

Pravděpodobnost a Statistika

Alena Černíková

alena.cernikova@ujep.cz

8. prosince 2023

- **Zápočet**

dva domácí úkoly – jednoduché opakování příkladů ze cvičení

seminární práce – krátký ucelený text na zpracování dvou proměnných (od zadání až po závěr)

odevzdávat na univerzitní OneDrive – bude upřesněno později

- **Zkouška** – ústní

tři příklady u počítače

jedna teoretická otázka

- 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 zkoušku udělá nebo 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 (Ω, \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 $A \in \mathcal{A}$, $A_1, A_2, \dots \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ pro všechna $i \neq j$ splňuje

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

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 sje 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 psat, 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. \bar{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

Definice pravděpodobnosti

Náhodné jevy a jejich pravděpodobnosti

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

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

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. $H_1 \cup H_2 \cup \dots \cup H_k = \Omega$ 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 $P(B|A)$

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

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

termíny používané v medicíně

- **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 specifická testu

Příklad. Výzkumu se zúčastnilo 2000 pacientů, z nichž 50 bylo HIV pozitivních. Všichni podstoupili test na HIV. Test vyšel pozitivní pro 45 pozitivních pacientů a pro 200 negativních. Spočítejte senzitivitu a specificku testu a také pravděpodobnost, že člověk bude skutečně HIV pozitivní, pokud mu vyjde pozitivní test.

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

- **Senzitivita testu** – $P(\text{test je pozitivní}|\text{osoba je nemocná}) = 45/50 = 0.9$
- **Specifická 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

$$P(ON|TP) = \frac{P(ON \cap TP)}{P(TP)} = \frac{P(TP|ON)P(ON)}{P(TP|ON)P(ON) + P(TP|OZ)P(OZ)} =$$
$$= \frac{\text{Senzitivita} * \text{podíl nemocných}}{\text{Senzitivita} * \text{podíl nemocných} + (1 - \text{Specifická}) * \text{podíl zdravých}} = \frac{0.9 * 0.025}{0.9 * 0.025 + 0.102 * 0.975} = 0.184$$

- **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.

- **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 mohou 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 mohou vybrat 5 žáků ke zkoušení z dvaceti přítomných ve třídě?

Náhodná procházka

Uvažujme částici, která se pohybuje po celočíselné přímce \mathbb{Z} a označme jako S_k její polohu v čase $k = 0, 1, 2, \dots$.

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 .

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

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

Tento model byl podrobně zkoumán a 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 dělíme podle typu proměnné na

- **Spojité** – 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

Další charakteristiky pro diskrétní i spojitá rozdělení

- Střední hodnota

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

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

- Rozptyl

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

Vlastnosti střední hodnoty a rozptylu

- Střední hodnota

$$E(a + bX) = a + bEX$$

$$E(X + Y) = E(X) + E(Y)$$

- Rozptyl

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X, Y)$$

kde $\text{cov}(X, Y)$ je kovariance počítaná jako

$\text{cov}(X, Y) \sum_{i=1}^n (X_i - E(X))(Y_i - E(Y))p_i q_i$ nebo

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

Pravděpodobnostní rozdělení

Binomické rozdělení – zástupce diskretních 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

$$E(X) = np,$$

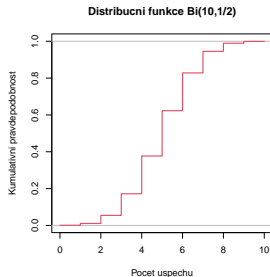
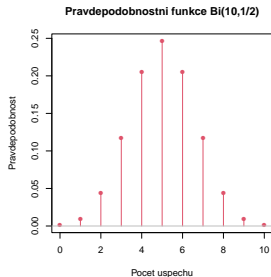
$$\text{Var}(X) = np(p - 1)$$

Pravděpodobnostní rozdělení

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

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

Poissonovo rozdělení – zástupce diskretních 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 $k = 0, 1, 2, \dots$

