Statistika v Biologii

Alena Černíková

alena.cernikova@ujep.cz

5. listopadu 2026

Podmínky zápočtu

- **tři domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat přes moje internetové stránky – kapitola Statistika v Biologii, odkaz Úkoly
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  zpracování vzájemného vztahu tří proměnných
  souvislý text obsahující všechny podstatné statistické informace

Obsah kurzu

1. Co je statistika
2. Popisné statistiky
3. Pravděpodobnostní rozdělení
4. Bodový a intervalový odhad
5. Testování hypotéz
6. Jednovýběrový,párový a dvouvýběrový test
7. Analýza rozptylu
8. Korelační koeficient a lineární regrese

Co je statistika

Statistika je přesná věda o nepřesných číslech.

Cílem je zjistit chování náhodné veličiny v určité populaci. Celou populaci změřit neumíme. Uděláme náhodný výběr, na kterém změříme sledovanou veličinu, a na základě zjištěných informací děláme závěry pro celou populaci.

Co je statistika

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.*

- Průměrná výška náhodného výběru (200 lidí) vyšla 175 cm.
- Je to nejlepší odhad průměru celé populace, který můžeme v daný moment poskytnout.
- Pravděpodobnost, že jsme přesně trefili populační průměr je 0.
- Odhad je zatížen chybou.

Základní pojmy

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

Čištění dat

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

Typy proměnných

Náhodná veličina se často nazývá proměnná.

Abychom správně určili, co přesně máme pro proměnnou počítat, je třeba nejprve určit její typ.

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

Popisné statistiky

Jak popisujeme jednotlivé typy proměnných

- **Číselné proměnné**
  - popisné statistiky polohy – průměr, medián, vybrané percentily (kvartily, extrémy)
  - popisné statistiky variability – rozptyl, směrodatná odchylka, mezikvartilové rozpětí, koeficient variace
  - popisné statistiky tvaru rozdělení – šikmost, špičatost
  - grafické 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

Popisné statistiky polohy

Potřebujeme stručně popsat, jakých hodnot veličina nabývá

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

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.*

Popisné statistiky polohy

Označme

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

Popisné statistiky polohy

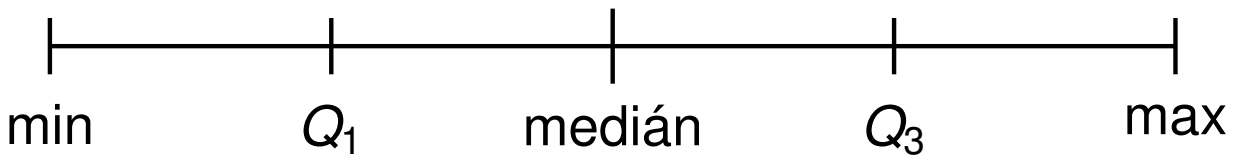
- **průměr** – sečtu všechny hodnoty a vydělím počtem pozorování

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

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

Popisné statistiky polohy

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



Popisné statistiky polohy

Výpočet kvartilů podle **R**

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

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

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

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

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

Grafické popisné statistiky

Pro popis číselné proměnné se používají 2 typy grafů

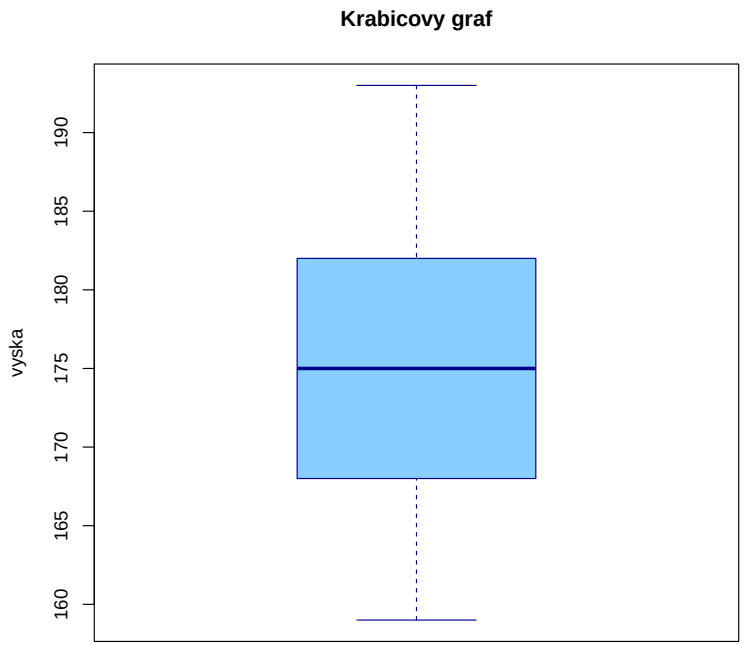
- **Krabicový graf**
  jsou v něm zobrazeny vybrané percentily (medián a kvartily), tykadla dosahují k nejvzdálenějšímu neodlehlému pozorování (odlehlá pozorování se vyznačují zvlášť)
  odlehlé pozorování je takové, které je od bližšího kvartilu dále než jeden a půl násobek mezikvartilového rozpětí
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

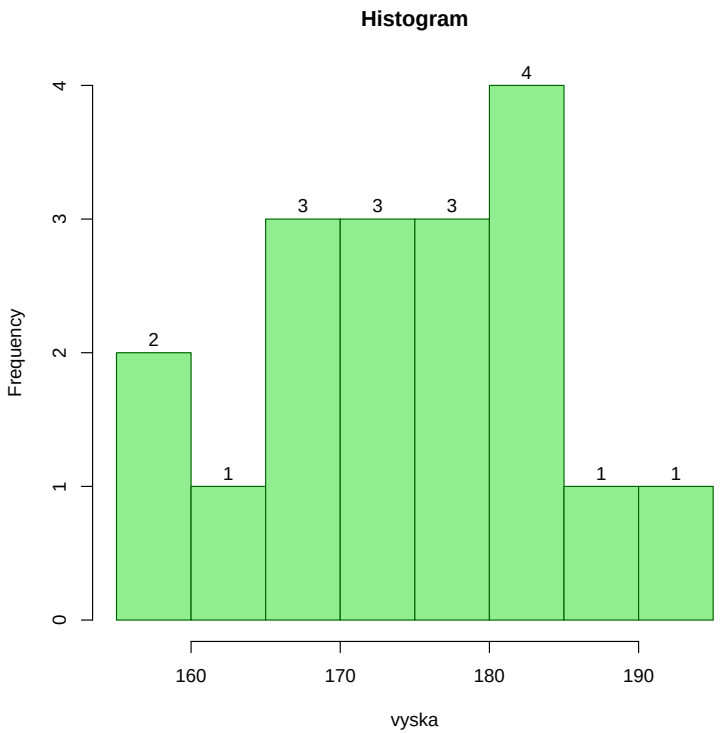
$$k = 1 + 3.32 \log_{10}(n)$$

kde n je počet pozorování

Grafické popisné statistiky

Ukázka grafů pro číselnou proměnnou





Grafické popisné statistiky

Jak interpretovat grafy

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

Popisné statistiky variability

- Rozptyl a **směrodatná odchylka**

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

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

Popisné statistiky variability

- **Mezikvartilové rozpětí**

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

kde Q_3 je třetí kvartil a Q_1 je první kvartil

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

Popisné statistiky variability

- **Variační koeficient**

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

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

Popisné statistiky tvaru rozdělení

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

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

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

Popisné statistiky tvaru rozdělení se neurčují u vícevrcholových rozdělení

Popisné statistiky tvaru rozdělení

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

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

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

Popisné statistiky tvaru rozdělení

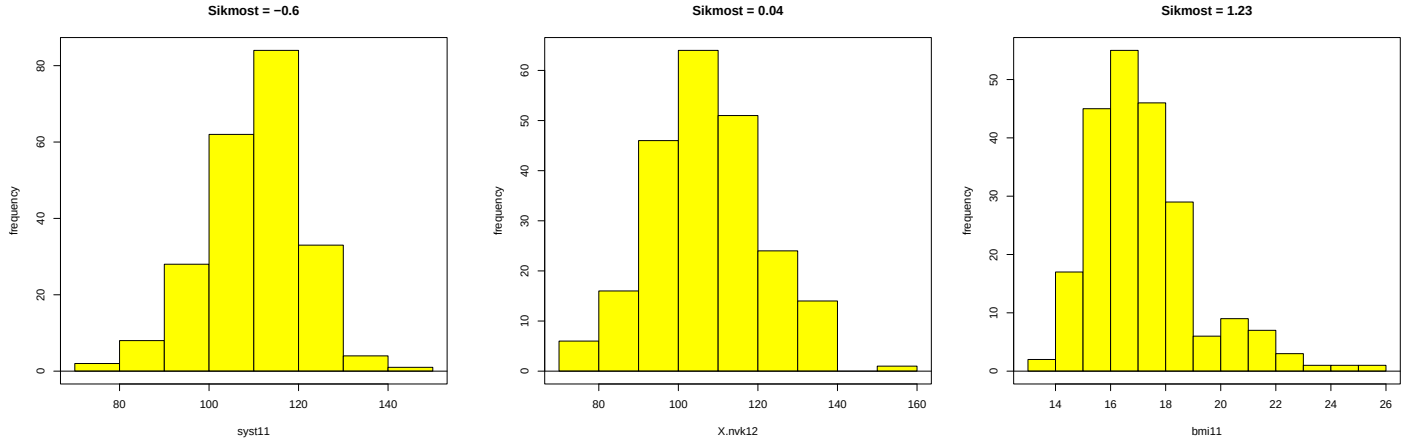
- **Špičatost** – průměr ze čtvrtých mocnin z-skórů mínus 3

