozdělení

Popisné statistiky variability – výsledky

- rozptyl – 88.81
- směrodatná odchylka – 9.42
- mezikvartilové rozpětí – 13.75
- variační koeficient – 0.054

Popisné statistiky tvaru rozdělení – výsledky

- šikmost – 0.027
- špičatost – -1.04

Otázky na promyšlení

- Kdy kterou charakteristiku použít a proč
- Jaké mají jednotlivé statistiky rozměry
- Jak se jednotlivé statistiky mění v závislosti na posunutí a změně měřítka u původní veličiny

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.

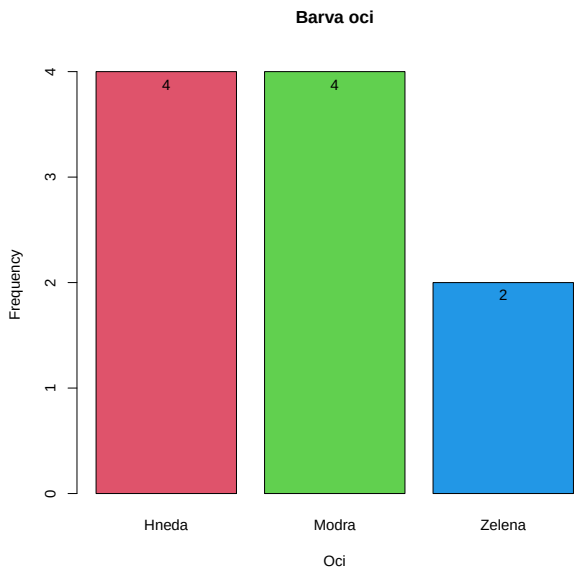
Barva	Značení	Absolutní	Relativní %
Modrá	n_1	4	40%
Hnědá	n_2	4	40%
Zelená	n_3	2	20%
Celkem	n	10	100%

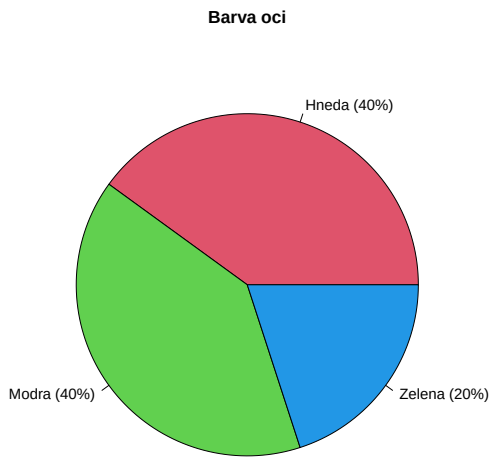
Relativní četnost vypočteme jako $p_j = \frac{n_j}{n}$

