PAS

2024 / 2025

Obsah

1. Pravděpodobnost

1.1 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 kostce padne sudé číslo, výška člověka bude větší než 170 cm, student zkoušku udělá…

Elementární jev

– nejmenší možné náhodné jevy, 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; značíme se **Ω**

1.2 Definice pravděpodobnosti

Klasická definice

- Mějme prostor elementárních náhodných jevů **Ω**; systém množin ***A*** – algebru[1] na tomto prostoru; pravděpodobností pak nazveme reálnou funkci ***P(A)*** definovanou na algebře ***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 ***(Ω, A, P)*** se nazývá klasický pravděpodobnostní prostor

- Př.*: 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 ***A*** jsou pak všechny možné výsledky pokusu, např.:
  v prvním hodu padne orel, celkem padne orel právě 5x, orel padne méně než 3x atd…

  – funkce P pak přiřadí každému z těchto výsledků hodnotu mezi 0 a 1 – pravděpodobnost

1 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.

Kolmogorova definice

- Nechť $\boldsymbol{\Omega}$ je neprázdná množina, nechť $\boldsymbol{A}$ je σ-algebra náhodných jevů definovaných na $\boldsymbol{\Omega}$; pravděpodobností se nazývá reálná funkce $\boldsymbol{P(A)}$ definovaná na $\boldsymbol{A}$, která pro $A \in A$, $A_1, A_2, \ldots \in A$, $A_i \cap A_j = \emptyset$ pro všechna $i \neq j$ splňuje:

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

- rozšíření klasické definice na spočetné, potažmo nespočetné (reálné) množiny $\boldsymbol{\Omega}$

- Př.*: 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 ½, že orel nepadne, takže skutečně nemusí padnout)
  – jako množinu $\boldsymbol{\Omega}$ pak uvažujeme všechny přirozená čísla $\boldsymbol{\Omega} = N \cup \infty$

- Př. 2: *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 $\boldsymbol{\Omega}$ teď tvoří všechna reálná čísla na intervalu od 0 do délky okruhu

1.3 Výpočet pravděpodobnosti

– pravděpodobnost množiny A je dána vztahem: $P(A) = \frac{|A|}{|\Omega|}$

- $|\boldsymbol{A}|$ velikost množiny $\boldsymbol{A}$ = počet možností příznivých jevů $\boldsymbol{A}$; délka úsečky $\boldsymbol{A}$
- $|\boldsymbol{\Omega}|$ velikost celého prostoru elementárních jevů $\boldsymbol{\Omega}$ = počet všech možností; velikost celé uvažované úsečky

- Př.*: 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
  – zajímá nás 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

1.3.1 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 ∅

– jev, který neobsahuje ani jeden elementární jev
– Př.: na kostce padne mínus jedna

Jev opačný k jevu A

– tj. 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

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

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

poslední rovnost lze zobecnit na n = ∞

1.3.2 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

$$P(A) = P(A|B) \qquad \text{neboli} \qquad P(A)P(B) = P(A \cap B)$$

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

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)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \qquad \text{neboli} \qquad P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_{j=1}^{k} P(A|H_j)P(H_j)}$$

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

1.4 Senzitivita a specificita testu

– charakteristiky popisující kvalitu nejčastěji u medicínských testů

- **senzitivita testu**
  – pravděpodobnost, že test vyjde pozitivně, pokud je osoba nemocná
  – ***P***(test je pozitivní | osoba je nemocná)

- **specificita testu**
  – pravděpodobnost, že test vyjde negativně, pokud je osoba zdravá
  – ***P***(test je negativní | osoba je zdravá)

– Př.: *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(test je pozitivní | osoba je nemocná) = 45 / 50 = 0.9

- specificita testu
  – P(test je negativní | osoba je zdravá) = 1750 / 1950 = 0.897

- Jsem nemocný, když mám pozitivní test?
  – P(osoba je nemocná | test je pozitivní) = 45 / 245 = 0.184
  – pomocí Bayesovy věty:

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

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

1.5 Vybrané klasické pravděpodobnostní modely

Uspořádaný výběr s vracením

- máme urnu, ve které je $\boldsymbol{M}$ 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 $\boldsymbol{\Omega}$ má velikost $\boldsymbol{M^m}$

- 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 $\boldsymbol{M}$ rozlišitelných koulí a kouli po každém tahu vrátím;
  když mi nezáleží na pořadí, pak prostor elementárních jevů $\boldsymbol{\Omega}$ 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 $\boldsymbol{M}$ rozlišitelných koulí, tahám náhodně $\boldsymbol{m}$ koulí;
  koule tam nevracím, ale zaznamenávám pořadí, v jakém byly taženy;
  pak prostor elementárních jevů $\boldsymbol{\Omega}$ 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 $\boldsymbol{M}$ rozlišitelných koulí;
  koule nevracím a nezáleží mi na pořadí, v jakém je táhnu;
  prostor elementárních jevů $\boldsymbol{\Omega}$ má pak velikost: $\binom{M}{m}$

- Př.: *Kolika způsoby mohu vybrat 5 žáků ke zkoušení z dvaceti přítomných ve třídě?*

Náhodná procházka

- uvažujme částici, která se pohybuje po celočíselné přímce $\boldsymbol{Z}$ a označme jako S_k její polohu v čase k = 0, 1, 2, ...
- předpokládejme, že na začátku je částice v bodě 0, tj. $S_0 = 0$
  a je-li v nějakém časové okamžiku $\boldsymbol{k}$ v bodě $\boldsymbol{a}$, tak v čase $\boldsymbol{k + 1}$ je v bodě $\boldsymbol{a - 1}$ s pravděpodobností ½ a v bodě $\boldsymbol{a + 1}$ také s pravděpodobností ½

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

- rozhodování o pohybu částice v každém bodě je náhodné a nezávislé na předchozích krocích
- určeme rozdělení pravděpodobností, kde se bude částice nacházet v čase $\boldsymbol{n}$

