Pokročilé statistické metody pro biology

Alena Černíková

alena.cernikova@ujep.cz

7. ledna 2026

Podmínky zápočtu

- **tři domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat na disk přes mou stránku
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  uvidíme, jestli letos bude :)
  závěrečná zpráva ze statistického výzkumu

Obsah kurzu

Statistický výzkum

- Statistický výzkum popisuje cílovou populaci na základě náhodného výběru
- rozlišujeme výběrové a populační charakteristiky
- **Zpráva ze statistického výzkumu** obsahuje
  - Cíle výzkumu
  - Rozsah výzkumu (počet pozorování)
  - Popisné statistiky klíčových proměnných
  - Výsledky a interpretace složitějších proměnných
  - Vhodné je přidat několik stěžejních grafů

Popisné statistiky

- **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, relativní, případně kumulativní četnosti

Popisné statistiky tvaru rozdělení

Popisné statistiky tvaru rozdělení se počítají ze standardizovaných proměnných, tak zvaných **Z-skórů**

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

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

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

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

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

Popisné statistiky tvaru rozdělení

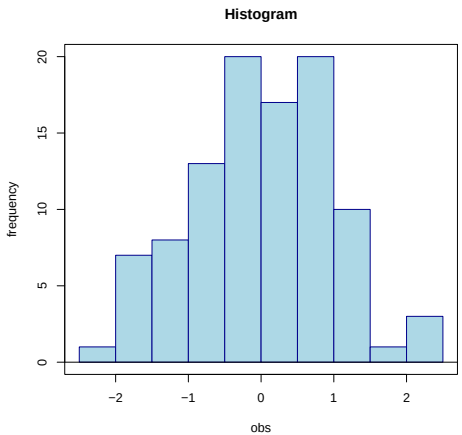
Vlastnosti

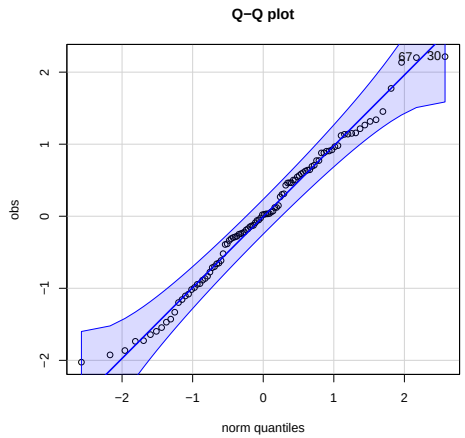
- **Šikmost**
  - nulová – přibližně symetrické rozdělení
  - kladná – rozdělení protažené doprava
  - záporná – rozdělení protažené doleva
- **Špičatost**
  - nulová – špičatost normálního rozdělení
  - kladná – špičatější než normální rozdělení
    uprostřed vyšší sloupce, těžké chvosty
  - záporná – plošší než normální rozdělení
    všechny sloupce podobně vysoké

Testování normality

Jak otestovat normalitu

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





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

Testování normality

Nejčastěji používané číselné testy normality

- **Shapiro-Wilkův** – test odpovídající pravděpodobnostnímu grafu
  porovnává, jak si odpovídají teoretické percentily pro normální rozdělení a percentily naměřené pro sledovanou proměnnou
- **Kolmogorovův-Smirnovův** – test je založen na maximálním rozdílu empirické distribuční funkce a distribuční funkce normálního rozdělení
- **Andersonův-Darlingův** – test je založen na váženém průměru druhé mocniny rozdílu empirické distribuční funkce a distribuční funkce normálního rozdělení

Značení

V průběhu semestru budeme využívat následující značení

- X, Y náhodné veličiny
- n počet pozorování
- i, j pořadový index pozorování / skupiny
- X_i, Y_i konkrétní realizace veličin X, Y
- p_i, q_i pravděpodobnosti konkrétních hodnot u diskrétních rozdělení
- $f(x), f(y)$ hustoty spojitých veličin X, Y
- $f(x, y)$ sdružená hustota veličin X, Y