Grafické popisné statistiky nominální proměnné

Sloupcový a koláčový graf

- zobrazují se v absolutních počtech, nebo v procentech

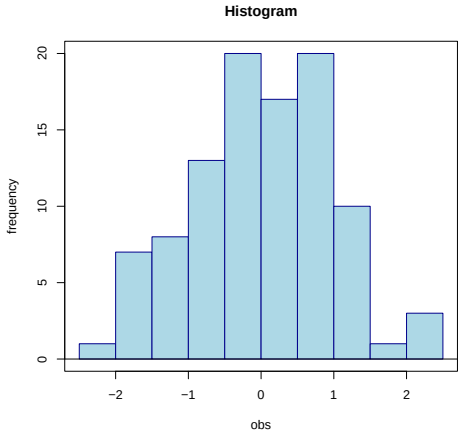


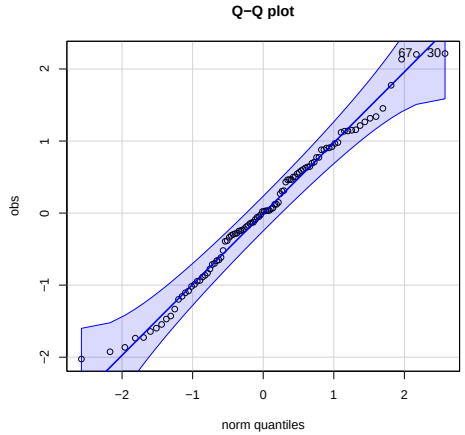


Testy normality

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

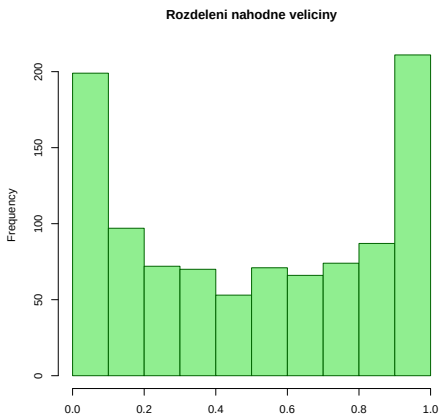
Centrální limitní věta

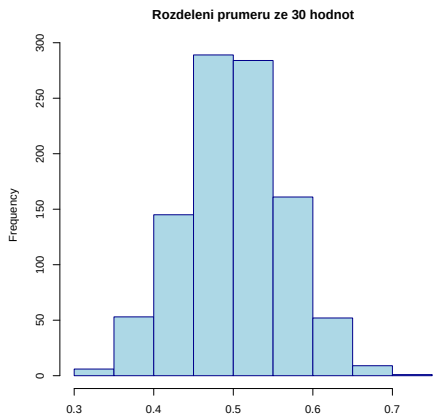
Věta

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 jejích rozdělení pro počet těchto náhodných veličin rostoucí nade všechny meze.

V praxi to znamená, že výběrový průměr dobře odhaduje skutečnou střední hodnotu a že se jedná o tzv. nestranný odhad.

Nestranný odhad

- střední hodnota odhadu se rovná odhadovanému parametru
- $\mathrm{E}\overline{X} = \mathrm{E}X$

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 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

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

Intervalový odhad střední hodnoty

Chceme interval, ve kterém se s vysokou pravděpodobností bude nacházet skutečný populační průměr/ skutečná střední hodnota.
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

Základem pro interval spolehlivosti je fakt, že výběrový průměr má asymptoticky normální rozdělení (CLV)

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

kde

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

Interval spolehlivosti pro střední hodnotu má tvar

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

kde $q(1-\alpha/2)$ je kvantil teoretického rozdělení

- znám-li skutečný rozptyl σ^2
  používá se kvantil standardního normálního rozdělení $N(0,1)$
- musím-li odhadnout σ^2 pomocí výběrového rozptylu
  používá se kvantil t-rozdělení o $n-1$ stupních volnosti

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

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

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

$$172.34, 173.9$$

Se splehlivostí 95% bude skutečný populační průměr výšky mužů ležet v intervalu od 172.34 cm do 173.9 cm.