- prostor elementárních jevů je $\boldsymbol{\Omega} = \{0, 1\}^n$ a jeho velikost je $|\Omega| = 2^n$
- stav v čase $\boldsymbol{n}$ je pak jednoznačně určen počtem pohybů doprava
- označme tento počet kroků doprava jako $\boldsymbol{k}$ (nutně $\boldsymbol{n - k}$ kroků musí být doleva)
  a skončím v bodě $\boldsymbol{S_n = k - (n - k)}$
- pro pravděpodobnost stavu $\boldsymbol{S_n}$ 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 $\boldsymbol{a \in Z}$
- budeme-li zkoumat její návrat do bodu 0, pak k němu dojde s pravděpodobností 1, ale střední doba čekání na tento okamžik bude nekonečná

2. Pravděpodobnostní rozdělení

2.1 Základní charakteristiky

2.1.1 dělení pravděpodobnostního rozdělení

– pravděpodobnostní rozdělení dělíme dle 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, …

2.1.2 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

2.1.3 Střední hodnota a rozptyl

– další charakteristiky pro diskrétní i spojitá rozdělení je střední hodnota a rozptyl

Střední hodnota

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

– vlastnosti:

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

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:

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

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

2.2 Vybraná diskrétní rozdělení

Binomické rozdělení

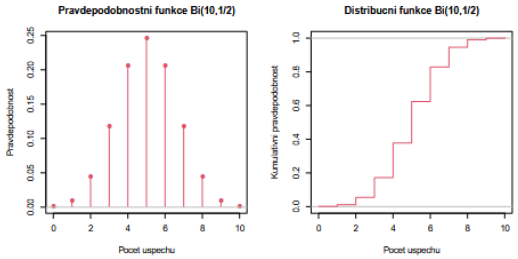
– *mějme náhodný pokus, který může skončit jedním ze dvou výsledků: úspěch / neúspěch;*
opakujme tento pokus mnohokrát a počítejme počet úspěchů;
počet úspěchů má binomické rozdělení

– 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: $E(X) = np, \qquad Var(X) = np(p-1)$

– Př.: *Házíme 10x mincí a počítáme, kolikrát padla panna.*

- počet pokusů je ***n*** = 10, pravděpodobnost úspěchu ***p*** = ½
- máme tedy rozdělení ***Bi(10, ½)***
- pravděpodobnostní a distribuční funkce:



- střední hodnota a rozptyl:

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

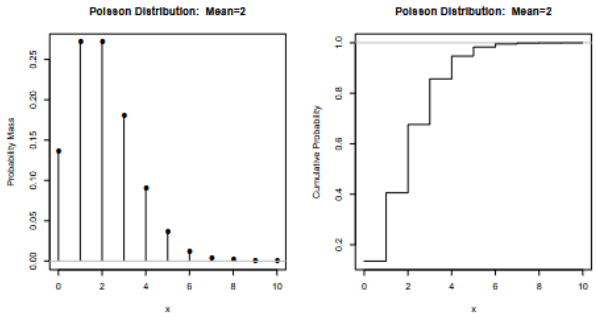
Poissonovo rozdělení

– *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(λ)***, kde ***λ*** – parametr rozdělení, intenzita

– hodnoty pravděpodobnostní funkce pro k = 0, 1, 2, . . . $p(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
– střední hodnota a rozptyl: $E(X) = \lambda, \qquad Var(X) = \lambda$

– pravděpodobnostní a distribuční funkce Poissonova rozdělení s parametrem λ = 2:



– předpokládejme binomická rozdělení ***Bi(n, pn)***, kde $np_n \to \lambda$,
pak tato binomická rozdělení konvergují k rozdělení Poissonovu s parametrem ***λ***

Hypergeometrické rozdělení

– 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

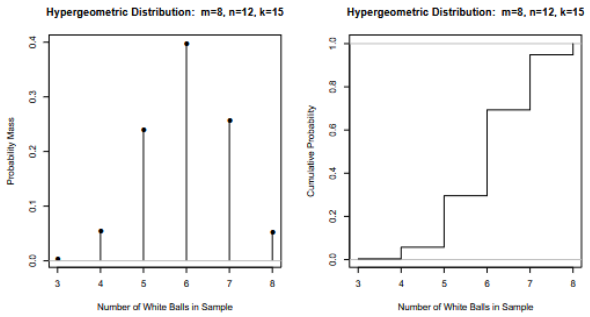
– značení ***Hy(N, A, n)***, kde ***N*** – počet koulí v urně;
A – počet označených koulí v urně;
n – počet tažených koulí

– hodnoty pravděpodobnostní funkce: $p(X = k) = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}$

– střední hodnota a rozptyl: $E(X) = \frac{nA}{N}, \qquad Var(X) = n\frac{A}{N}\left(1 - \frac{A}{N}\right)\left(\frac{N-n}{N-1}\right)$

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

- pravděpodobnostní a distribuční funkce:



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

2.3 Vybraná spojitá rozdělení

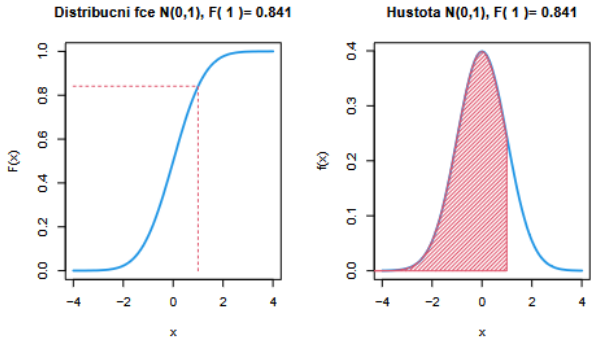
Normální rozdělení

– 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(μ, σ2)***, kde **μ** – střední hodnota; **σ^2** – rozptyl

– hustota normálního rozdělení má tvar (Gaussova křivka): $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

– 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:



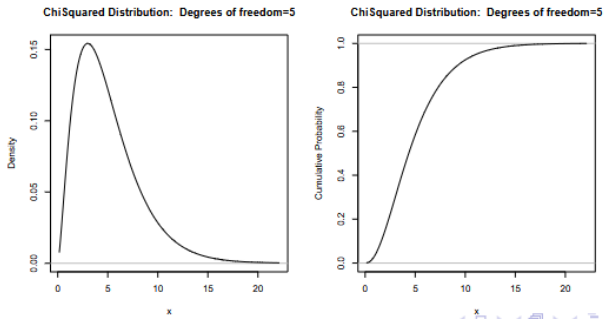
– předpokládejme binomické rozdělení ***Bi(n, p)***, kde $0.1 \leq \boldsymbol{p} \leq 0.9$,
pak pro $\boldsymbol{n} \to \infty$ toto rozdělení konverguje k normálnímu s parametry ***np***, ***np(1 − p)***

x^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 x^2-rozdělení o ***n*** stupních volnosti

– dále je to rozdělení některých testových statistik, zejména těch, týkajících se rozptylu

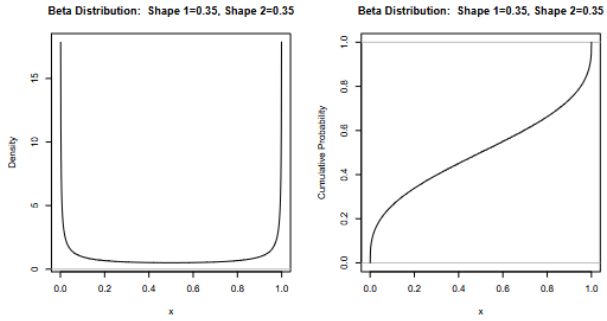
– hustota a distribuční funkce x^2-rozdělení o 5 stupních volnosti:



Beta rozdělení

– rozdělení pravděpodobností nějakého jevu; např.: sledujeme pravděpodobnost, že vybraný 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):



3. Statistika

3.1 Definice a základní pojmy

– 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

– Př.: *Zajímá nás průměrná výška dospělých lidí v celé ČR. 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.*

Náhodná veličina

– jakákoliv veličina, kterou měříme *(v př. výška)*

Populace

– soubor, pro nějž chceme udělat nějaký závěr, *(v př. dospělí obyvatelé ČR)*

Náhodný výběr

– v porovnání s populací malý soubor pozorování, jde o nezávislé, stejně rozdělené náhodné veličiny, *(v př. výběr 200 lidí)*

Populační charakteristika

– charakteristika popisující populaci, *(v př. populační průměr)*

Výběrová charakteristika

– charakteristika spočítaná na výběru, pomocí níž odhadujeme populační ekvivalent, *(v př. výběrový průměr)*

3.2 Typy proměnných

- **číselné (kvantitativní, numerické) proměnné**
  – př.: výška, váha, věk, atd…
  - **spojité**
  - **diskrétní**

- **kategorické (kvalitativní) proměnné**
  – př. barva, kraj, povolání, nebo také známka ve škole; číslo, které padne na kostce, atd…
  - **nominální** – neuspořádané, př. barva, kraj
    – déle dělíme na polytomické, dichotomické…
  - **ordinální** – uspořádané, př. známka, číslo na kostce

3.2.1 popis jednotlivých typů 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.
- pro malé počty kategorií i absolutní a relativní četnost

3.2.2 problémy v datech

– chybějící pozorování

- snažíme se, aby jich bylo co nejmén ˇe,
- když jich je málo, tak pracujeme bez nich – v ˇetšina statistických
- metod implementovaných v r ˚uzných softwarech si s tím poradí
- je možné je doplnit na základ ˇe n ˇejakého modelu (imputation

– odlehlé hodnoty

- kontrola, zda nedošlo k chyb ˇe m ˇeˇrení
- pokud ne, tak z popisných statistik se v ˇetšinou nevynechávají,
- ale je dobré zmínit, že se jedná o odlehlé hodnoty
- pro popis prom ˇenné je pak lépe zvolit ukazatele necitlivé na
- odlehlé pozorování
- ze složit ˇejších analýz se ˇcasto vynechávají

4. Popisné statistiky

4.1 Popisné statistiky číselné proměnné

Popisné statistiky polohy

– Př.: *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.*

- **průměr** z ***n*** hodnot značených X_1, X_2, X_3, …, X_n: $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- **medián** z uspořádané řady
  = hodnota prostřední podle velikosti, nebo průměr prostředních dvou

- **kvartily** z uspořádané řady
  – hodnoty v jedné (¼) a ve třech čtvrtinách (¾)

  – 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 chceme *p*-tým percentilem oddělit
      - $\boldsymbol{X_{(k)}}$ – hodnoty z uspořádané řady, *k*-tý nejmenší prvek
      - ***q*** – koeficient, kterým se násobí uspořádané hodnoty do váženého průměru

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

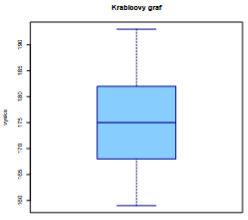
- **extrémy** – minimum a maximum

Grafické popisné statistiky

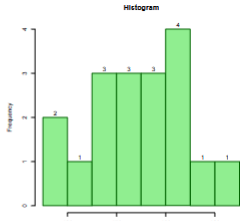
– používáme dva typy grafů:

- **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 $\boldsymbol{1.5(Q_3 - Q_1)}$

- **histogram**
  – počet sloupců je určen vybraným pravidlem, nejčastěji se používá ***Sturgesovo pravidlo***
  $\boldsymbol{k = 1 + 3.32\ log_{10}(n)}$ kde ***n*** je počet pozorování



1. Krabicový graf



2. Histogram

výsledky pro výšky dospělých mužů:

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

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**

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

- **variační koeficient**

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

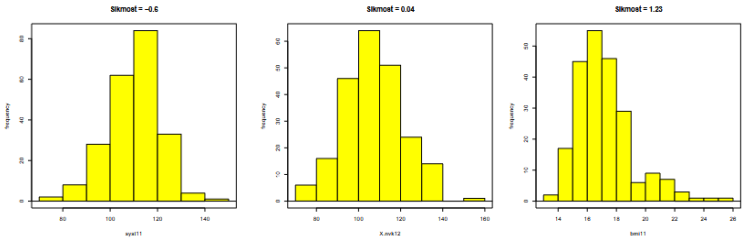
Popisné statistiky tvaru rozdělení

– počítají se ze standardizovaných proměnných, tak zvaných **Z-skórů** $Y_i = \frac{X_i - \overline{X}}{\mathrm{sd}(X)}$

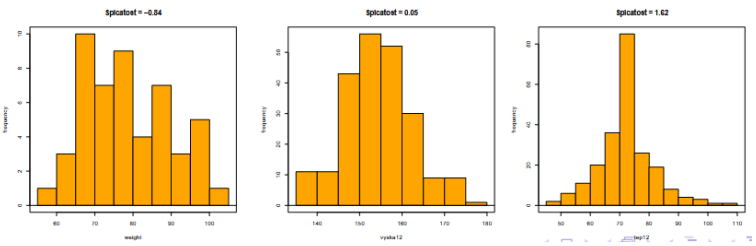
- **šikmost**
  – průměr ze třetích mocnin z-skórů $\mathrm{Skew}(X) = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\mathrm{sd}(X)}\right)^3 = \frac{\sum_{i=1}^{n} Z_i^3}{n}$

- **špičatost**
  – průměr ze čtvrtých mocnin z-skórů $\mathrm{Kurt}(X) = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\mathrm{sd}(X)}\right)^4 - 3 = \frac{\sum_{i=1}^{n} Z_i^4}{n} - 3$

záporná, nulová a kladná šikmost



záporná, nulová (špičatost normálního rozdělení a kladná špičatost



4.2 Popisné statistiky kategorické proměnné

Číselné popisné statistiky nominální proměnné

– Př.: *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.*

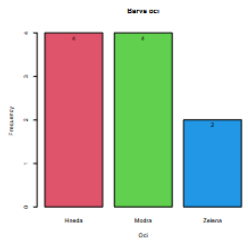
– tabulka absolutních a relativních četností:

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

Grafické popisné statistiky nominální proměnné

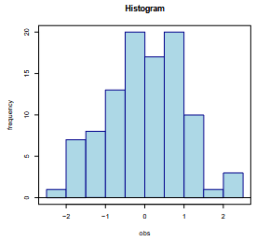
– sloupcový a koláčový graf – v absolutních počtech nebo procentech

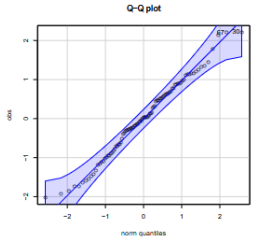


4.3 Testy normality

– většina statistických postupů odvozena pro normální rozdělení → třeba zjistit, zda ho veličina má či ne

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





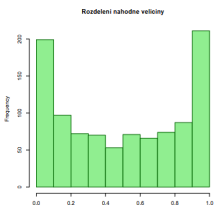
- **číselné testy**
  – např.: Shapiro-Wilkův, Andersonův-Darlingův, Kolmogorovův-Smirnovův, Lillieforsův a další

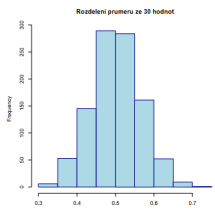
Centrální limitní věta

– definice: *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 sčítáme / průměrujeme, tím spíše bude mít průměr normální rozdělení

– rozdělení průměru 30-ti hodnot z beta rozdělení:





Zákon velkých čísel

– definice: *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: $E\overline{X} = EX$

5. Odhady populačních charakteristik

5.1 Odhad střední hodnoty

Bodový odhad střední hodnoty

– Př.: *Mějme situaci, kdy potřebujeme odhadnout průměrnou výšku dospělých lidí v celé ČR. 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?

- ***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

– tento interval závisí na:
- ***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, 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

– interval spolehlivosti
- základem 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 pozorován

- 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 − α/2)*** je kvantil teoretického rozdělení
  - znám-li skutečný rozptyl ***σ2*** → používá se kvantil standardního normálního rozdělení N(0, 1)
  - musím-li odhadnout ***σ2*** pomocí výběrového rozptylu → používá se kvantil *t*-rozdělení o ***n − 1*** stupních volnosti

– výpočet 95% intervalu spolehlivosti pro uvedený příklad:
- teoretický rozptyl není znám; známe: $\overline{X} = 173.12, \mathrm{sd(X)} = 8.9, \mathrm{n} = 500, \alpha = 0.05.$

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

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

$$172.34, 173.9$$

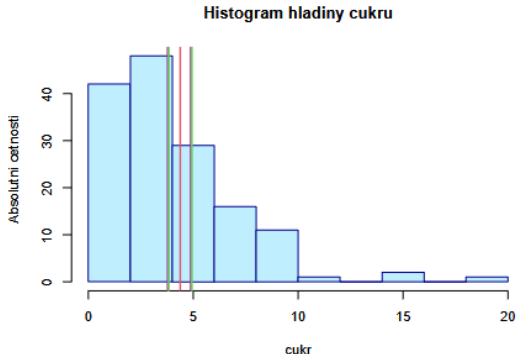
Se spolehlivostí 95% bude skutečný populační průměr výšky mužů ležet v intervalu od 172.34 cm do 173.9 cm

Bootstrapový interval spolehlivosti pro střední hodnotu

- 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 spočteme výběrový průměr
- meze bootstrapového intervalu spolehlivosti jsou ***α/2*** a ***1 − α/2***-tý kvantil z vektoru průměrů

– počet bootstrapových výběrů má být minimálně B = 1000, lépe B = 10000

Př.: *Uvažujme měření hladiny cukru v krvi. Bylo změřeno 150 mužů s průměnou 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).