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

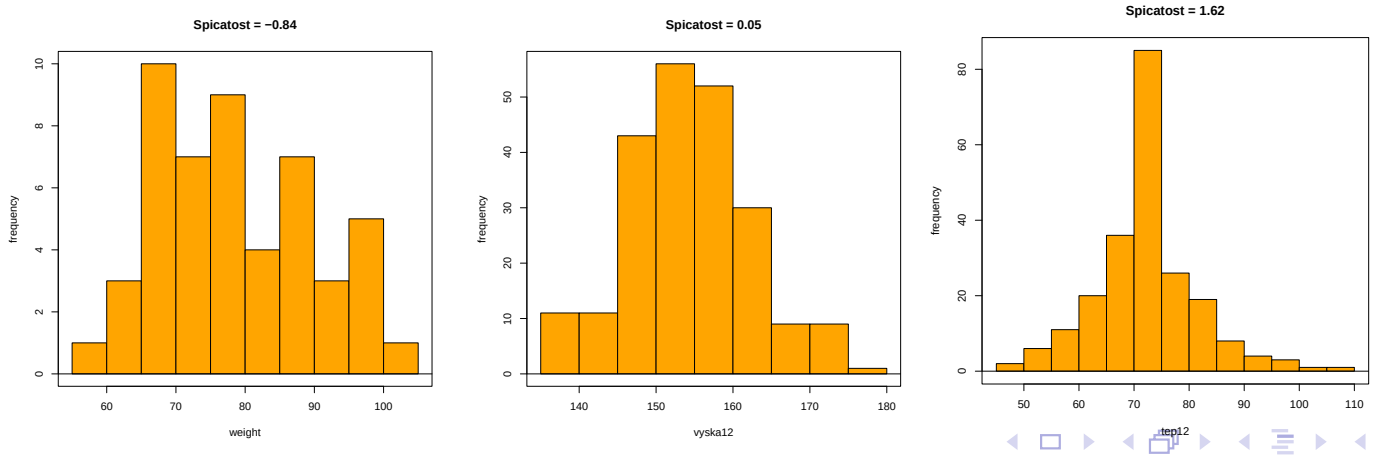
  - porovnává špičatost vůči normálnímu rozdělení
    - nula . . . normální rozdělení
    - záporná hodnota . . . plošší než normální rozdělení
    - kladná hodnota . . . špičatější než normální rozdělení
  - přibližně normální rozdělení má špičatost v intervalu $< -0.5; 0.5 >$
  - špičatost u výrazně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

Popisné statistiky tvaru rozdělení

Ukázka záporné, nulové a kladné šikmosti



Ukázka záporné, nulové a kladné špičatosti



Popisné statistiky výšky

```
> numSummary(Kojeni[,"vyskaM", drop=FALSE], statistics=c("mean", "sd", "se(mean)",
      "IQR", "quantiles", "cv", "skewness", "kurtosis"), quantiles=c
     (0,.25,.5,.75,1), type="3")
   mean     sd se(mean) IQR    cv  skewness kurtosis    0% 25% 50% 75% 100% n
166.9697 6.1103 0.6141   7 0.0366   0.0215 0.2156     153 164 167 171 187 99
```

Popisné statistiky polohy

- průměr: mean $=$ 166.97 cm
- minimum: 0% $=$ 153 cm
- maximum: 100% $=$ 187 cm
- medián: 50% $=$ 167 cm
- dolní kvartil: 25% $=$ 164 cm
- horní kvartil: 75% $=$ 171 cm

počet pozorování: $n = 99$

Popisné statistiky variability

- směrodatná odchylka: sd $=$ 6.11 cm
- střední chyba průměru: se(mean) $=$ 0.6141 cm
- mezikvartilové rozpětí: IQR $=$ 7 cm
- variační koeficient: CV $=$ 0.0366

Popisné statistiky tvaru rozdělení

- šikmost: skewness $=$ 0.0215
- špičatost: kurtosis $=$ 0.2156

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.

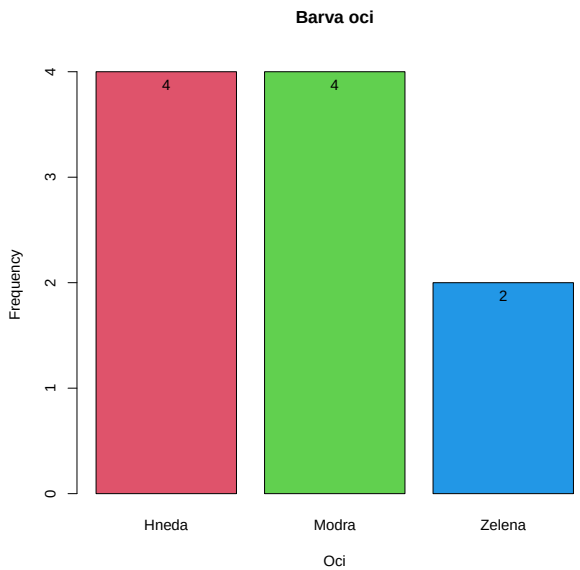
Barva	Absolutní	Relativní (%)
Modrá	4	40%
Hnědá	4	40%
Zelená	2	20%
Celkem	10	100%

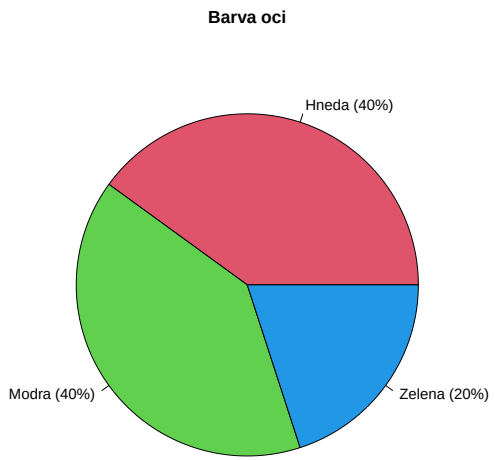
Označme n_j četnosti v jednotlivých kategoriích a n celkový počet pozorování, pak relativní četnost p_j spočteme jako

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

Grafické popisné statistiky

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





Popisné statistiky barvy očí

```
> local({
         .Table <- with(Data, table(BarvaOci))
         cat("\ncounts:\n")
         print(.Table)
         cat("\npercentages:\n")
         print(round(100*.Table/sum(.Table), 2))
       })

counts:
BarvaOci
Hneda Modra   Zelena
4     4       2

percentages:
BarvaOci
Hneda Modra   Zelena
40     40     20
```

Popisné statistiky kategorické proměnné

- absolutní četnosti: counts
- relativní četnosti: percentages

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení popisuje pravděpodobnosti možných výsledků náhodného pokusu.

- **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 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

Příklad

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*

Jak se vypočte **pravděpodobnost náhodného jevu** A?

$$P(A) = \frac{\text{počet příznivých možností}}{\text{počet všech možností}}$$

Náhodné jevy

- **Jev jistý** Ω: soubor všech elementárních jevů, tj. celý prostor možných výsledků, $P(\Omega) = 1$
  Př. na kostce padne číslo od jedné do šesti
- **Jev nemožný** $\emptyset$: jev, který neobsahuje ani jeden elementární jev, $P(\emptyset) = 0$
  Př. na kostce padne mínus jedna
- **Jev opačný** k jevu A, tj. $\overline{A}$: soubor elementárních jevů, které nastanou právě když nenastane jev A, $P(\overline{A}) = 1 - P(A)$
  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

Vztah dvou jevů

- **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*

Věta o celkové pravděpodobnosti

- 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

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

  - prostor všech jevů si rozdělíme na části (např. populaci na muže a ženy)
  - v každé z nich spočítáme pravděpodobnost jevu A (např. pst. že člověk vystuduje vysokou školu)
  - pst jevu A v celém prostoru/ populaci je pak součet pstí v jednotlivých částech (podmíněné psti) krát velikost (pravděpodobnost) těchto částí
  - tj. pst, že člověk vystuduje VŠ můžeme spočítat jako pst, že muž vystuduje VŠ krát pst, že dotyčný je muž, plus pst, že žena vystuduje VŠ krát pst, že dotyčný je žena