Bootstrapový interval spolehvosti pro střední hodnotu

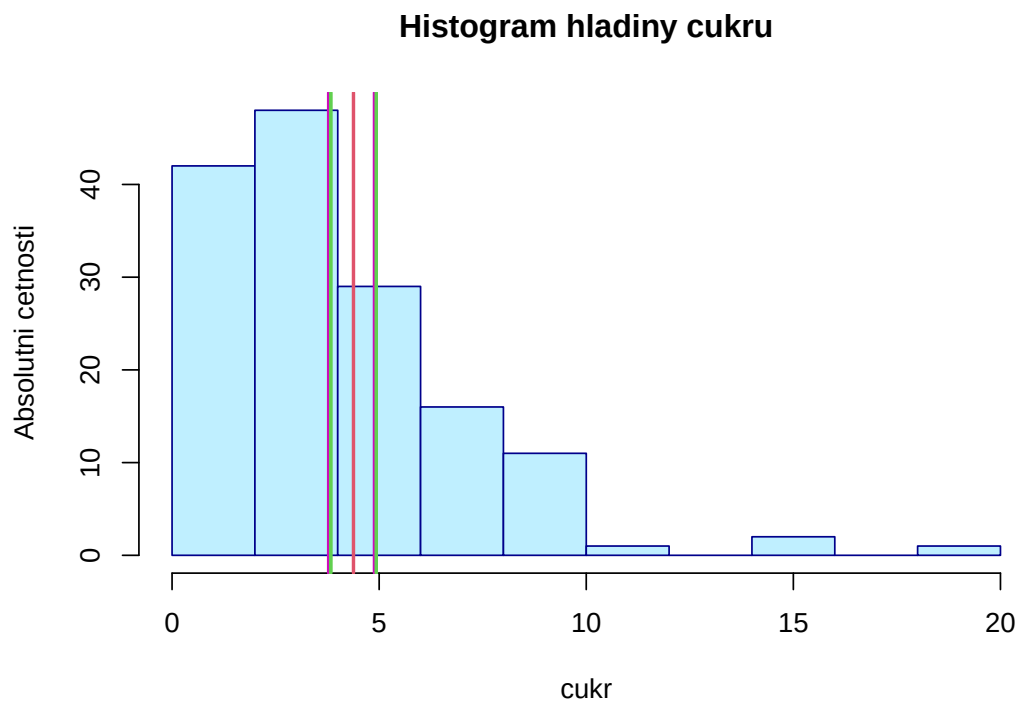
Konstrukce intervalu

- 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 $B = 10000$.

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 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

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

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

- ověřit normalitu testované proměnné
- souvislost mezi statistikou T a intervalem spolehlivosti

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.

Znaménkový test

Neparametrický test o střední hodnotě

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

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
- 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

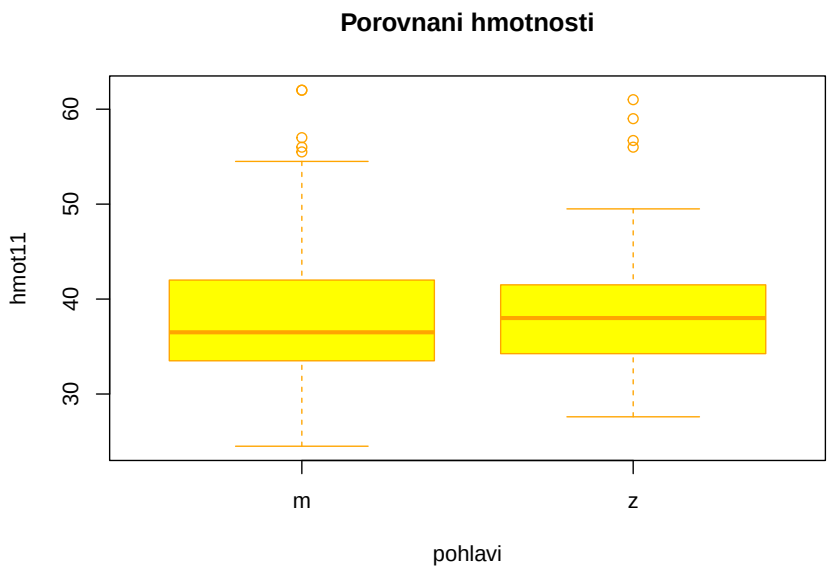
- testová statistika $F = \frac{\text{Var(X)}}{\text{Var(Y)}} = \frac{45.1}{50.6} = 0.89$
- p-hodnota $= 0.56 > \alpha = 0.05$
- nulovou hypotézu nezamítáme
- rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly

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.