5.2 Odhad pravděpodobnosti

– předpokládejme binomické rozdělení s parametrem ***p***, který chci odhadnout z dat

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

- nejlepším bodovým odhadem psti je relativní četnosti: $\hat{p} = 20/100 = 1/5$
- 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

5.3 Odhad rozptylu

– jako bodový odhad populačního rozptylu používáme výběrový rozptyl: $\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

– 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

– $\mathbf{x}^2$ není symetrické

– intervalový odhad pro rozptyl je: $\left(\frac{(n-1)s^2}{\chi^2_n(1-\alpha/2)}, \frac{(n-1)s^2}{\chi^2_n(\alpha/2)} \right)$

6. Testování hypotéz

6.1 Základní pojmy

– 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 (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 (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

Výsledky 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* – **hladina významnosti (α)**
  – zamítneme H_0, přestože platí
  – závažnější z obou chyb

- *chyba druhého druhu* – **(β)**
  – **ne**zamítneme H_0, přestože neplatí
  – hodnota $1 - \beta$ se nazývá **síla testu**
  – za dané hladiny významnosti chceme test co nejsilnější

	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

– dle toho, co testujeme, a dle typu dat vybereme vhodný statistický test, kterým budeme o platnosti testovaných hypotéz rozhodovat
– rozhodnutí můžeme udělat bud' 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

6.2 Jednovýběrový test

Jednovýběrový t-test

– nejjednodušším testem je jednovýběrový test o střední hodnotě
– testujeme $\boldsymbol{H_0}$: střední hodnota = μ_0 proti jedné ze tří alternativ:
- $\boldsymbol{H_1}$: střední hodnota ≠ μ_0
- $\boldsymbol{H_1}$: střední hodnota < μ_0
- $\boldsymbol{H_1}$: střední hodnota > μ_0

– není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.0$

– ***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 $\boldsymbol{n-1}$ stupních volnosti
- testovou statistiku $\boldsymbol{T}$ 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

– Př.: *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 ČR je menší než 150 cm?*

- testované hypotézy:
  H_0 : průměrná výška = 150 cm
  H_1 : průměrná výška < 150 cm

- testujeme na hladině významnosti $\alpha = 0.05$

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

- 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 $\boldsymbol{p} = 0.005 < 0.05$, 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:
- $\mathbf{H_0}$: medián = m_0
- $\mathbf{H_1}$: medián ≠ m_0 *nebo* > m_0 *nebo* < m_0

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

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

  za platnosti $\boldsymbol{H_0}$ má ***U*** normální rozdělení ***N(0, 1)***

– Př.: *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

- spočteme rozdíly $\boldsymbol{X_i - m_0}$: -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 $\boldsymbol{\alpha}$(= 0.05) a $\boldsymbol{H_0}$ tedy **ne**zamítáme
- použitím U-transformace dostaneme ***U*** = −0.632 a p-hodnotu 0.527
- závěr: Střední hodnota věku otců může být 33.

Wilcoxonův jednovýběrový test

– neboli Mann-Whitneyův test
– neparametrický test
- používá se, když proměnná nemá normální rozdělení
- založen na pořadích
- silnější než znaménkový test
- testované hypotézy zůstávají stejné (testuje hodnotu mediánu)

– postup:
- spočítají se rozdíly od testované hodnoty $\boldsymbol{X_i - m_0}$
- určí se jejich znaménko
- určí se pořadí absolutních hodnot rozdílů
- spočítá se součet těchto pořadí pařících kladným rozdílům
- označme $\boldsymbol{S^+}$ součet pořadí kladných rozdílů a $\boldsymbol{S^-}$ součet pořadí záporných rozdíl, 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 $\boldsymbol{H_0}$ normální ***N(0, 1)*** rozdělení

– Př.: *Stejný příklad 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

- spočteme rozdíly $\boldsymbol{X_i - m_0}$: -3, -5, **3**, **5**, -5, -7, -4, **4**, -8, **17**
  a jejich absolutním hodnotám přiřadíme pořadí 1.5, 6, **1.5**, **6**, 6, 8, 3.5, **3.5**, 9, **10**
- sečteme kladné (tučné) pořadí $\boldsymbol{S^+} = 21$ a záporné pořadí $\boldsymbol{S^-} = 34$
- testová statistika vychází $\boldsymbol{U} = -0.66$ a p-hodnota 0.51 je větší než $\boldsymbol{\alpha}(= 0.05)$, $\mathbf{H_0}$ tedy **ne**zamítáme
- závěr: Střední hodnota věku otců může být 33.

6.3 Párový test

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

– postup:
- při aplikaci testu je důležité udržet párová data u sebe
- 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

– Př.: *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): $T = \frac{\overline{X} - \mu_0}{\mathrm{sd}(X)}\sqrt{n} = \frac{6.1 - 5}{2.8}\sqrt{222} = 5.9$

- 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

6.4 Dvouvýběrový test

– porovnává střední hodnotu dvou nezávislých výběrů
– testované hypotézy:
- $\boldsymbol{H_0}$: střední hodnota $\boldsymbol{X}$ – střední hodnota $\boldsymbol{Y} = 0$
- $\boldsymbol{H_1}$: střední hodnota $\boldsymbol{X}$ – střední hodnota $\boldsymbol{Y} \neq 0,\ < 0,\ > 0$

– kontrolují se zde 2 předpoklady
- normalitu dat
- shodu rozptylů

– 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 normální 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 $\boldsymbol{H_0}$ má F-rozdělení o $\boldsymbol{n_1 - 1}$ a $\boldsymbol{n_2 - 1}$ 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