Bayesův vzorec

- Jak vypočítat podmíněnou pravděpodobnost $P(B|A)$ ze znalosti $P(A|B)$

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

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*

  - podle vzorce pro výpočet podmíněné pravděpodobnosti je v čitateli pravděpodobnost průniku jevů

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

  - podle vzorce pro úplnou pravděpodobnost je ve jmenovateli pravděpodobnost jevu A

Senzitivita a specificita testu

Charakteristiky medicínských testů označují podmíněné pravděpodobnosti

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

Senzitivita a specificita testu

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

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

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

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

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

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné na

- **Spojitá**: pro číselné proměnné nabývající všech reálných hodnot v určitém intervalu,
  př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní**: pro číselné proměnné které nabývají jasně oddělitelných hodnot
  mohou nabývat i nekonečně hodnot, najčastěji počty
  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
  – definovaná pro všechna rozdělení
- **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
  – pravděpodobnost jedné konkrétní hodnoty u spojitého rozdělení je 0
  – derivace funkce distribuční
- **Kvantilová funkce**: $Q(p) = t : P(X \leq t) = p$
  – inverzní funkce k funkci distribuční
  – odděluje p% nejmenších hodnot

Střední hodnota a rozptyl

Další charakteristiky pro diskrétní i spojitá rozdělení

- Střední hodnota

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

- Rozptyl

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

Binomické rozdělení (diskrétní)

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

Značení $Bi(n, p)$, kde

- n – počet pokusů,
- p – pravděpodobnost úspěchu

Hodnoty **pravděpodobnostní funkce**

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

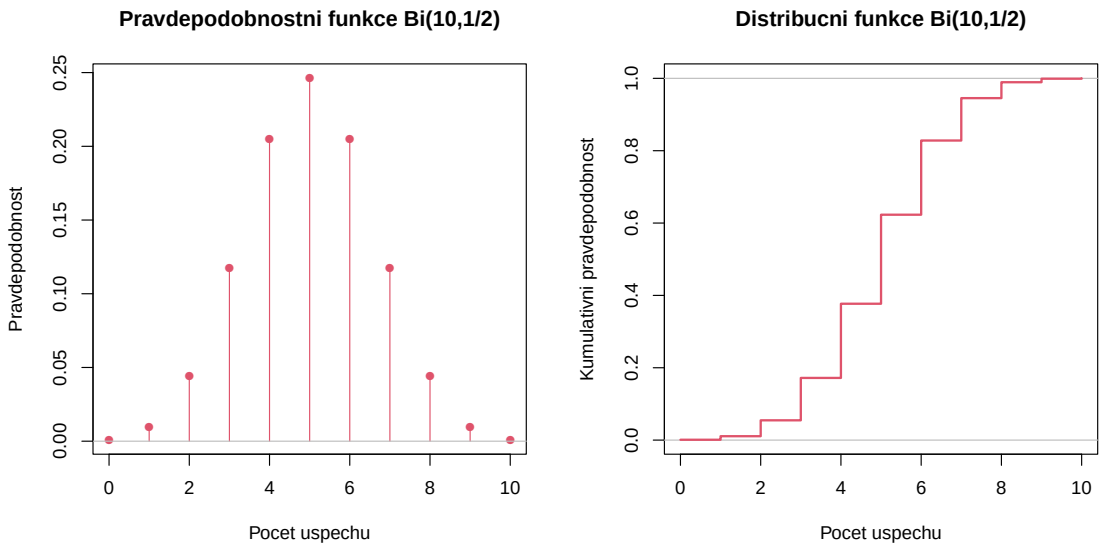
Střední hodnota a **rozptyl**

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

Binomické rozdělení (diskrétní)

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

Pravděpodobnostní a distribuční funkce.



Střední hodnota a rozptyl

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

Normální rozdělení (spojité)

Známé rozdělení s hezkými vlastnostmi. Toto rozdělení má výška lidí určitého věku, IQ, ...

Značení $N(\mu, \sigma^2)$, kde

- μ – střední hodnota
- σ^2 – rozptyl

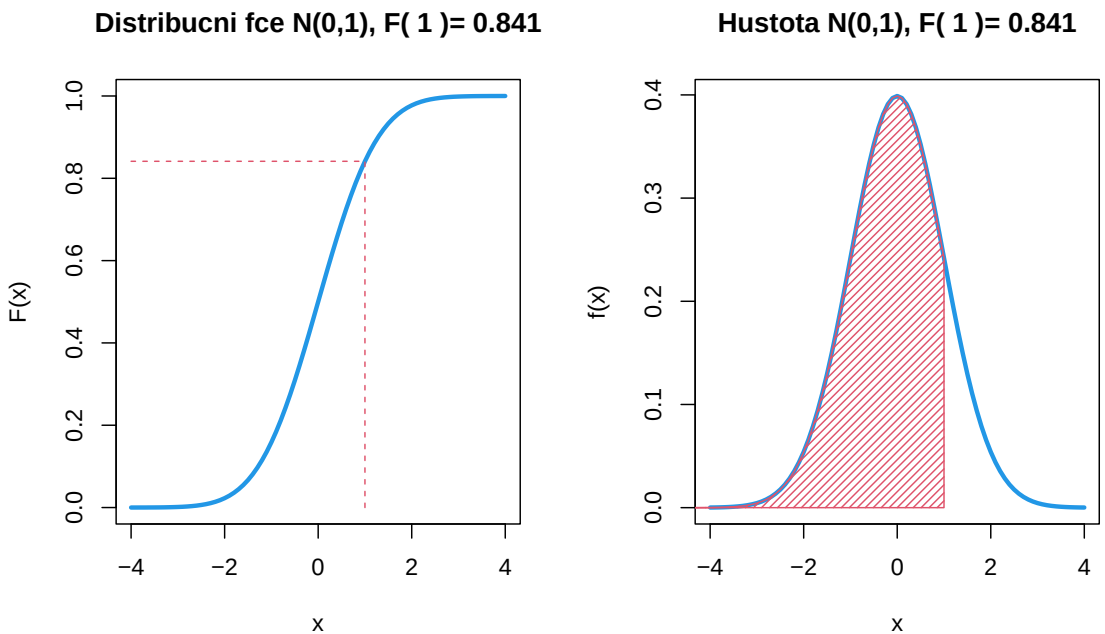
Hustota normálního rozdělení má tvar

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

Je to tak zvaná Gaussova křivka.
Ve statistice se nejčastěji používá standardní normální rozdělení $N(0, 1)$.

Normální rozdělení (spojité)

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

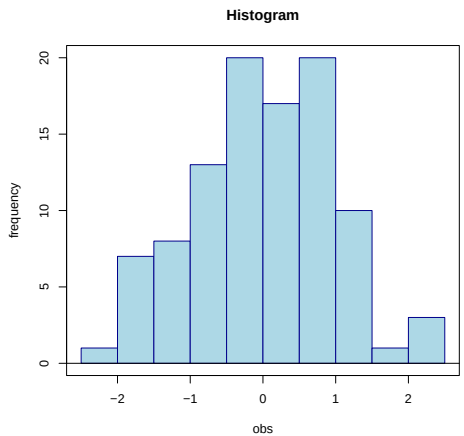


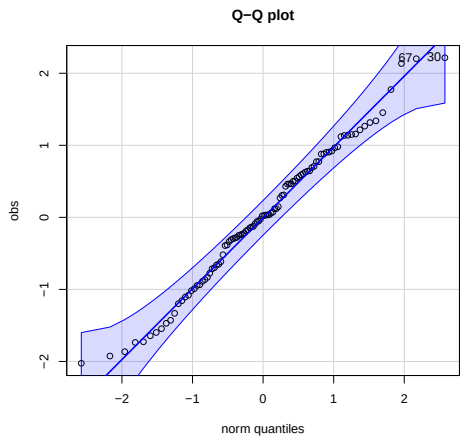
Testování normality

- normální, neboli Gaussovo rozdělení je ve statistice stěžejní
- objevuje se v běžném životě (výška, IQ, hmotnost bochníku chleba, atd.)
- základní statistické postupy, které budeme brát na tomto kurzu, jsou založené na normálním rozdělení
- je potřeba umět normalitu **otestovat**

Grafické testy normality

- základní grafické testy jsou histogram a **pravděpodobnostní graf**





- histogram by měl mít tvar Gaussovy křivky
- v pravděpodobnostním grafu by body měly ležet na přímce

Číselné testy normality

- normalitu lze odhadovat i z popisných statistik tvaru rozdělení: šikmost i špičatost by měly být blízko nuly
- příklady číselných testů Shapiro-Wilkův, Andersonův-Darlingův, Kolmogorovův-Smirnovův, Lillieforsův a další
- tyto testy vrací p-hodnotu
- pokud je **p-hodnota větší než** 0.05**, pak data mají přibližně normální rozdělení**

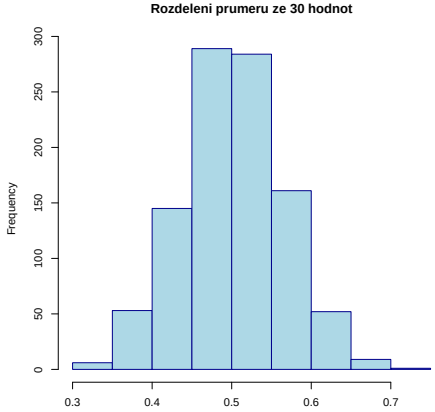
Centrální limitní věta

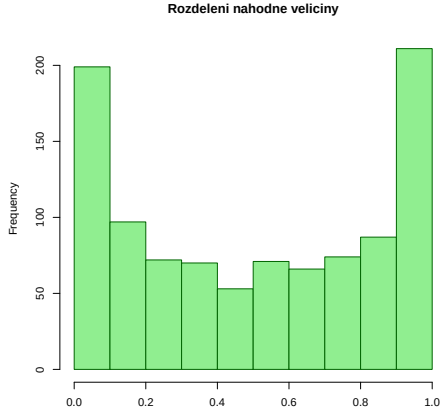
Věta

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

Tedy: 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.