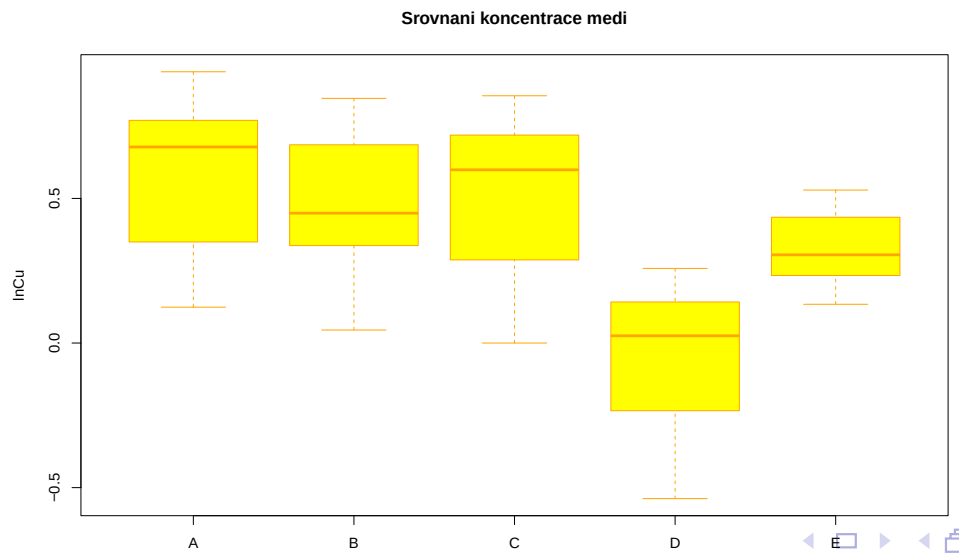
Analýza rozptylu – ANOVA

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.

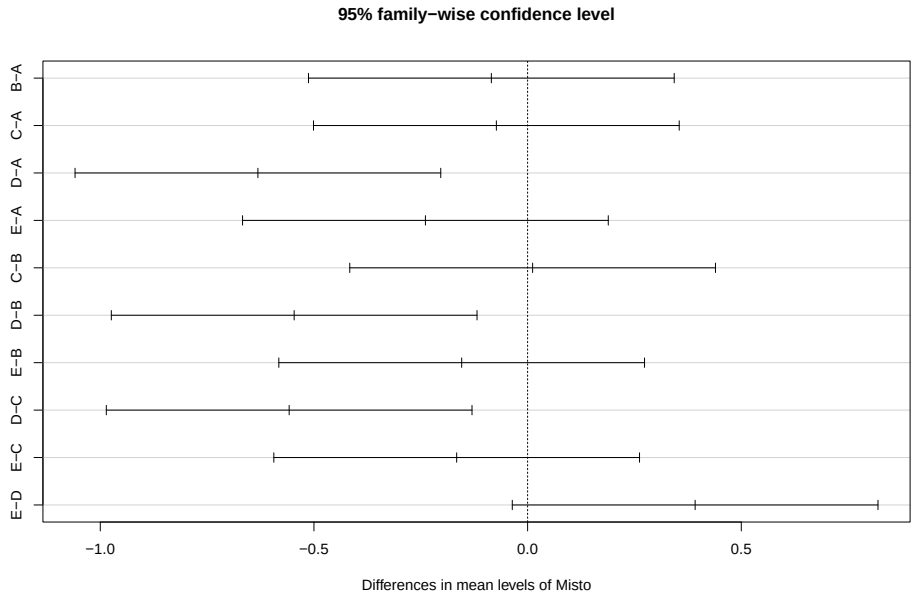
Analýza rozptylu – ANOVA

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

Vyhodnocení hypotéz probíhá opět pomocí porovnáná variability mezi výběry, ale tentokrát s variabilitou zbytkovou. Zbytková variabilita se od celkové variability v rámci výběrů liší tím, že je od snížena o variabilitu způsobenou rozdíly mezi jedinci. Konkrétně se tato zbytková variabilita získá následovně

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

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)
- proněnná 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

Předpokladem χ^2-testu je, že všechny očekávané četnosti jsou větší než 5. Pokud předpoklad není splněn, používá se **Fisherův exaktní test**, známý též jako **Fisherův faktoriálový test**. Tento test počítá přímo p-hodnotu, tj. pravděpodobnost, že za platnosti H_0 bude pozorována právě naše tabulka četností.
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.