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

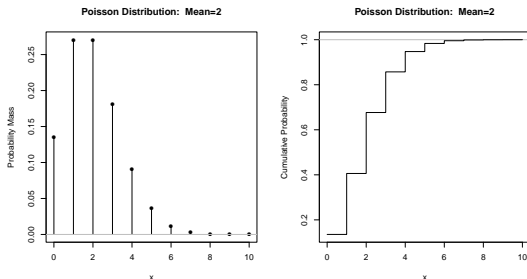
Střední hodnota a rozptyl

$$E(X) = \lambda,$$

$$\text{Var}(X) = \lambda$$

Poissonovo rozdělení

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



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

Pravděpodobnostní rozdělení

Hypergeometrické rozdělení – zástupce diskrétních 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

$$E(X) = \frac{nA}{N},$$

$$\text{Var}(X) = n \frac{A}{N} \left(1 - \frac{A}{N}\right) \left(\frac{N-n}{N-1}\right)$$

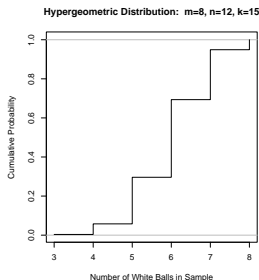
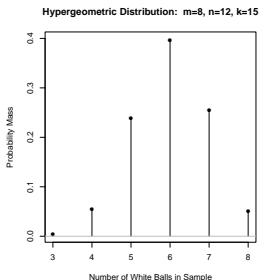


Pravděpodobnostní rozdělení

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

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

Normální rozdělení – zástupce spojitých 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

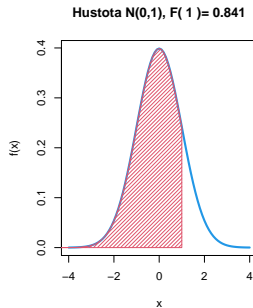
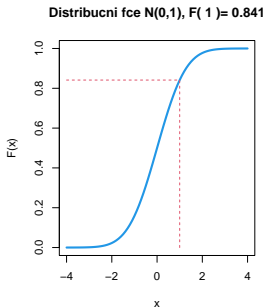
Hustota normálního rozdělení má tvar

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

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 $n \rightarrow \infty$ toto rozdělení konverguje k normálnímu s parametry $np, np(1 - p)$.

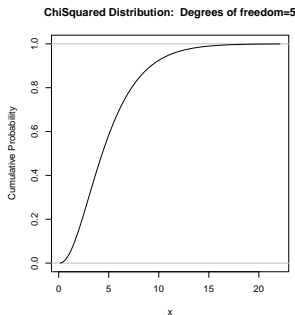
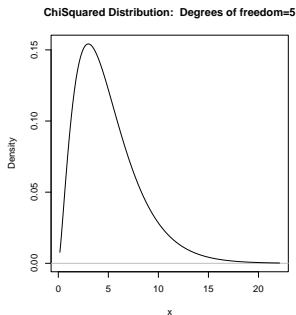
Pravděpodobnostní rozdělení

χ^2 -rozdělení – zástupce spojitych rozdělení

Rozdělení kvadratických forem. Náhodná veličina

$Y = X_1^2 + X_2^2 + \dots + X_n^2$, kde $X_i \sim N(0, 1)$ jsou nezávislé, má

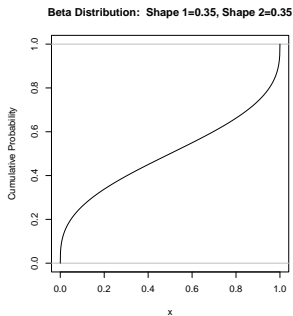
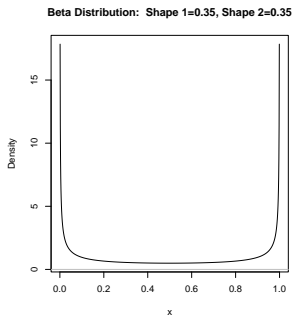
χ^2 -rozdělení o n stupních volnosti. Dále je to rozdělení některých testových statistik, zejména těch, týkajících se rozptylu. Hustota a distribuční funkce χ^2 -rozdělení o 5 stupních volnosti



Pravděpodobnostní rozdělení

Beta – zástupce spojitých rozdělení

Rozdělení pravděpodobností nějakého jevu. Např. sledujeme pravděpodobnost, že vybrný 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



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.*

- **Nahodná veličina** – jakákoliv veličina, kterou měříme, zde výška
- **Populace** – soubor, pro nějž 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.