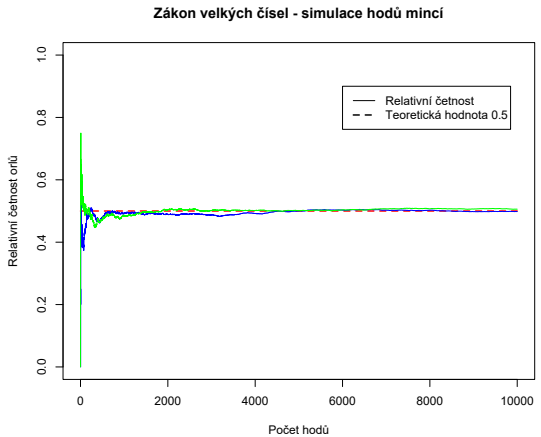
Zákon velkých čísel

Věta

Výběrový průměr nezávislých, stejně rozdělených náhodných veličin se střední hodnotou μ konverguje k této střední hodnotě pro počet náhodných veličin rostoucí nade všechny meze.

Tedy: čím více hodnot průměrujete, tím blíže je průměr k odhadované střední hodnotě.

Ukázka, jak konverguje relativní četnost orlů při opakovaném hodu mincí, k teoretické hodnotě $p = 0.5$.



Bodový odhad střední hodnoty

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

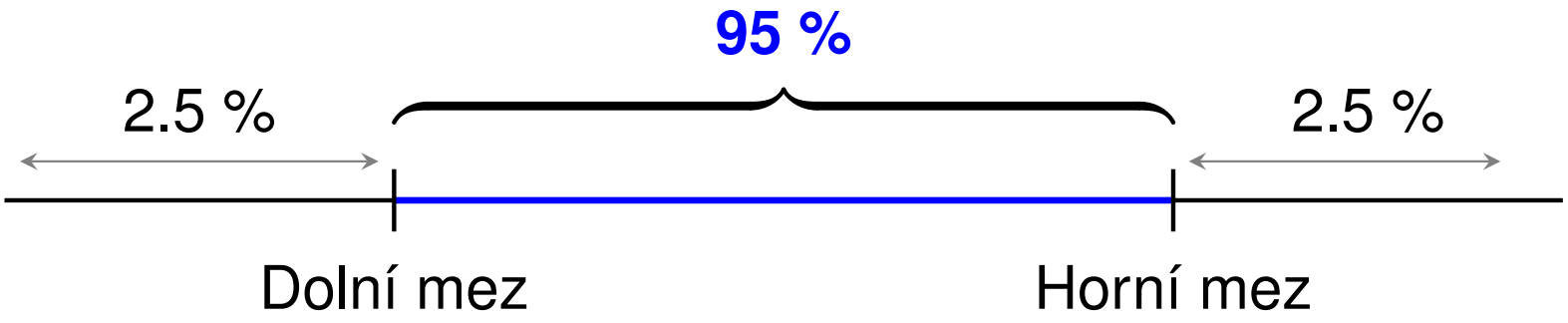
- **nejlepší bodový odhad** je výběrový průměr $\overline{X} = 173.12$ cm
- jaká je pravděpodobnost, že se populační průměr bude rovnat přesně tomuto číslu?
- chyba odhadu, tzv. střední chyba odhadu průměru

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

Intervalový odhad

Nyní chceme spojit bodový odhad a jeho chybu do jedné informace

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



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

Intervalový odhad střední hodnoty

Chceme interval, ve kterém se s vysokou pravděpodobností bude nacházet populační průměr/ skutečná střední hodnota. Z čeho se interval spolehlivosti počítá

- 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 potřebuji

Intervalový odhad střední hodnoty

Výběrový průměr má normální rozdělení

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

- μ je teoretická střední hodnota
- σ je teoretická směrodatná odchylka
- n je počet pozorování
- $N(0,1)$ je standartní normální rozdělení

Interval spolehlivosti pro střední hodnotu, když **znám** skutečný **rozptyl** dat

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

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

Intervalový odhad střední hodnoty

Běžnější je případ, že **rozptyl neznám** a musím ho nahradit výběrovou směrodatnou odchylkou. Pak platí

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

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

meze intervalu spolehlivosti jsou

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

- $t_{n-1}(1-\alpha/2)$ je kvantil t-rozdělení o $n-1$ stupních volnosti
- kvanlity t-rozdělení jsou větší než kvantily normálního rozdělení
- interval spolehlivosti je širší než v předchozím případě

Intervalový odhad střední hodnoty

V R-ku je interval spolehlivosti pro průměr součástí výstupu k jednovýběrovému t-testu

```
> with(Ichs, (t.test(hmot, alternative='two.sided', mu=0.0, conf.level=.95)))

One Sample t-test

data: hmot
t = 110.46, df = 203, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
80.64702 83.57847
sample estimates:
mean of x
82.11275
```

Interval spolehlivosti

- S spolehlivostí 95% se skutečný populační průměr hmotnosti mužů nachází v intervalu od 80.65 kg do 83.58 kg
- v nastavení t-testu je možné měnit i požadovanou spolehlivost (Confidence level)
- důležité je nastavení oboustranné alternativy

Intervalový odhad pravděpodobnosti

Chceme odhadnout parametr binomického rozdělení (p).

Příklad. *Házíme 100 krát kostkou, z těchto 100 hodů mi šestka padla 20 krát. Jaká je pravděpodobnost, že padne 6.*

- **nejlepším bodovým odhadem** je relativní četnost $\hat{p} = n_u/n$
  - n_u je počet úspěchů
  - n je počet pokusů
- **intervalový odhad** vychází z předpokladu, že

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

  pro $n\hat{p}(1-\hat{p}) > 9$
- meze intervalu spolehlivosti jsou

$$\hat{p} \pm z(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}$$

Základy testování hypotéz

Testování hypotéz se využívá, když potřebujeme ověřit platnost nějakého tvrzení, např.

- Nový lék je lepší než ten stávající.
- Průměrná výška lidí je 175 cm.
- Průměrná výška lidí se za posledních 50 let zvýšila.
- Výnosy z jednotlivých druhů jabloní se liší.
- Krevní tlak závisí na hmotnosti.

Testované hypotézy

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme H_0
  – je v ní vždy pouze jedna varianta, např.
  - nový lék je stejný jako ten stávající (rozdíl je roven nule)
  - výška se rovná 175 cm
  - výnosy všech druhů jabloní jsou stejné (rozdíl je roven nule)
- Alternativní hypotézu – značíme H_1
  – obsahuje více možností (interval hodnot), např.
  - nový lék je lepší než ten stávající (rozdíl je větší než nula)
  - výška se liší od 175 cm
  - výnosy druhů jabloní se liší (rozdíl se nerovná nule)

Testované hypotézy

Nejčastěji testované dvojice hypotéz

- **Test o střední hodnotě**
  - H_0: střední hodnota se rovná číslu μ_0
  - H_1: střední hodnota se nerovná číslu μ_0
- **Porovnání výběrů** / skupin
  - H_0: mezi skupinami není rozdíl
  - H_1: mezi skupinami je rozdíl
- **Test nezávislosti**
  - H_0: proměnné spolu nesouvisí
  - H_1: proměnné spolu souvisí

Závěr 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

Jiný závěr udělat nemohu, z čehož plyne, že

- to co mě zajímá (to, co chci prokázat), **musí být v alternativě**
- platnost nulové hypotézy nelze prokázat
- na interpretaci závěru testu je kladen speciální důraz

.