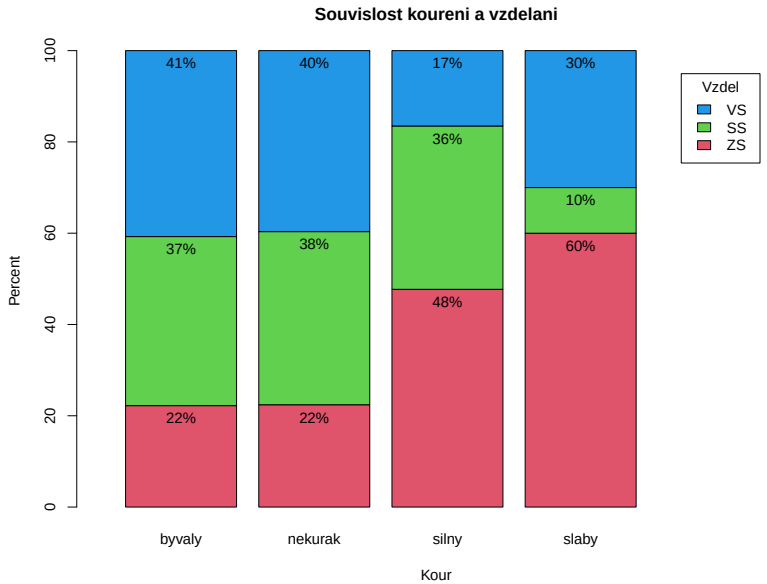
χ^2-test nezávislosti

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

Testem nezávislosti jsme zjišťovali

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

Testová statistika vyšla 21.286. Porovnáváme ji s kvantilem χ^2-rozdělení $\chi^2_6 = 12.59$. Jelikož testová statistika vyšla vyšší, tak zamítáme nulovou hypotézu. P-hodnota testu vyšla 0.00163, což je menší než $\alpha = 0.05$.
Jelikož však nejsou splněny předpoklady testu, měli bychom vypočítat ještě p-hodnotu Fisherova exaktního testu. Ta vychází 0.00084.
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. Výsledky je možné zapsat do čtyřpolní tabulky

	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. Nejprve definujme **šanci** "mít chřipku proti nemít chřipku"jako

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

Poměr šancí je pak podíl této šance v jedné populaci ku šanci v druhé populaci.

Poměr šancí

Pro naši tabulku je pak **poměr šancí** definovaný jako

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

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

Pokud chceme otestovat, že šance na chřipku jsou stejné ve městě jako na venkově, testujeme

- $H_0 : OR = 1$
- $H_1 : OR \neq 1$

Testová statistika tohoto testu 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 vychází 3.27, kritická hodnota 1.96 a p-hodnota 0.001. Jelikož testová statistika je větší než kritická hodnota a p-hodnota je menší než* α, *zamítáme nulovou hypotézu. 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**. Podle typu dat vybíráme konkrétní korelační koeficient, který na měření vzájemné souvislosti použijeme:

- **Pearsonův** korelační koeificient – 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

Korealční koeficient měří lineární závislost 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 koeficient

O statistické významnosti závislosti rozhodujeme testem

- 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

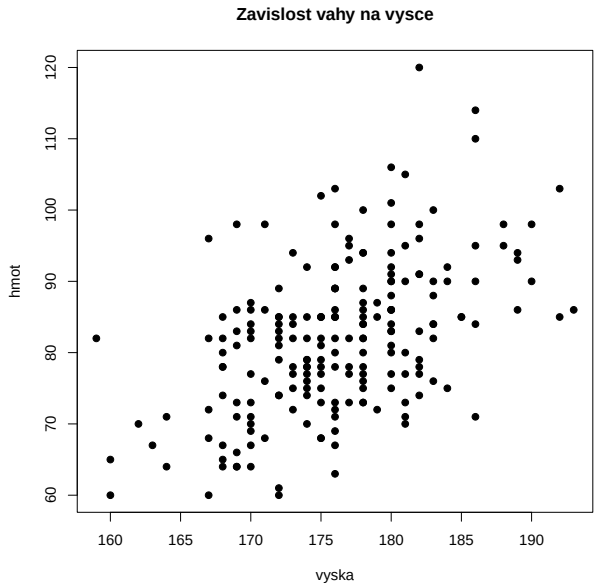
$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

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

Korelační koeficient

Příklad. *Do výzkumu bylo zahrnuto 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční. U těchto mužů byly měřeny různé charakteristiky. Souvisí spolu výška a hmotnost těchto mužů?*

Nejprve grafické znázornění



Z grafu je patrná rostoucí závislost mezi oběma proměnnými.