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

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 charakteristiky – krabicový graf, histogram

- **Nominální proměnné**

- číselné charakteristiky – 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 – aneb co dělat když

- **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 $X_1, X_2, X_3, \dots, X_n$

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

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

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

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

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

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*

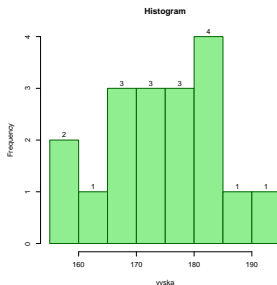
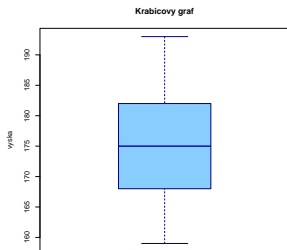
$$k = 1 + 3.32 \log_{10}(n)$$

kde n je počet pozorování

Popisné statistiky polohy – výsledky

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

Grafy



Popisné statistiky variability

- Rozptyl a směrodatná odchylka

$$\text{Var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}, \quad \text{sd}(X) = \sqrt{\text{Var}X}$$

- Mezikvartilové rozpětí

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

kde Q_3 je třetí kvartil a Q_1 je první kvartil

- Variační koeficient

$$\text{cv}(X) = \frac{\text{sd}(X)}{\bar{X}}$$

Popisné statistiky tvaru rozdělení

Pro obě statistiky (šíkmost i špičatost) je třeba nejprve spočítat standardizované proměnné, tak zvané **Z-skóry**

$$Y_i = \frac{X_i - \bar{X}}{\text{sd}(X)}$$

- **Šíkmost** – průměr ze třetích mocnin z-skóru

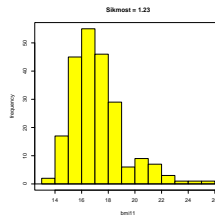
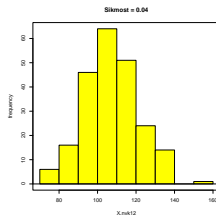
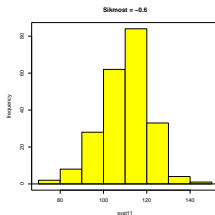
$$\text{Skew}(X) = \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{\text{sd}(X)} = \frac{\sum_{i=1}^n Z_i^3}{n}$$

- **Špičatost** – průměr ze čtvrtých mocnin z-skóru minus 3

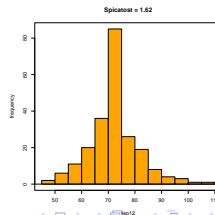
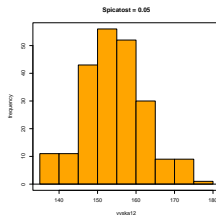
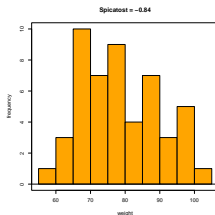
$$\text{Kurt}(X) = \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{\text{sd}(X)} - 3 = \frac{\sum_{i=1}^n Z_i^4}{n} - 3$$

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



Ukázka záporné, nulové (špičatost normálního rozdělení) a kladné špičatosti



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

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	Absolutní	Relativní %
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

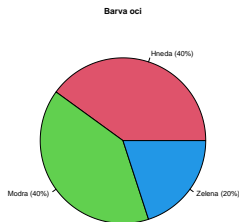
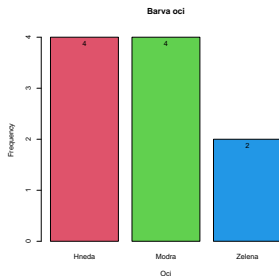
Jak vypočítat **relativní četnost**?

Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

$$p_j = \frac{n_j}{n}$$

Grafické popisné statistiky

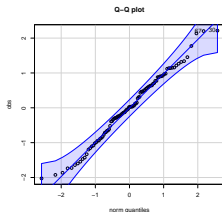
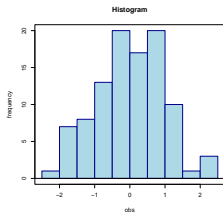
Sloupcový a koláčový graf – je možné je popisovat v absolutních počtech, nebo v procentech



Většina statistických postupů, odhadů a testů je odvozena právě pro normální rozdělení. Je proto dobré zjistit, zda náhodná veličina normální rozdělení má či nemá.

K tomuto účelu se využívají

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



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

Centrální limitní věta

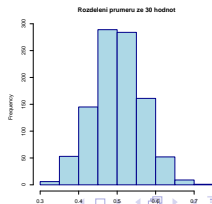
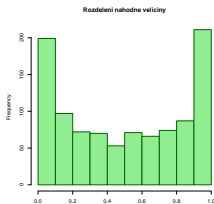
U testů pracujících s průměrem stačí, když má normální rozdělení průměr. Většinu případů lze tedy vyřešit pomocí **Centrální limitní věty**.

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.

Ukázka, jak vypadá rozdělení průměru 30-ti hodnot z beta rozdělení v porovnání s rozdělením samotným.



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 $\bar{X} = 173.12$
- jaká je pravděpodobnost, že se populační průměr bude rovnat přesně tomuto číslu?
- jaká je chyba tohoto odhadu
- střední chyba odhadu průměru

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

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

Výpočet intervalu spolehlivosti vychází z faktu, že výběrový průměr má normální rozdělení

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

kde μ je odhadovaná teoretická střední hodnota, σ je teoretická směrodatná odchylka a n je počet pozorování.

Když znám skutečný rozptyl dat, pak interval spolehlivosti pro střední hodnotu má tvar

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

kde $z(1 - \alpha/2)$ je kvantil standardního normálního rozdělení.

Častější je případ, že rozptyl neznám, pak

$$\frac{\bar{X} - \mu}{\text{sd}(X)/\sqrt{n}} \sim t_{n-1},$$

a meze intervalu spolehlivosti pak jsou

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

kde $t_{n-1}(1 - \alpha/2)$ je 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. Víme $\bar{X} = 173.12$, $\text{sd}(X) = 8.9$, $n = 500$, $\alpha = 0.05$. Interval spolehlivosti tedy je

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

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

$$172.34, 173.9$$

Výše uvedený intervalový odhad je založen na teoretických znalostech. V poslední době se těší oblibě **Bootstrapový interval spolehlivosti**, který nepotřebuje teorii, ale využívá dostupnosti výkonné výpočetní techniky.

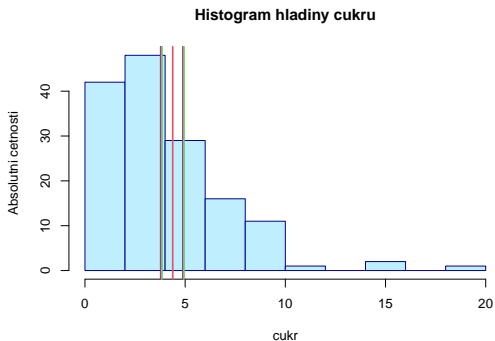
Konstrukce

- uvažujme dostupný náhodný výběr jako "základnu" dat
- realizujme B bootstrapových výběrů velikosti n na této základně - výběr s opakováním (každá naměřená hodnota má být vybrána $1/n$)
- z každého výběru spočteme 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 spolehlivosti vychází velmi podobně jako ten klasický.

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).



Intervalový odhad pravděpodobnosti

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

Příklad. *Házíme 100 krát kostkou, z těchto 100 hodů mi šestka padla 20 krát a zajímá mne interval spolehlivosti pro pravděpodobnost, že padne 6.*

Nejlepším bodovým odhadem pravděpodobnosti je relativní četnost, tj. $\hat{p} = 20/100 = 1/5$.