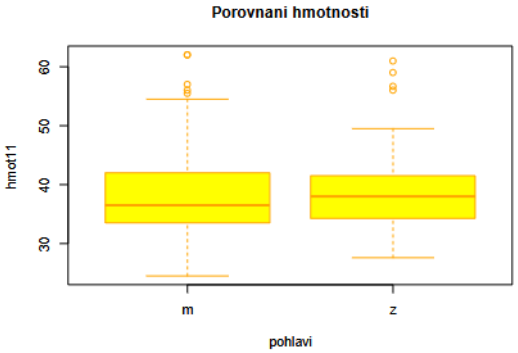
Dvouvýběrový t-test

– Př.: *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 > α = 0.05
  – nulovou hypotézu **ne**zamítáme
  – rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly

- testujeme:
  - H_0 : hmotnost hochů a hmotnost dívek se neliší
    hmotnost hoch ˚u – hmotnost dívek = 0
  - H_1 : hmotnost hoch ˚u a dívek se liší
    hmotnost hoch ˚u – hmotnost dívek ≠ 0

- grafické porovnání:



- 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 **ne**zamítám
- p-hodnota = 0.3151 > α = 0.05
- závěr: Na hladině významnosti 5% jsem neprokázala, že by se hmotnost jedenáctiletých hoch ˚u a dívek lišila.

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 ***v*** 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 ***v*** není celé číslo

Wilcoxonův dvouvýběrový test

– používá se pro porovnání dvou nezávislých výběrů, které nesplňují předpoklad normality
– test 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

– 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

– Př.: *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

- srovnáme všechny hodnoty do řady: **41**, 45, 53, 56, **62**, **66**, **76**, 77, **79**, 79, **80**, 81, **84**
- následně jim přiřadíme pořadí: **1**, 2, 3, 4, **5**, **6**, 7, 8, 9.5, **9.5**, 11, 12, **13**
- poté vypočteme $T_1 = 38.5$; $\mathbf{T_2 = 52.5}$
  $U_1 = 24.5$; $\mathbf{U_2 = 17.5}$
  $U_0 = 0.5$
  $p = 0.6678$

- p-hodnota > α a tedy **ne**zamítáme nulovou hypotézu
- neprokázal se rozdíl mezi studenty v Ústí a v Liberci

6.5 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

– máme 2 předpoklady: normalitu a shodu rozptylů

– 3 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

6.5.1 Klasická ANOVA

– pro shodné rozptyly porovnává variabilitu mezi výběry s variabilitou v rámci výběrů
– označení:

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

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

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

– výstupem ***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 $\boldsymbol{k-1}$ a $\boldsymbol{n-k}$ 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, pak :

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

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

– ***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 x^2*-rozdělení* o $\boldsymbol{k-1}$ stupních volnosti

Párové srovnání

– vzájemné srovnání všech dvojic výběrů se 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

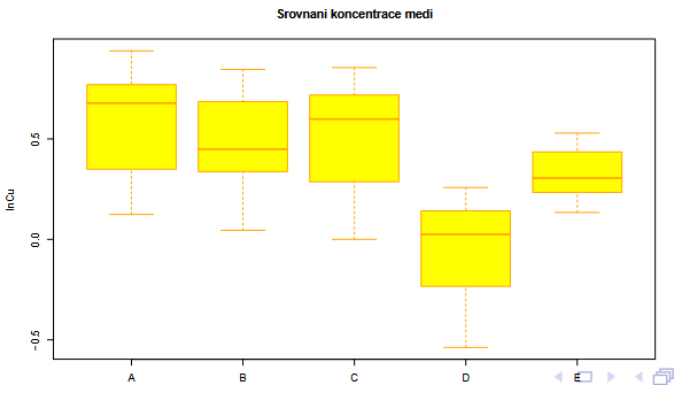
– testové statistiky mají tvar: $Q = \frac{|\overline{X}_{i.} - \overline{X}_{j.}|}{s^*}$, kde $s^* = \sqrt{\frac{SSe}{n(n-k)}}$

– rozdělení těchto statistik nazýváme ***studentizované rozpětí*** a má své vlastní tabelované kritické hodnoty

Př. – ANOVA: *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:



- pro výběr správné verze analýzy rozptylu otestujme nejprve shodu rozptylů ve všech výběrech
- tyto rozptyly vyšly postupně: 0.10, 0.08, 0.10, 0.08, 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
- p-hodnota větší než $\alpha = 0.05$ → nulovou hypotézu **ne**zamítáme a můžeme použít klasickou ANOVU pro shodné rozptyly

- tabulka analýzy rozptylu:

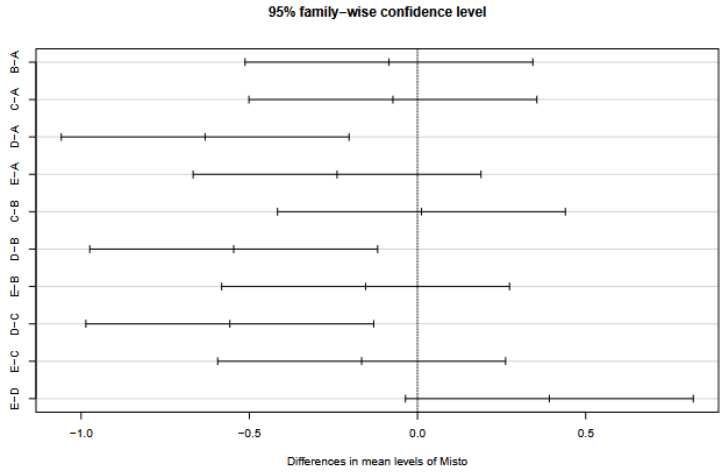
	Součty čtverců	Stupně volnosti	Průměrné čtverce	Testová statistika	*p*-hodnota
Rybník	1.796	4	0.4491	5.896	0.00127
Chyba	2.285	30	0.0762		
Celkem	4.081	34			

- p-hodnota vyšla menší než $\alpha = 0.05$ → nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší

- párové srovnání vrátí 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

- graf pro párové srovnání
  pro dvojici rybníků, kde interval spolehlivosti neobsahuje svislou čárkovanou čáru (nulu), pak v 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.