Chyby při testování hypotéz

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 – nejčastěji se volí $\alpha = 0.05$
- **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

	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

Postup testování

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)
- 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
  - testovou statistiku je možné ručně spočítat a kritické hodnoty jsou tabelovány
- statistické softwary vrací jako výsledek testu ***p*-hodnotu**
- *p*-hodnota se také nazývá *aktuální dosažená hladina testu*
- počítá se kombinací hodnoty testové statistiky a příslušné kritické hodnoty
- **definice** *p*-hodnoty
  - pravděpodobnost, že za platnosti H_0
    nastal výsledek, jaký nastal,
    nebo jakýkoliv jiný, který ještě více odpovídá alternativě

Probírané testy

- **jednovýběrový test**: testuje střední hodnotu v jednom výběru
  Statistics → Means → Single Sample t-test
- **párový test**: porovnává střední hodnotu ve dvou závislých výběrech (párovaná data)
  Statistics → Means → Paired t-test
- **dvouvýběrový test**: porovnává střední hodnotu ve dvou nezávislých výběrech
  Statistics → Means → Independent Samples t-test
- **ANOVA**: porovnává střední hodnotu ve více nezávislých výběrech
  Statistics → Means → One way ANOVA
- χ^2 **test**: testuje závislost dvou kategorických proměnných
  Statistics → Contingency tables → Two way table
- **korelační koeficient**: testuje závislost dvou číselných proměnných
  Statistics → Summaries → Correlation test

Statistický test

Co musí obsahovat každý číselný test

- graf (u testů předpokladů není bezpodmínečně nutné)
- formulaci testovaných hypotéz
- porovnání p-hodnoty a hladiny významnosti α (nejčastěji $\alpha = 0.05$)
- zamítnutí nebo nezamítnutí nulové hypotézy
- interpretaci výsledku

U vlastních testů je potřeba kontrolovat splnění předpokladů

Testované hypotézy

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

Testujeme nulovou hypotézu

- 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

Testová statistika jednovýběrového t-testu je

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

- za platnosti nulové hypotézy má t-rozdělení o $n - 1$ stupních volnosti
- její hodnotu porovnáváme s kvantily t-rozdělení (kritické hodnoty)
- je-li rozdíl mezi tím, co jsme naměřili ($\overline{X}$) a nulovou hypotézou μ_0 příliš velký, nulovou hypotézu ZAMÍTÁME

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

Jednovýběrový t-test

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

Testované hypotézy

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

Testujeme na hladině významnosti $\alpha = 0.05$.

Jednovýběrový t-test

```
> with(Deti, (t.test(vyska11, alternative='less', mu=150, conf.level=.95)))

One Sample t-test

data: vyska11
t = -2.5618, df = 221, p-value = 0.005539
alternative hypothesis: true mean is less than 150
95 percent confidence interval:
-Inf 149.5664
sample estimates:
mean of x
148.7793
```

Testová statistika

P-hodnota

Výběrový průměr

Jednovýběrový t-test

Vyhodnocení testu

- testová statistika vyšla

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

- porovnám ji s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$
- jelikož $|T| = 2.5618 > 1.65$, tak nulovou hypotézu zamítáme
- p-hodnota vyšla $p = 0.005 < 0.05 \Rightarrow$ zamítáme H_0
- nakonec je potřeba ověřit normalitu

Závěr: Prokázali jsme, že průměrná výška jedenáctiletých dětí je menší než 150 cm.

Párový test

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

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

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

Závislé výběry poznám tak, že data tvoří přirozené páry.

Párový t-test

Při aplikaci testu je důležité udržet párová data u sebe, (abyste neporovnávali Vaší pravou ruku se sousedovou levou).

Postup testování

- V prvním kroku jsou pro všechny páry vypočteny **rozdíly**:

$$R_i = X_i - Y_i$$

- střední hodnota těchto rozdílů je testována jednovýběrovým testem

Párový t-test

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

Testované hypotézy

- H_0 : výška ve 12 letech – výška v 11 letech $=$ 5 cm
- H_1 : výška ve 12 letech – výška v 11 letech $>$ 5 cm

Párový t-test

Vyhodnocení testu

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

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

- porovnáváme ji s kvantilem t-rozdělení $t_{221}(1 - 0.05) = 1.65$
- jelikož $|T| = 5.9 > 1.65$ tak, nulovou hypotézu zamítáme
- p-hodnota vyšla $p = 7.26 \times 10^{-9} < 0.05 \Rightarrow$ zamítáme H0
- nakonec je potřeba ověřit normalitu rozdílu

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

Dvouvýběrový test

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

- Je nový lék lepší než ten stávající?
- Jsou muži vyšší než ženy?
- Liší se od sebe dva druhy hnojiva?

Testované hypotézy

- H_0 : průměr 1. skupiny – průměr 2. skupiny $= \mu_0$
- H_1 : průměr 1. skupiny – průměr 2. skupiny $\neq \mu_0$, $> \mu_0$,$< \mu_0$

Dvouvýběrový test testuje rozdíl průměrů.

Dvouvýběrový *t*-test

Budeme zde brát dva typy dvouvýběrového testu:

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

Oba tyto testy předpokládají normální rozdělení obou průměrů.

Test shody rozptylů

K tomu, abychom mohli vybrat správnou verzi testu, je potřeba otestovat **shodu rozptylů** v obou výběrech.
Testované hypotézy

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

Testová statistika je

$$F = \frac{\mathrm{Var(X)}}{\mathrm{Var(Y)}}$$

za platnosti H_0 má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti, kde n_1, n_2 jsou rozsahy prvního, respektive druhého výběru.

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika testu má tvar

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

kde

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

za platnosti nulové hypotézy má tato statistika t-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.

Welchův t-test

Testová statistika testu má tvar

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

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

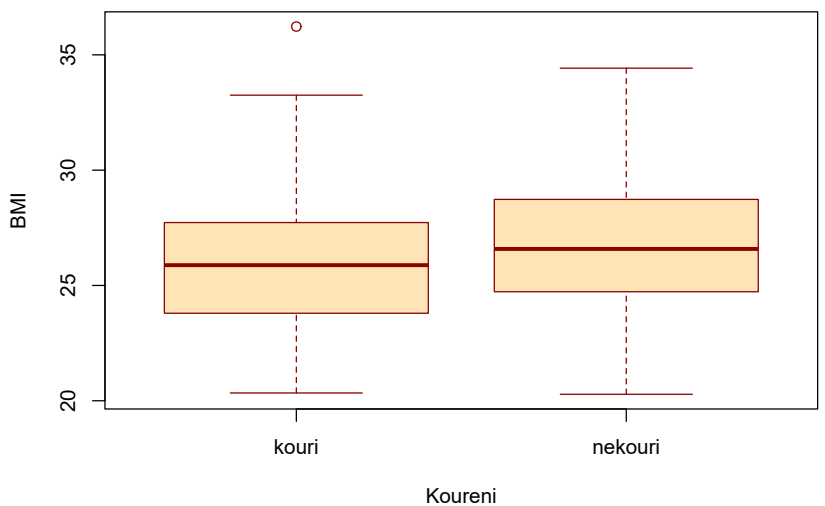
$$\nu = \frac{(\mathrm{Var(X)}/\mathrm{n_1} + \mathrm{Var(Y)}/\mathrm{n_2})^2}{\frac{(\mathrm{Var(X)}/\mathrm{n_1})^2}{n_1 - 1} + \frac{(\mathrm{Var(Y)}/\mathrm{n_2})^2}{n_2 - 1}}.$$

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

Dvouvýběrový t-test

Příklad. *Ve výběru mám 204 mužů středního věku. Z nich 109 kouří a 95 nekouří. Cílem výzkumu je zjistit, zda se Body Mass Index (BMI) liší u kuřáků a nekuřáků. Průměrné BMI u kuřáků je* 25.95 *se směrodatnou odchylkou* 2.8*. U nekořáků je průměrné BMI* 26.98 *a směrodatná odchylka* 3.1*.*

Grafické porovnání



Test shody rozptylů

Testované hypotézy

- H_0 : rozptyly ve skupinách jsou stejné
- H_1 : rozptyly se ve skupinách liší

Vyhodnocení testu

- test shody rozptylů má testovou statistiku

$$F = \frac{\text{Var(X)}}{\text{Var(Y)}} = \frac{7.84}{9.5} = 0.825$$

- p-hodnota vyšla $p = 0.3325 > 0.05$ a nulovou hypotézu nezamítáme
- rozptyly ve skupinách jsou přibližně stejné
- můžeme použít dvouvýběrový t-test pro shodné rozptyly

F-test shody rozptylů

```
> var.test(BMI ~ Koureni, alternative='two.sided', conf.level=.95, data=Ichs)

F test to compare two variances

data: BMI by Koureni
F = 0.82514, num df = 108, denom df = 94, p-value = 0.3335
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.5553633 1.2193409
sample estimates:
ratio of variances
0.8251421
```

Testová statistika

P-hodnota

Interval spolehlivosti pro podíl rozptylů

Podíl porovnávaných rozptylů

Dvouvýběrový t-test

Testované hypotézy

- H_0 : BMI kuřáků a BMI nekuřáků se neliší
  BMI kuřáků – BMI nekuřáků = 0
- H_1 : BMI kuřáků a BMI nekuřáků se liší
  BMI kuřáků – BMI nekuřáků $\neq$ 0

Dvouvýběrový t-test

Vyhodnocení testu

- Testová statistika testu vychází

$$T = \frac{\overline{X} - \overline{Y} - \mu_0}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{25.95 - 26.98}{0.4126}\sqrt{\frac{109 \times 95}{109 + 95}} = -2.49$$

- porovnáme ji s kvantilem t-rozdělení $t_{202}(1 - 0.025) = 1.97$
- jelikož $|T| = 2.4966 > 1.97$ tak nulovou hypotézu zamítáme
- p-hodnota vyšla $p = 0.01334 < 0.05 \Rightarrow$ zamítáme H_0
- nakonec je potřeba ověřit normalitu v obou výběrech

Závěr: Na hladině významnosti 5% jsem prokázala, že BMI u kuřáků a nekuřáků se liší.