Za předpokladu dostatečně velkých hodnot n a p , konkrétně $n\hat{p}(1 - \hat{p}) > 9$, kde \hat{p} je odhad parametru p platí, že

$$p = (\hat{p} - p) / \sqrt{p(1 - p)/n} \sim N(0, 1)$$

tj. 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.

Statistické odhady jsou jen jednou částí statistiky, někdy je potřeba otestovat 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.

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.
- **Alternativní hypotézu** – značíme H_1
– obsahuje více možností (např. interval) – př. nový lék je lepší než ten stávající, výnosy druhů jabloní se liší

Základy testování hypotéz

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

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

Jiný závěr udělat nemohu, to co mě zajímá (to, co chci prokázat), musí být v alternativě. 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

Základy testování hypotéz

	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

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**

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.

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 $\alpha = 0.05$

Testová statistika jednovýběrového t-testu je

$$T = \frac{\bar{X} - \mu_0}{\text{sd}(X)} \sqrt{n}$$

a za platnosti nulové hypotézy má tato statistika t -rozdělení o $n - 1$ stupních volnosti.

Testovou statistiku T porovnáваме s kritickými hodnotami t -rozdělení (tj. kvantily), na základě čehož buď můžeme přímo rozhodnout o zamítnutí nebo nezamítnutí nulové hypotézy, nebo můžeme spočítat p -hodnotu a test vyhodnocovat na základě ní.

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

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 jedenáctiletých 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

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

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

V případě, že data zásadně porušují předpoklad normality, je potřeba použít některý z neparametrických testů, které nejsou založeny na průměru. Nejjednodušším z nich je **znaménkový test**. Testujeme

- $H_0 : \text{medián} = m_0$
- $H_1 : \text{medián} \neq m_0, > m_0, < m_0$

Pro každé pozorování spočteme rozdíl $X_i - m_0$ a spočítáme, kolik těchto rozdílů je kladných. Tento součet označme jako Z . Za platnosti nulové hypotézy má testová statistika Z binomické rozdělení $Bi(n, 1/2)$, kde n je počet pozorování. Pro velká n je možné použít i transformaci

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

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

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

Spočtíme 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 $> \alpha (= 0.05)$ a H_0 tedy **nezamítáme**.

Použitím U -transformace dostaneme $U = -0.632$ a p -hodnotu 0.527. **Střední hodnota věku otců může být 33.**

Wilcoxonův jednovýběrový test

Znaménkový test porovnává pouze počet hodnot ležících pod mediánem a těch, co leží nad ním. Nezohledňuje však vzdálenost od mediánu. To dělá Wilcoxonův test, neboli **Mann-Whitneyův** test. Ten už je založen na pořadích. Testované hypotézy zůstávají stejné.

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 tento součet S^+ a obdobně označme S^- součet pořadí pro záporné rozdíly, musí platit $S^+ + S^- = n(n+1)/2$.

Pro větší n lze užít transformaci

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

kteřá má za platnosti H_0 $N(0, 1)$ 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 $0.51 > \alpha (= 0.05)$ a H_0 tedy **nezamítáme**. Střední hodnota věku otců může být 33.

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.

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.

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).

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

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 $p = 7.26 \cdot 10^{-9}$, 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

V případě, že chceme porovnat dva závislé výběry, které nesplňují předpoklad normality, používá se párový Wilcoxonův test.

Testované hypotézy zůstávají stejné jako u párového t-testu a i postup je obdobný.

V prvním kroku se spočítají rozdíly v rámci párů, tj. pro každé $X_i, Y_i, i = 1, \dots, n$

$$R_i = X_i - Y_i$$

Pokud tyto rozdíly nemají normální rozdělení, použije se pro ně jednovýběrový Wilcoxonův test.

Dvouvýběrový t-test

Porovnáváme-li střední hodnotu dvou **nezávislých** výběrů, používá se **dvouvýběrový test**.