Odhadované charakteristiky

Nejčastěji odhadujeme následující charakteristiky

- **Pravděpodobnost** náhodného jevu, π_i
- **Střední hodnotu**, E(X)
  definice
  - diskrétní rozdělení: $E(X) = \sum_{i=1}^{n} X_i p_i$
  - spojité rozdělení: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$

  s vlastnostmi:
  - $E(aX + b) = aE(X) + b$
  - $E(X + Y) = E(X) + E(Y)$

Odhadované charakteristiky

Nejčastěji odhadujeme následující charakteristiky

- **Rozptyl**, $\mathrm{Var}(X)$
  definice
  - diskrétní rozdělení: $\mathrm{Var}(X) = \sum_{i=1}^{n}(X_i - E(X))^2 p_i$
  - spojité rozdělení: $\mathrm{Var}(X) = \int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx$

  s vlastnostmi:
  - $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
  - $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{cov}(X, Y)$
- **Korelace**, $\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$ definice
  - diskrétní rozdělení: $\mathrm{cor}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - E(X))(Y_i - E(Y))p_i q_i}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$
  - spojité rozdělení: $\mathrm{cor}(X, Y) = \frac{\int_{-\infty}^{\infty}(x - E(X))(y - E(Y))f(x,y)dxdy}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$

Odhad pravděpodobnosti

- nejlepším bodovým odhadem pravděpodobnosti je relativní četnost $p_i = n_i/n$
- nestranný odhad
- náhodná veličina $p = (p_i - \pi_i)/\sqrt{\pi_i(1-\pi_i)/n}$ konverguje k normálnímu rozdělení $N(0,1)$ pro $n \to \infty$
- intervalový odhad pro pravděpodobnost je

$$p_i \pm z(1-\alpha/2)\sqrt{\frac{p_i(1-p_i)}{n}}$$

- pro použití tohoto intervalu musíme mít dostatečně velké n a p_i, má platit $np_i(1-p_i) > 9$

Odhad střední hodnoty

- nejlepším bodovým odhadem střední hodnoty je výběrový průměr $\overline{X} = \sum_{i=1}^{n} X_i/n$
- nestranný odhad
- platí Centrální limitní věta – pro rostoucí počet pozorování konverguje rozdělení výběrového průměru k normálnímu pro $n \to \infty$
- střední chyba průměru je SEM $=$ sd(X)$/\sqrt{\text{n}}$
- intervalový odhad pro průměr je

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

Odhad rozptylu

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

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

Odhad korelačního koeficientu

- nejlepším bodovým odhadem korelačního koeficientu je výběrový Pearsonův korelační koeficient

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

- máme-li dvourozměrné normální rozdělení a odhadovaný korelační koeficient $|\rho| < 0.5$ pak je interval spolehlivosti pro korelační koeficient

$$\mathrm{Cor}(X, Y) \pm z(1 - \alpha/2)\frac{1 - \mathrm{Cor}(X, Y)^2}{\sqrt{n - 3}}$$

Odhad korelačního koeficientu

- nejsou-li splněny podmínky výše, pak je intervalový odhad pro korelační koeficient odvozen z faktu, že náhodná veličina

$$Z = \frac{1}{2}\ln\left\{\frac{1+\mathrm{Cor}(X,Y)}{1-\mathrm{Cor}(X,Y)}\right\} \sim N\left(\frac{1}{2}\ln\left\{\frac{1+\rho}{1-\rho}\right\} + \frac{\rho}{2(n-1)}, \frac{1}{n-3}\right)$$

- interval spolehlivosti tedy je

$$\mathrm{tgh}(Z \pm z(1-\alpha/2)/\sqrt{n-3})$$

kde $\mathrm{tgh}(x) = (e^x - e^{-x})/(e^x + e^{-x})$

Rozsah výběru

Určení rozsahu výběru na základě **požadované délky intervalu spolehlivosti**

Předpokládejme, že chceme realizovat výzkum, jehož cílem je odhadnout střední hodnotu s požadovanou přesností. Délka intervalu spolehlivosti nesmí přesáhnout hodnotu 2Δ.
Platí

$$\Delta \geq z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

Rozsah výběru pak musí splňovat

$$n \geq \left(z(1-\alpha/2)\frac{\sigma}{\Delta}\right)$$

Testování hypotéz

Opakování:

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu, značíme H_0
  – patří sem jedna hodnota
- Alternativní hypotézu, značíme H_A
  – patří sem interval hodnot

Nejběžnější testované hypotézy

- H_0 : mezi skupinami není rozdíl
  H_1 : mezi skupinami je rozdíl
- H_0 : proměnné spolu nesouvisí
  H_1 : proměnné spolu souvisí
- H_0 : data mají normální rozdělení
  H_1 : data nemají normální rozdělení

Testování hypotéz

Na základě testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu, pak platí alternativa
- Nezamítneme nulovou hypotézu

Při rozhodování můžeme udělat chybu

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

Testování hypotéz

Vyhodnocení testu je založené na porovnání ***p*-hodnoty** a **hladiny významnosti** (α):