Dvouvýběrový t-test

```
> t.test(BMI~Koureni, alternative='two.sided', conf.level=.95, var.equal=TRUE,
      data=Ichs)

Two Sample t-test

data: BMI by Koureni
t = -2.4966, df = 202, p-value = 0.01334
alternative hypothesis: true difference in means between group kouri and group
      nekouri is not equal to 0
95 percent confidence interval:
-1.8405703 -0.2161728
sample estimates:
mean in group kouri mean in group nekouri
25.95205            26.98042
```

Testová statistika

P-hodnota

Interval spolehlivosti pro rozdíl průměrů

Výběrový průměr nekuřáků

Výběrový průměr kuřáků

Analýza rozptylu – ANOVA

Porovnáváme-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.

Analýza rozptylu – ANOVA

Stejně jako u dvouvýběrového t-testu, budeme i zde brát dvě varianty analýzy rozptylu

- klasická analýza rozptylu (ANOVA) – pro shodné rozptyly ve všech výběrech
- Welchova verze analýzy rozptylu – pro různé rozptyly ve výběrech

Předpokladem obou těchto verzí ANOVY je normalita dat ve všech výběrech či ekvivalentně normalita residuí odpovídajícího lineárního modelu.

Klasická ANOVA

Označme

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

Analýza rozptylu rozkládá celkovou variabilitu (čitatel rozptylu)

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

Klasická ANOVA

Rozklad variability na

- variabilitu vysvětlenou výběry (mezi výběry) SS_A
  jedná se o rozdíl mezi skupinovými průměry
- variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SS_e
  jedná se o odchylky od průměru v rámci jedné skupiny

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

Klasická ANOVA

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

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

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

Welchova ANOVA

Welchova ANOVA je založena na vážené variabilitě mezi výběry

$$F \frac{\frac{1}{k-1}\sum_{j=1}^{k} w_j(\overline{X}_j - \overline{X}_w)^2}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2}$$

kde

$$w_j = \frac{n_j}{s_j^2}, \qquad w = \sum_{j=1}^{k} w_j, \qquad \overline{X}_w = \frac{\sum_{j=1}^{k} w_j \overline{X}_j}{w}$$

Testová statistika má F-rozdělení o $k-1$ a ν stupních volnosti

$$\nu = \frac{k^2-1}{3\sum_{j=1}^{k}\left(\frac{1}{n_j-1}\right)\left(1-\frac{w_j}{w}\right)^2}$$

Bartlettův test

Test shody rozptylů pro více než 2 rozptyly.
Testované hypotézy

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

Označme $\mathrm{Var}(\mathrm{X_j})$ výběrový rozptyl v j-tém výběru a

$$S^2 = \frac{\sum_{j=1}^{k}(n_j - 1)\mathrm{Var}(\mathrm{X_j})}{n-k},$$

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

Testová statistika Bartlettova testu

$$B = \frac{1}{C}\left((n-k)\ln S^2 - \sum_{j=1}^{k}(n_j - 1)\ln \mathrm{Var}(\mathrm{X_j})\right)$$

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

Párové srovnání

Porovnání dvojic výběrů

- nelze pomocí dvouvýběrových testů – příliš roste chyba prvního druhu
- prostřednictvím tzv. párového srovnání
- nejčastěji **Tukeyho test**, respektive **Tukey HSD test** pro různě velké výběry

Testované hypotézy

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

Párové srovnání

Testová statistika má tvar

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

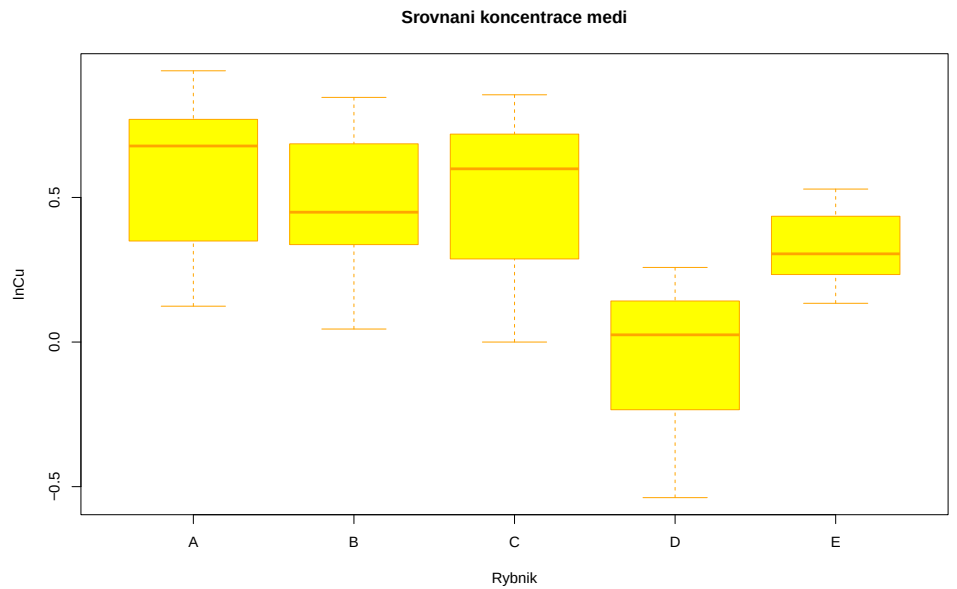
Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty. Vyhodnocení

- na základě p-hodnot nebo intervalů spolehlivosti pro rozdíly všech dvojic
- přestože ANOVA ukáže významný rozdíl mezi skupinami, nemusí se tento projevit v párovém srovnání

Analýza rozptylu – ANOVA

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

Grafické porovnání středních hodnot



Analýza rozptylu – ANOVA

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

Výstupy z R-ka

Bartlettův test

```
> bartlett.test(lnCu ~ Misto, data=Anova2)

Bartlett test of homogeneity of variances

data: lnCu by Misto
Bartlett's K-squared = 3.6675, df = 4, p-value = 0.4529
```

P-hodnota

Tabulka analýzy rozptylu

Součty čtverců

```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value  Pr(>F)
Misto        4  1.796  0.4491  5.896 0.00127 **
Residuals   30  2.285  0.0762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
64 observations deleted due to missingness
```

P-hodnota

Analýza rozptylu – ANOVA

Testované hypotézy

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

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. Nakonec je potřeba ověřit normalitu residuí nebo jednotlivých výběrů

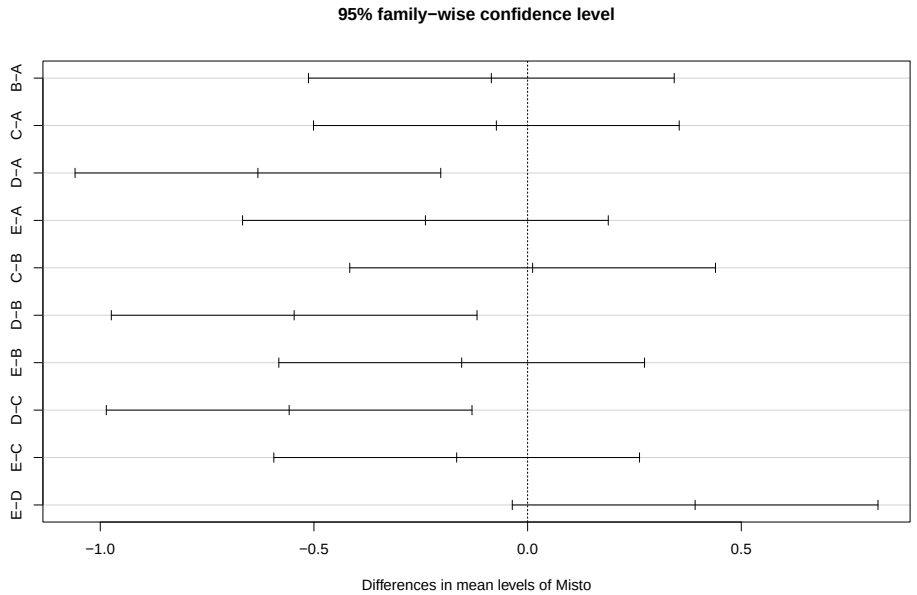
Analýza rozptylu – ANOVA

Párové srovnání vrátí následující tabulku

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

Analýza rozptylu – ANOVA

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



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

Pearsonův korelační koeficient

Lineární vztah dvou číselných proměnných zkoumá **korelační koeficient**. **Pearsonův korelační koeficient** vypočteme jako