Budeme zde brát dva typy dvouvýběrového t-testu:

- Dvouvýběrový t-test pro shodné rozptyly
- Welchův dvouvýběrový test pro různé rozptyly

K tomu, abychom mohli vybrat správnou verzi testu, je potřeba otestovat shodu rozptylů v obou výběrech. Testuje se

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

Testová statistika má za platnosti H_0 F -rozdělení.

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

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika tohoto testu má tvar

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

kde

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

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.

Testová statistika tohoto testu má tvar

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

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

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

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

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?

Nejprve otestujeme shodu rozptylů, testová statistika vychází

$$F = \frac{\text{Var}(X)}{\text{Var}(Y)} = \frac{45.1}{50.6} = 0.89$$

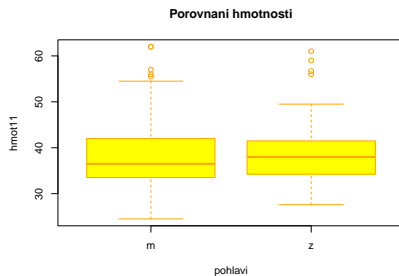
P-hodnota testu vyšla 0,56, což je více než $\alpha = 0.05$. **Nulovou hypotézu tudíž nezamítáme**, rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly.

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 testu vychází

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

Tuto testovou statistiku porovnáváme s kvantilem t-rozdělení $t_{220}(1 - 0.025) = 1.97$ (kvantil pro oboustrannou alternativu). Jelikož testová statistika je v absolutní hodnotě menší než tento kvantil, tak **nulovou hypotézu nezamítám**.

P-hodnota testu vyšla 0.3151, tedy číslo větší než $\alpha = 0.05$

Závěr: Na hladině významnosti 5% jsem neprokázala, že by se hmotnost jedenáctiletých hochů a dívek lišila.

Wilcoxonův dvouvýběrový test

Pro porovnání dvou nezávislých výběrů, které nesplňují předpoklad normality, se používá Wilcoxonův dvouvýběrový test. Testujeme

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

Test je založen na pořadích hodnot sdruženého výběru. Postup

- oba výběry 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.

Porovnááme-li střední hodnotu ve více než ve dvou nezávislých výběrech, používá se **analýza rozptylu**. Testujeme

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

Myšlenka spočívá v porovnání variability **mezi výběry** s variabilitou **v rámci výběrů**.

Klasická (níže uvedená) ANOVA je určena pro normálně rozdělená data a výběry se shodnými rozptyly. Existuje i Welchova obdoba pro různé rozptyly ve skupinách a neparametrická verze pro data, která nemají normální rozdělení.

Analýza rozptylu – ANOVA

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

Analýza rozptylu rozkládá celkovou variabilitu

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

na variabilitu vysvětlenou výběry (mezi výběry) SS_A a variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SS_e . Platí

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

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

	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	p -hodnota
Faktor A	SSA	$df_A = k - 1$	$MSA = \frac{SSA}{df_A}$	$F = MSA / MSe$	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

Předpokladem analýzy rozptylu je shoda rozptylů ve všech výběrech. Tento předpoklad můžeme zkontrolovat např. prostřednictvím **Bartlettova testu**.

Testujeme

- H_0 rozptyly jsou shodné
- H_1 rozptyly se liší

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

$$S^2 = \frac{\sum_{i=1}^k (n_i - 1) \text{Var}(X)_i}{n - k},$$
$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right)$$

Testová statistika

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

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

Zajímá-li nás, které konkrétní dvojice výběrů se od sebe významně liší, nelze toto zjistit větším počtem běžných dvouvýběrových testů, neboť by tím příliš vzrostla chyba prvního druhu (tj. neudržela by se celková hladina významnosti). Je nutné použít párové srovnání, např. **Tukeyův test**, případně **Tukey HSD test** pro různě velké výběry.

Testuje se

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

pro všechny dvojice i a j .