6.5.2 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 x^2-rozdělení

6.5.3 Dunnův test

– párové srovnání pro data, která nemají normální rozdělení
– testová statistika porovnávající *i-tý* a *j-tý* výběr:

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

– 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

6.6 ANOVA pro opakované měření

– pro porovnání několika závislých vývěrů
– příklady, kdy použít:
- 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.

– testujeme hypotézy:
- H0 : Střední hodnoty výběrů se neliší
- H1 : Střední hodnoty výběrů se liší

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

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

– test je pak založen na porovnání SS_A a SS_z

6.6.1 Fiedmanův test

– neparametrický test porovnávající závislé výběry
– postup:
- stanoví se pořadí hodnot v rámci každého jedince
- pro každý výběr se spočte součet a průuměr pořadí
- označme tyto průměry $\overline{r}_{.i}$

– testová statistika:

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

za platnosti nulové hypotézy má x^2-rozdělení o $\boldsymbol{k-1}$ stupních volnosti

6.7 Test dobré shody

– test o pravděpodobnostním rozdělení kategorické ***proměnné***

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

– multinomické rozdělení je dáno pravděpodobnostmi: $P(X_1 = c_1, \ldots, X_k = c_k) = \frac{n!}{c_1! \cdot \ldots \cdot c_k!} p_1^{c_1} \cdot \ldots \cdot p_k^{c_k}$

– dále platí, že: $E(X_i) = np_i, \qquad Var(X_i) = np_i(1 - p_i)$

– testované hypotézy:

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

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

za platnosti H_0 má x^2-rozdělení o $\boldsymbol{k - 1}$ stupních volnosti

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

– Př.: *Házíme 50 krát šestistěnnou kostkou a počítáme, kolikrát padla která hodnota. Jednička padla 8x, dvojka 5x, trojka 12x, čtyřka 7x, pětka 9x a šestka také 9x. 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

– Př.: Naměřili jsme hodnoty: $c_1 = 8, c_2 = 5, c_3 = 12, c_4 = 7, c_5 = 9, c_6 = 9$
Teoretická hodnota $\boldsymbol{n\pi_i} = 50 \times 1/6 = 8.3333$. 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 x^2-rozdělení o 5-ti stupních volnosti je $\chi^2_5 = 11.07$
  p-hodnota p = 0.6569
- testová statistika je větší než kritická hodnota a p-hodnota menší než α, tedy **ne**zamítáme nulovou hypotézu
- závěr: Neprokázali jsme, že by kostka byla falešná.

6.8 Test nezávislosti pro kategorická data

6.8.1 x^2-test nezávislosti

– vztah dvou kategorických proměnných popisujeme ***tabulkou absolutních četností***
– označíme:

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

– kontingenční tabulka absolutních četností má tvar:

	Y_1	...	Y_l	
X_1	$n_{1,1}$	...	$n_{1,l}$	$n_{1.}$
⋮		⋱		⋮
X_k	$n_{k,1}$	...	$n_{k,l}$	$n_{k.}$
	$n_{.1}$	...	$n_{.l}$	n

– testované hypotézy:

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

– test je založen na porovnání

- pozorovaných četností $\boldsymbol{n_{ij}}$
- očekávaných četností $\boldsymbol{n_{i.}n_{.j}/n}$
  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 x^2-rozdělení o $\boldsymbol{(k-1)(l-1)}$ stupních volnosti

6.8.2 Fisherův exaktní test

– test nezávislosti pro malá data

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

– pro čtyřpolní tabulku:

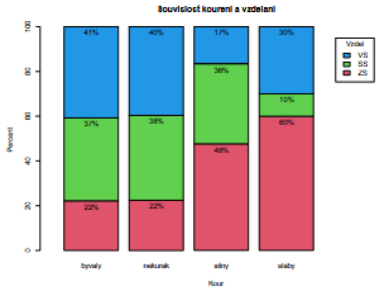
	Y_1	Y_2	
X_1	n_{11}	n_{12}	$n_{1.}$
X_2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

se p-hodnota vypočítá následujícím způsobem: $$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

– pro větší tabulky je test složitější

– Př.: *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



- vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu (můžeme zobrazovat pomocí řádkových nebo sloupcových procent)

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

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

- závěr: Prokázali jsme, že kouření se vzděláním souvisí.

6.9 Poměr šancí

– uvažujme dvouhodnotovou veličinu ve dvou populacích (např. sledujeme výskyt chřipky ve městě a na venkově)

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

– rozdíl mezi populacemi je možné popsat poměrem šancí

- šance "mít chřipku proti nemít chřipku"

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

- poměr šancí je podíl šancí v obou populacích (kolikrát je větší šance na chřipku ve městě než na venkově)

– definice poměru šancí: $$OR = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

– testované hypotézy:

- H_0 : OR = 1, šance jsou stejné
- H_1 : OR ≠ 1, šance se v populacích liší

– testová statistika je rovna: $$Z = \frac{\ln(OR)}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

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

– Př.: *Uvažujme následující čtyřpolní tabulku:*

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

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

6.10 Korelační koeficient

– použití, když je cílem výzkumu zjistit, zda spolu lineárně souvisí dvě číselné proměnné
– rozlišujeme tři ***základní korelační koeficienty***:

- **Pearsonův korelační koeficient**
  – používá se pokud obě proměnné mají přibližně normální rozdělení

- **Spearmanův korelační koeficient**
  – používá se, pokud máme obě proměnné spojité, ale alespoň jedna z nich nemá normální rozdělení

- **Kendallův korelační koeficient**
  – používá se, pokud pracujeme s kategorickými uspořádanými (ordinálními) veličinami

– korelační koeficient měří ***lineární závislost*** dvou proměnných hodnotou z intervalu < −1, 1 >:
- Cor(X, Y) = −1 značí absolutní nepřímou závislost
- Cor(X, Y) = 0 značí lineární nezávislost/ nekorelovanost
- Cor(X, Y) = 1 značí absolutní přímou závislost