Korelační koeficient

Dále jsme testovali

- 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. 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 dvě porovnávané proměnné X a Y. Nyní uvažujme všechny dvojice naměřených hodnot (X_i, Y_i) a (X_j, Y_j). Pokud pro danou dvojici platí, že $X_i < X_j \& Y_i < Y_j$ nebo $X_i > X_j \& Y_i > Y_j$, pak označme tuto dvojici jakou souhlasnou, pokud platí $X_i < X_j \& Y_i > Y_j$ nebo $X_i > X_j \& Y_i < Y_j$, označme ji za nesouhlasnou.

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

Kendallův korelační koeficient

Konkrétně je **Kendallovo** τ definováno jako

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

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

Výše uvedený koeficient funguje dobře, pokud v datech nejsou stejné hodnoty. Pokud se stejné hodnoty vyskytnou, používají se následující obdoby tohoto koeficientu.
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

Vztah mezi dvěma spojitými proměnnými lze hodnotit i z pohledu **lineární regrese**, která zkoumá příčinnou závislost. V tomto případě máme

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

Výsledkem je odhad lineárního modelu ve tvaru

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

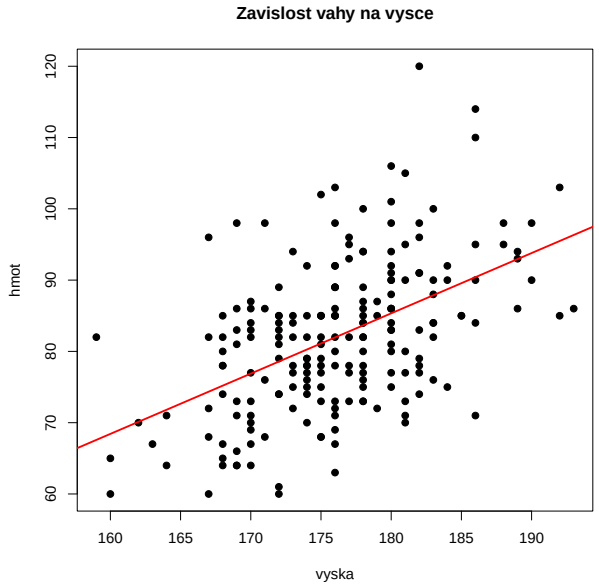
kde

- 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.

Lineární regrese

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 $\hat{Y}_i$ se nazývají odhady, nebo též predikce, b_0 a b_1 jsou odhady regresních koeficientů. Pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné.

Pro hodnotu x_0 nezávisle proměnné X očekáváme hodnotu

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

např. ze známé výšky můžeme predikovat očekávanou hmotnost.

Lineární regrese

Koeficient determinace

Zajímavý ukazatel je koeficient determinace

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

Říká, kolik procent variability závisle proměnné se modelem vysvětlí. Jinými slovy, z kolika procent závisle proměnná závisí na X a z kolika na něčem jiném.
Na základě modelu lze též zkonstruovat **test nezávislosti**. Testujeme

- 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.*

Odhadli jsme model ve tvaru

$$Y_i = -66.85 + 0.85X_i$$

Střední chyba odhadu lineárního členu vyšla 0.1 a testová statistika tedy 8.19. Tu jsme porovnali s kvantilem t-rozdělení $t_{220}(1 - 0.975) = 1.97$. Jelikož je testová statistika větší, tak zamítáme nulovou hypotézu. P-hodnota testu vyšla $< 2.9 * 10^{-14}$, což je menší než $\alpha = 0.05$. 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).