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

Libovolný korealční koeficient nabývá hodnot mezi -1 a 1 a platí, že

- absolutní nepřímá závislost má $Cor(X,Y) = -1$
- lineární nezávislost/ nekorelovanost má $Cor(X,Y) = 0$
- absolutní přímá závislost má $Cor(X,Y) = 1$

Pearsonův korelační koeficient

O statistické významnosti závislosti rozhodujeme testem

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

Za platnosti nulové hypotézy platí, že testová statistika

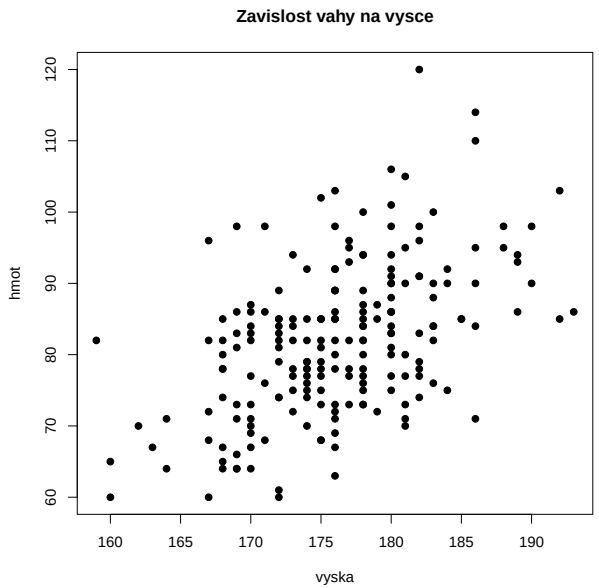
$$T = \frac{\mathrm{Cor}(X, Y)}{\sqrt{1 - \mathrm{Cor}(X, Y)^2}} \sqrt{(n-2)}$$

má t-rozdělení o $n - 2$ stupních volnosti.
Předpokladem použití Pearsonova korelačního koeficientu je normalita obou prpoměnných

Pearsonův korelační koeficient

Příklad. *Výzkumu se účastnilo 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční, u nichž byly měřeny fyzické údaje. Souvisí spolu výška a hmotnost těchto mužů?*

Nejprve grafické porovnání



Z grafu je patrná rostoucí závislost mezi oběma proměnnými.

Pearsonův korelační koeficient

Testované hypotézy

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

Korelační koeficient vyšel 0,5 a testová statistika

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

Testová statistika je větší než kvantil t-rozdělení $t_{202}(1 - 0.975) = 1.97$. P-hodnota testu vyšla 2.926×10^{-14}, což je menší než $\alpha = 0.05$. Nulovou hypotézu tedy zamítáme.
Závěr: Závislost mezi váhou a výškou je průkazná a přímá.
Nakonec je potřeba ověřit normalitu obou proměnných.

Pearsonův korelační koeficient

```
> with(Ichs, cor.test(hmot, vyska, alternative="two.sided", method="pearson"))

Pearson's product-moment correlation

data: hmot and vyska
t = 8.1904, df = 202, p-value = 2.926e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.3885799 0.5958041
sample estimates:
cor
0.4992987
```

Testová statistika

P-hodnota

Interval spolehlivosti pro korelační koeficient

Výběrový korelační koeficient

Lineární regrese

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

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

Výsledkem je odhad lineárního modelu ve tvaru

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

kde

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

Lineární regrese

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

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

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

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

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

Koeficient determinace

Často nazývaný *R-squared*

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

- ukazatel kvality modelu
- procento variability závisle proměnné vysvětlené modelem tj. z kolika procent závisí Y na X a z kolika na něčem jiném

Regresní koeficienty

Interpretace regresních koeficientů

- b_0 – kde protíná regresní přímka x-ovou osu
  – kolik by vyšlo Y, kdyby X bylo 0
- b_1 – o kolik se v průměru změní závisle proměnná Y, když se nezávisle proměnná X zvýší o 1

Test nezávislosti

- H_0 : Y na X lineárně nezávisí, $\beta_1 = 0$
- H_1 : Y na X lineárně závisí, $\beta_1 \neq 0$

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

- test vychází stejně jako u korelačního koeficientu

Lineární regrese

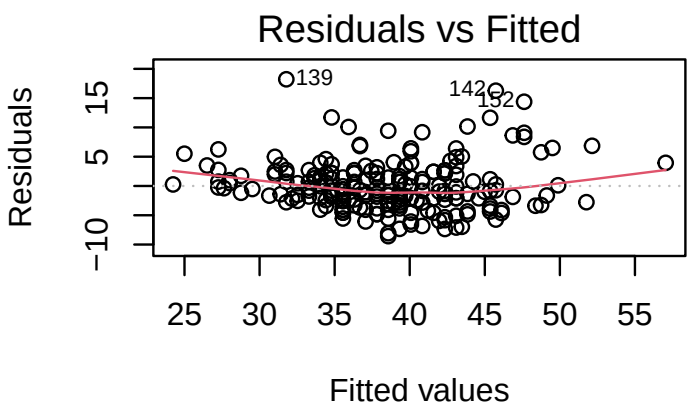
I lineární regrese má své **předpoklady**

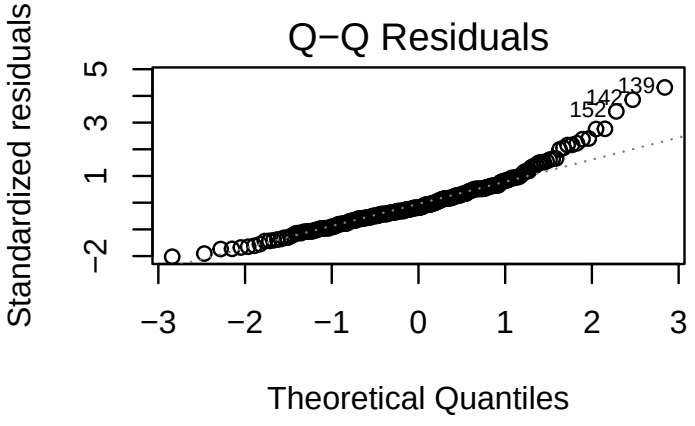
- Mezi proměnnými je skutečně lineární vztah
- Residua jsou nezávislá
- Residua mají normální rozdělení
- Stabilita rozptylu residuí
- V datech nejsou vlivná pozorování

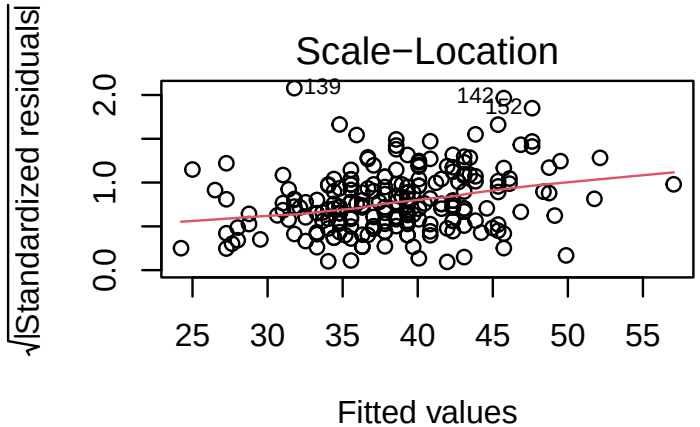
Jednotlivé předpoklady můžeme hodnotit buď na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy.

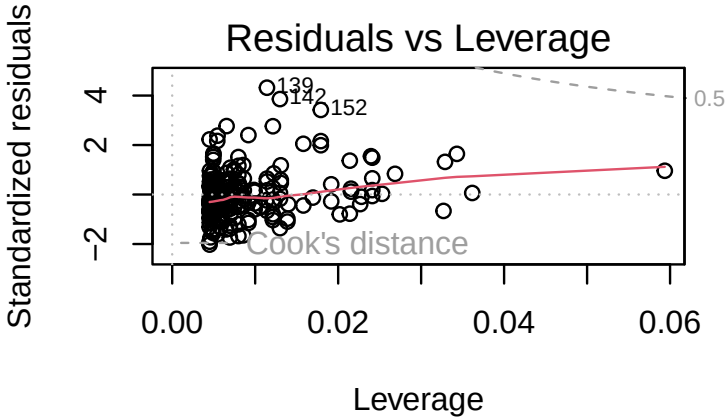
Lineární regrese

Ukázka grafických testů předpokladů









Lineární regrese

Ukázka grafických testů předpokladů

- **1. graf**: lineární vztah – červená čára nemá mít trend
  na grafu je čára do oblouku – není ideální
- **2. graf**: normalita residuí – body mají ležet na přímce
  na grafu leží body na oblouku – není ideální
- **3. graf**: stabilita rozptylu – červená čára nemá mít trend
  na grafu čára roste – není ideální
- **4. graf**: body nemají překročit meze (čárkované křivky)
  na grafu body nepřekračují meze – vše je v pořádku

Lineární regrese

Příklad. *Pokračujme příkladem závislosti hmotnosti na výšce u mužů s jedním rizikovým faktorem ischemické choroby srdeční.*

Odhadli jsme model ve tvaru

$$Y_i = -66.85 + 0.85 X_i$$

- lineární člen ukazuje na přímou závislost: když se výška zvýší o 1 cm, hmotnost se v průměru zvýší o 0.85 kg
- koeficient determinace je $R^2 = 0.25$
- p-hodnota testu vyšla $2.93 \times 10^{-14} < \alpha = 0.05$
- tedy zamítáme nulovou hypotézu

Závěr: U mužů s jedním rizikovým faktorem ischemické choroby srdeční závisí hmotnost na výšce. Závislost je přímá a vysvětlí se jí 25% variability závisle proměnné.