Testová statistika má tvar

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

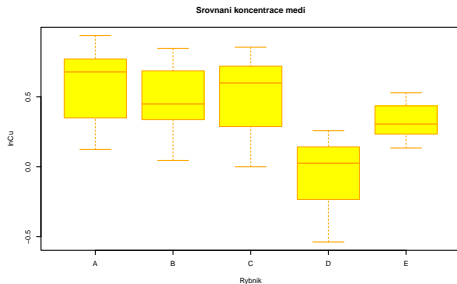
Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty.

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

Testujeme

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

Grafické porovnání



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

Testujeme

- 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.

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			

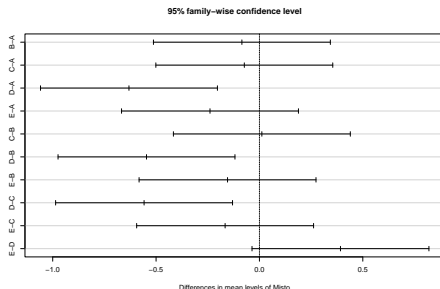
P-hodnota vyšla menší než $\alpha = 0.05$, což znamená, že **nulovou hypotézu zamítáme** a rybníky se mezi sebou významně liší.

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

V případě, že není splněn předpoklad normality při porovnání více než dvou nezávislých výběrů, používá se

Kruskal-Wallisova ANOVA. Kruskal-Wallisova ANOVA je přímým zobecněním Wilcoxonova dvouvýběrového testu.

Testujeme

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

Stejně jako u dvouvýběrového Wilcoxonova testu srovnáme všechny naměřené hodnoty do řady, určíme jejich pořadí a spočteme statistiky T_1, \dots, T_k , kde k je počet výběrů. Pak platí, že 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

V případě, že Kruskal-Wallisova ANOVA určí, že se výběry mezi sebou významně liší, je potřeba zjistit, které konkrétní dvojice výběrů se liší. K tomu může sloužit např. **Dunnův test**.

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

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

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.

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 porovnaná 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

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

Test je pak založen na porovnání SSA a SSz .

V případě, že porovnáváme závislé výběry, které nemají normální rozdělení, používá se **Friedmanův test**. Myšlenkou testu je, že každý jedinec přiřadí jednotlivým vzorkům pořadí od 1 do k (pro hodnoty naměřené v čase se určí pořadí v rámci každého jedince) a tato pořadí se pak sečtou a zprůměrují. Označme tyto průměry $\bar{r}_{.j}$. Ty jsou pak základem testové statistiky

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

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

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 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 kategoričnými uspořádanými (ordinálními) veličinami

Korelační koeficient měří lineární závislost hodnotou z intervalu $\langle -1, 1 \rangle$:

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

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

Pearsonův korelační koeficient vypočteme jako

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

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

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

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

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

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.

Dále jsme testovali

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

Pearsonův korelační koeficient vyšel 0,5 a testová statistika

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

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.

Jinak funguje **Kendallův korelační koeficient** (Kendalovo 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**.

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

Konkrétně je **Kendalovo** 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.

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 , $n_2 = \sum_i u_i(u_i-1)/2$ 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.

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

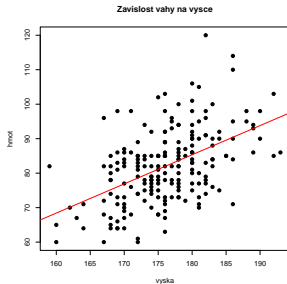
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

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

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.

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.

Koeficient determinace

Zajímavý ukazatel je koeficient determinace

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

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 $b_1/\text{se}(b_1) \sim N(0, 1)$, kde b_1 je odhad lineárního členu β_1 a $\text{se}(b_1)$ je jeho střední chyba.

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.