Lineární regrese

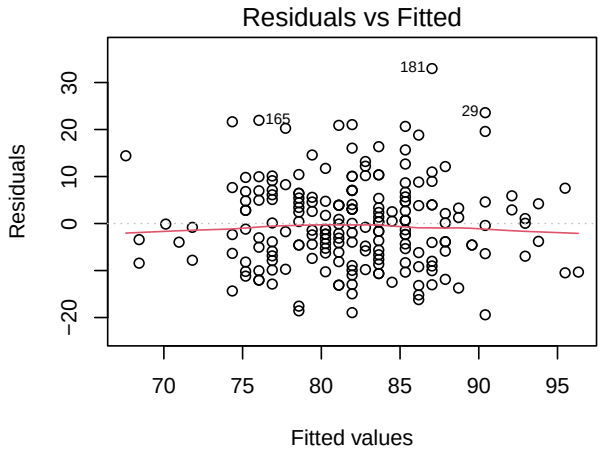
I lineární regrese má své **předpoklady**

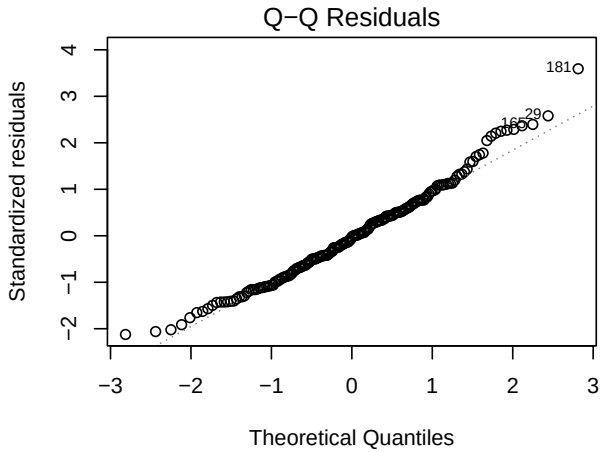
- Mezi proměnnými je skutečně lineární vztah
- Residua jsou nezávislá
- Residua mají normální rozdělení
- Stabilita rozptylu
- V datech nejsou vlivná pozorování

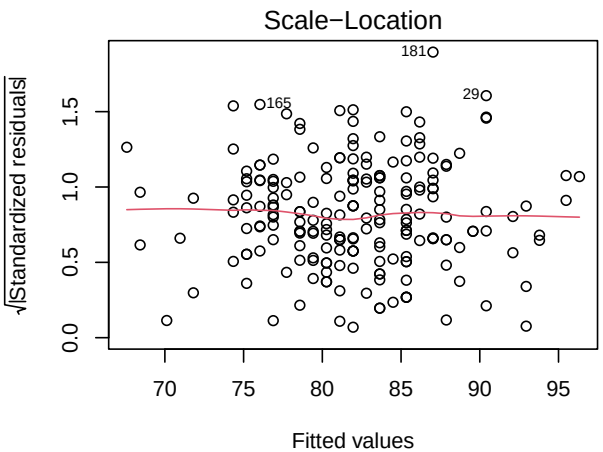
Jednotlivé předpoklady můžeme hodnotit buď na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy.

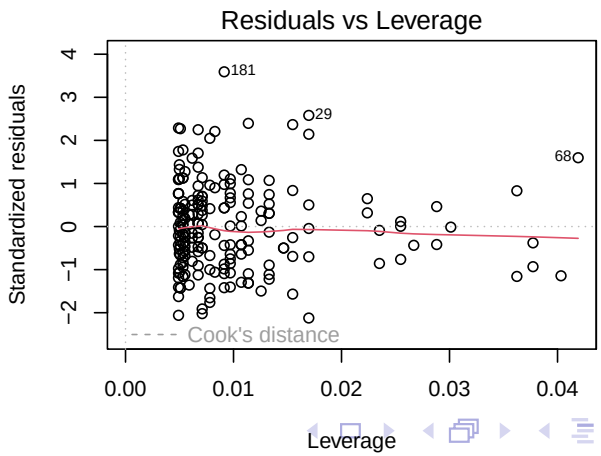
Lineární regrese

Ukázka grafických testů předpokladů









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)