Lineární regrese

```
> summary(RegModel.2)

Call:
lm(formula = hmot ~ vyska, data = Ichs)

Residuals:
Min     1Q Median     3Q    Max
-19.410  -6.362  -0.264   5.372  32.972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -66.8494    18.1990  -3.673  0.000307 ***
vyska         0.8455     0.1032   8.190  2.93e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.222 on 202 degrees of freedom
Multiple R-squared:  0.2493,    Adjusted R-squared:  0.2456
F-statistic: 67.08 on 1 and 202 DF,  p-value: 2.926e-14
```

Regresní koeficienty

P-hodnota

Koeficient determinace

Test dobré shody

Kategorická proměnná s více než dvěma kategoriemi má tzv. **Multinomické rozdělení**. Jedná se o zobecnění binomického rozdělení.

Označme

- k počet kategorií, kterých může náhodná veličina nabývat
- n počet pokusů pokus/ pozorování
- $X_1, \ldots, X_k$ počty, kolikrát nastala která kategorie v n pokusech

Pravděpodobnosti multinomického rozdělení jsou

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \cdot \ldots \cdot n_k!} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k}$$

Střední hodnota a rozptyl

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

Test dobré shody

Test o hodnotách jednotlivých pravděpodobností.
Testované hypotézy

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

Testová statistika má tvar

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

- za platnosti nulové hypotézy má χ^2-rozdělení o $k - 1$ stupních volnosti
- předpokladem je, že všechny očekávané četnosti , tj. hodnoty $n\pi_i$, jsou větší než 5.

Tímto testem je možné testovat i konkrétní rozdělení veličiny.

Test dobré shody

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

Testujeme hypotézy

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

Test porovnává

- *pozorované četnosti*
  $n_1 = 8, n_2 = 5, n_3 = 12, n_4 = 7, n_5 = 9, n_6 = 9$
- *očekávané četnosti* $n\pi_i = 50 \times 1/6 = 8.3333$.

Test dobré shody

Příklad. *Testová statistika*

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

- *kritická hodnota χ^2- rozdělení o 5-ti st. volnosti je $\chi^2_5 = 11.07$*
- *p-hodnota vyšla $p = 0.6569$*
- *testová statistika $X^2 < \chi^2_5$ a p-hodnota $< \alpha$*
- *tedy nezamítáme nulovou hypotézu*

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

χ^2-test nezávislosti

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

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

χ^2-test nezávislosti

Obecná kontingenční tabulka absolutních četností má tvar

	Y_1	$\dots$	Y_c	
X_1	$n_{1,1}$	$\dots$	$n_{1,c}$	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
X_r	$n_{r,1}$	$\dots$	$n_{r,c}$	$n_{r.}$
	$n_{.1}$	$\dots$	$n_{.c}$	n

Hodnoty uvedené v tabulce jsou tzv. *pozorované četnosti*.

χ^2-test nezávislosti

Test je založen na porovnání

- pozorovaných četností v tabulce
- očekávaných četností za platnosti nulové hypotézy

Testované hypotézy

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

Testová statistika má tvar

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

- za platnosti nulové hypotézy má χ^2-rozdělení o $(k-1)(l-1)$ stupních volnosti
- očekávané četnosti odpovídají definici nezávislosti $P(A \cap B) = P(A)P(B)$

Fisherův exaktní test

Pro malé četnosti se používá **Fisherův exaktní test**

- není-li splněn předpoklad χ^2-testu, tj. některé očekávané četnosti jsou menší než 5
- počítá přímo p-hodnotu, tj. pravděpodobnost, že za platnosti H_0 bude pozorována právě naše tabulka četností

Pro čtyřpolní tabulku

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

se p-hodnota vypočítá následujícím způsobem

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

Pro větší tabulky je test složitější.

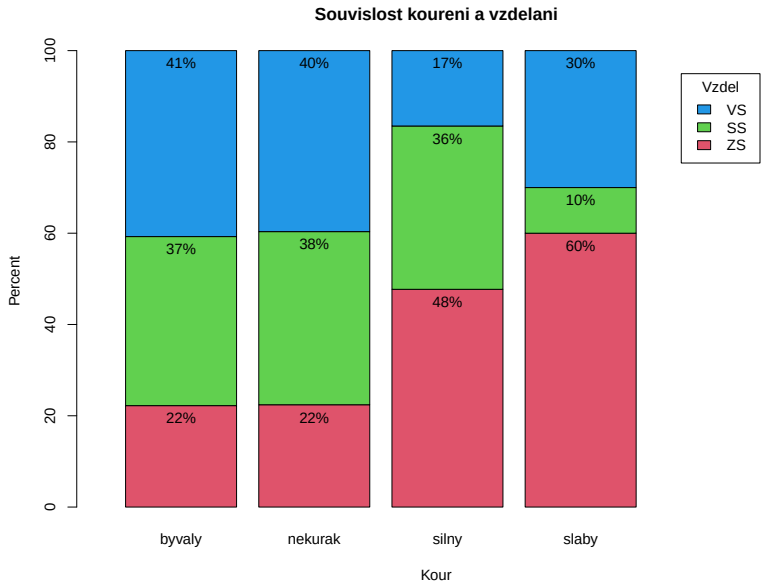
χ^2-test nezávislosti

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

χ^2-test nezávislosti

Vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

χ^2-test nezávislosti

Testem nezávislosti jsme zjišťovali

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

Výsledky testu

- testová statistika χ^2-testu vyšla 21.286
- kvantil χ^2-rozdělení $\chi^2_6 = 12.59$
- p-hodnota testu vyšla 0.00163
- **ale nejsou splněny předpoklady χ^2-testu**
- p-hodnota Fisherova exaktního testu 0.00084
- p-hodnota $< \alpha$ tedy zamítáme nulovou hypotézu

Závěr: Prokázali jsme, že kouření se vzděláním souvisí.

Lineární regrese

```
> local({
      + ...
      + })

Frequency table:
       Vzdel
Kour     SS VS ZS
byvaly   10 11  6
nekurak 22 23 13
silny    39 18 52
slaby     1  3  6

Pearson's Chi-squared test

data: .Table
X-squared = 21.286, df = 6, p-value = 0.00163


Fisher's Exact Test for Count Data

data: .Table
p-value = 0.00084
alternative hypothesis: two.sided
```

V okně Messages informace o splnění předpokladů

Tabulka absolutních četností

P-hodnota chí-kvadrát testu

P-hodnota Fisherova testu

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

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

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

Poměr šancí

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

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

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

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

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

Testová statistika tohoto testu je rovna

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

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

Poměr šancí

Pro poměr šancí je možné spočítat i **interval spolehlivosti**

$$\ln(OR) \pm \left(\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \right) z(\alpha/2).$$

Co je možné tímto intervalem zjistit?
Např. můžeme vyhodnocovat, zda se tento poměr může rovnat nějaké konkrétní hodnotě.

Poměr šancí

Příklad. *Uvažujme následující čtyřpolní tabulku*

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

Šance mít chřipku ve městě vychází $58/17 = 3.41$, *šance mít chřipku na venkově vychází* $32/30 = 1.07$. *Poměr šancí ve městě vs. na venkově vychází* $3.41/1.07 = 3.2$. *Ve městě je více než třikrát větší šance mít chřipku než na venkově.*
Testová statistika vychází 3.27, kritická hodnota 1.96 a p-hodnota 0.001. Testová statistika je větší než kritická hodnota a p-hodnota je menší než α, *zamítáme nulovou hypotézu.*
Závěr: *Ve městě je významně větší šance dostat chřipku než na venkově.*

Poměr šancí

```
> fisher.test(.Table)

Fisher's Exact Test for Count Data

data: .Table
p-value = 0.00208
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
1.444080 7.155019
sample estimates:
odds ratio
3.170026
```

P-hodnota

Interval spolehlivosti pro poměr šancí

Poměr šancí