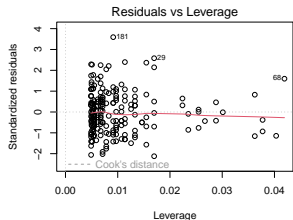
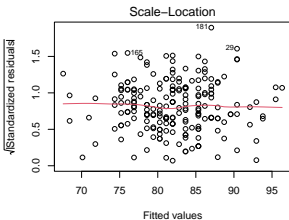
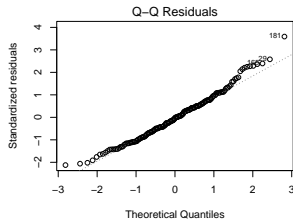
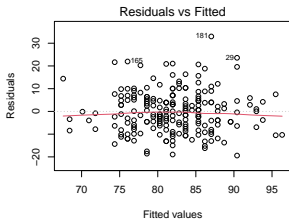
Závěr: Můžeme tedy říci, že u mužů s jedním rizikovým faktorem ischemické choroby srdeční hmotnost na výšce závisí. Závislost je přímá a vysvětlí se jí 25% variability závisle proměnné (hmotnosti).

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.

Ukázka grafických testů předpokladů



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)

Test dobré shody

Uvažujme jednu kategorickou proměnnou, která může nabývat více než dvou kategorií. Takováto proměnná má tzv. **Multinomické rozdělení**. Jedná se o zobecnění binomického rozdělení.

Máme k kategorií, kterých může náhodná veličina nabývat. Opakujeme n -krát pokus a počítáme, kolikrát byla každá z kategorií pozorována. Tyto počty označme jako proměnné X_1, \dots, X_k . Pravděpodobnost, že nastane i -tá kategorie označme p_i . Multinomické rozdělení je pak dáno pravděpodobnostmi

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

Dále platí, že

$$E(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i)$$

Test dobré shody

Chceme-li otestovat, že výše uvedené pravděpodobnosti nabývají nějakých konkrétních hodnot, tj. testujeme

- $H_0 : p_1 = \pi_1, \dots, p_k = \pi_k$
- $H_1 : \text{neplatí } p_1 = \pi_1, \dots, p_k = \pi_k$

použijeme χ^2 **test dobré shody**.

Testová statistika má tvar

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

a za platnosti nulové hypotézy má χ^2 -rozdělení o $k - 1$ stupních volnosti. Předpokladem testu je, že všechny očekávané četnosti, tj. všechny hodnoty $n\pi_i$, jsou větší než 5. Tímto testem můžeme i testovat, zda náhodná veličina má nějaké konkrétní rozdělení.

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

- $H_0 : p_1 = p_2 = \dots = p_6 = 1/6$
- $H_1 : \text{alespoň jedna z pravděpodobností } p_1, \dots, p_6 \text{ se nerovná } 1/6.$

Příklad. Naměřili jsme hodnoty

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

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

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**.

Neprokázali jsme, že by kostka byla falešná.

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

- X_1, \dots, X_k hodnoty jedné kategorické proměnné
- Y_1, \dots, 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

Kontingenční tabulka absolutních četností pak 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

Test nezávislosti je založen na porovnání pozorovaných četností v tabulce a četností očekávaných za platnosti nulové hypotézy. Testujeme

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

Testová statistika má tvar

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(\text{pozorovane}_{i,j} - \text{ocekavane}_{i,j})^2}{\text{ocekavane}_{i,j}} = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{i,j} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

Tato testová statistika má za platnosti nulové hypotézy χ^2 -rozdělení o $(k - 1)(l - 1)$ stupních volnosti. Očekávané četnosti se dopočítávají z definice nezávislosti $P(A \cap B) = P(A)P(B)$.

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

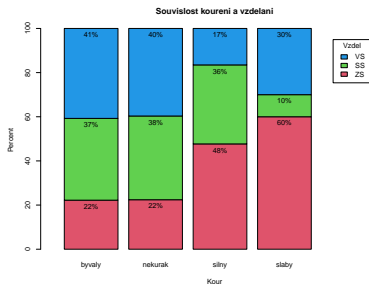
$$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

Pro větší tabulky je test složitější.

Příklad. U 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční bylo zjišťováno i 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

Vztah dvou kategoričkových proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

Testem nezávislosti jsme zjišťovali

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

Testová statistika vyšla 21.286. Porovnááme ji s kvantilem χ^2 -rozdělení $\chi_6^2 = 12.59$. 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.

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.

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.

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.*