– hodnota korelačního koeficientu říká, jak těsný je vztah mezi proměnnými

Pearsonův korelaccní koeficient

– výpočet:

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

Spearmanův korelační koeficient

– se počítá dle stejného vzorce, jen místo původně naměřených hodnot se do něj vkládají pořadí

6.10.1 Korelační test

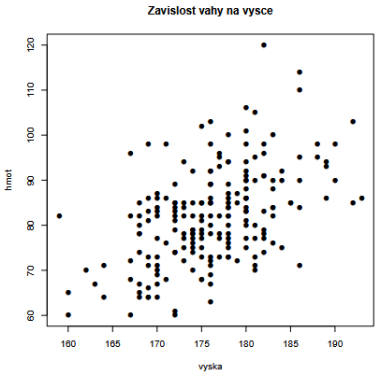
– testované hypotézy:
- H_0 : korelační koeficient = 0
- H_1 : korelační koeficient ≠ 0,
- H_1 : korelační koeficient > 0,
- H_1 : korelační koeficient < 0

– 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

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

- graf
  (z grafu je patrná rostoucí závislost mezi oběma proměnnými)



- testované hypotézy
  - H_0 : váha a výška spolu nesouvisí, korelační koeficient = 0
  - H_1 : váha a výška spolu souvisí, korelační koeficient ≠ 0

- pearsonův korelační koeficient vyšel 0,5
- testová statistika: $$T = \frac{\text{Cor}(X,Y)}{\sqrt{1-\text{Cor}(X,Y)^2}}\sqrt{(n-2)} = \frac{0.5}{\sqrt{1-0.25}}\sqrt{202} = 8.19.$$

  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 α = 0.05
  → nulovou hypotézu tedy zamítáme
- závěr: Souvislost mezi váhou a výškou je průkazná. Spearmanův korelační koeficient vyšel 0.48.

6.10.2 Kendallův korelační koeficient (Kendallovo τ)

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

Kendallovo τ

– založeno na rozdílu počtu souhlasných (n_s) a počtu nesouhlasných (n_n) dvojic

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

– rozptyl tohoto koeficientu: $$\text{Var}(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

– testová statistika ***τ /Var(τ)*** má za platnosti nulové hypotézy asymptoticky ***N(0, 1)*** rozdělení

– pokud se v datech opakují hodnoty
- pro proměnné se ***stejným*** počtem možných hodnot $$\tau_B = \frac{n_s - n_n}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$
  kde $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i-1)/2$ a t_i jsou počty shodných hodnot u proměnné X, $n_1 = \sum_i u_i(u_i-1)/2$ a u_i jsou počty shodných hodnot u proměnné Y.
- pro proměnné s ***různým*** počtem možných hodnot $$\tau_C = \frac{2(n_s - n_n)}{n^2\frac{m-1}{m}},$$
  kde ***m*** je minimální počet hodnot u obou proměnných

– výpočet rozptylů a následných testových statistik pro τ_B a τ_C je složitý (výpočet pomocí softwarů)

7. Statistické modely

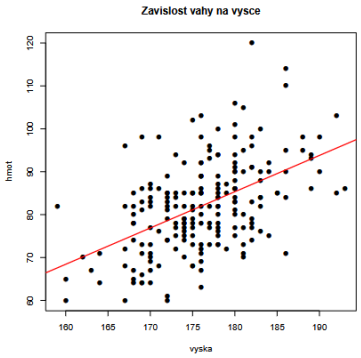
7.1 Lineární regrese

– popisuje příčinnou závislost
– terminologie
- nezávisle proměnná X – příčina
- závisle proměnná Y – důsledek

– odhad lineárního modelu ve tvaru: $Y_i = \beta_0 + \beta_1 X_i + e_i$

kde:
$\boldsymbol{Y_i}$ jsou hodnoty závisle proměnné
$\boldsymbol{X_i}$ jsou hodnoty nezávisle proměnné
$\boldsymbol{\beta_0}$ je absolutní člen
$\boldsymbol{\beta_1}$ je lineární člen
$\boldsymbol{e_i}$ jsou náhodné chyby

– grafický popis pomocí bodového grafu, nutné dodržení os pro proměnné:
- na ***x***-ovou osu se kreslí **ne**závisle proměnná
- na ***y***-ovou osu se kreslí závisle proměnná



– regresní model je rovnice přímky proložené daty

Metoda nejmenších čtverců

– odhad probíhá metodou nejmenších čtverců, která minimalizuje součet druhých mocnin residuí

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

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

– pomocí modelu je možné predikovat budoucí hodnoty závisle proměnné

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

(např. ze známé výšky můžeme predikovat očekávanou hmotnost)

Koeficient determinace

– procento variability závisle proměnné vysvětlené modelem: $R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} = \mathrm{cor}(X, Y)^2$

– na základě modelu lze též zkonstruovat test nezávislosti.
– testované hypotézy:
- H_0 : Proměnná ***Y*** (váha) na proměnné ***X*** (výšce) lineárně nezávisí, $\boldsymbol{\beta_1} = 0$
- H_1 : Proměnná ***Y*** (váha) na proměnné ***X*** (výšce) lineárně závisí, $\boldsymbol{\beta_1} \neq 0$

– test je založen na faktu, že $\boldsymbol{b_1 / se(b_1)} \sim \boldsymbol{N(0, 1)}$, kde $\boldsymbol{b_1}$ je odhad lineárního členu $\boldsymbol{\beta_1}$ a $\boldsymbol{se(b_1)}$ je jeho střední chyba

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

- odhadnutý model má tvar: $Y_i = -66.85 + 0.85X_i$

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

- zamítáme nulovou hypotézu nezávislosti
- koeficient determinace vyšel 0.2493

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

Předpoklady lineární regrese

– 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 bud' na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy

– ukázka grafických testů předpokladu
- 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)