- *p*-hodnota $\leq \alpha$ potom ZAMÍTÁME H_0
- *p*-hodnota $> \alpha$ potom NEZAMÍTÁME H_0

Definice ***p*-hodnoty**

- pravděpodobnost, že za platnosti H_0 nastal výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- Předpokládejme, že platí nulová hypotéza. Jak je potom pravděpodobný náš výsledek?
- jinak se nazývá aktuální dosažená hladina testu

Statistický test

Rozlišujeme několik typů testů

- **Parametrické testy**
  - předpokládají normální rozdělení
  - založené na odhadu testovaného parametru
  - t-testy, klasická ANOVA, Pearsonův korelační koeficient
  - bráno v základním kurzu
- **Neparametrické testy**
  - normalitu nepředpokládají
  - založené na pořadích
  - Wilcoxonův test, Kruskal-Wallisův test, Spearmanův korelační koeficient, atd.
- **Permutační testy**
  - nemají žádné požadavky na rozdělení vstupních dat
  - založené na přeuspořádání a náhodném generování z naměřených hodnot

Neparametrické testy

Přístup založený na **pořadích**

Příklad. *Uvažujme naměřené věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50. Data věků rodičů bývají sešikmena a často obsahují odlehlé hodnoty. Přiřadíme-li hodnotám pořadí podle velikosti, získáme řadu 6, 3.5, 7, 9, 3.5, 2, 5, 8, 1, 10. Takto získaná řada není sešikmená a nemá odlehlé hodnoty.*
Nevýhodnou je, že tyto testy bývají **slabší**.

Jednovýběrový test o střední hodnotě

Testované hypotézy

- H_0 : střední hodnota $= \mu_0$
- H_1 : střední hodnota $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Podle rozdělení dat

- pro normálně rozdělená data se používá **t-test**
- pro data, která nemají normální rozdělení, se používá **Wilcoxonův test**
  - Wlcoxonův test testuje hodnotu mediánu

Wilcoxonův jednovýběrový test

Postup jednovýběrového Wilcoxonova testu

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

Pro větší n lze užít transformaci

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

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

Wilcoxonův jednovýběrový test

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

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

Řešení:

- *spočtěme rozdíly* $X_i - m_0$*:* -3, -5, 3, 5, -5, -7, -4, 4, -8, 17
- *jejich absolutním hodnotám přiřaďme pořadí:* 1.5, 6, 1.5, 6, 6, 8, 3.5, 3.5, 9, 10
- *sečtěme kladné (modré) pořadí* $S^+ = 21$ *a záporné (červené) pořadí* $S^- = 34$
- *testová statistika vychází* $U = -0.66$
- *p-hodnota 0,51* $> \alpha (= 0.05)$ *a* H_0 *tedy* ***nezamítáme***
- *střední věk otců se významně neliší od hodnoty 33 let*

Rozsah výběru

Kolik pozorování je potřeba, aby jednovýběrový test splňoval

- hladina významnosti α
- síla testu $1 - \beta$
- očekávaný rozdíl od nulové hypotézy $\mu_1 - \mu_0$
- očekávaná směrodatná odchylka σ
- q značí kvantil standardního normálního rozdělení

Budeme potřebovat **rozsah výběru** n

$$n \geq \left(\frac{q(1-\alpha/2) + q(1-\beta)}{\mu_1 - \mu_0} \sigma \right)^2$$

Příklad. *Pro jednovýběrový t-test na hladině významnosti 0.05, jehož síla by byla 0.9 proti rozdílu od nulové hypotézy o velikosti 4, při očekávané směrodatné odchylce 7, potřebujeme n hodnot, kde n je*

$$n \geq \left(\frac{q(1-0.05/2) + q(0.9)}{4} 7 \right)^2 = 32.2$$

Pro Wilcoxonův test potřebujeme o 15% pozorování více.

Párový test

Test porovnávající dva závislé výběry, (data, která tvoří přirozené páry)

Testované hypotézy

- H_0 : střední hodnota rozdílu párů $= \mu_0$
- H_1 : střední hodnota rozdílu $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Postup párového testu

- spočítám rozdíly mezi všemi páry

$$R_i = X_i - Y_i$$

  kde X_i a Y_i jsou párová měření

- testuje se střední hodnota rozdílů R_i jednovýběrovým testem

Příklad. *Porovnávám věk otce a matky, srovnávám sílu pravé a levé ruky, srovnávám měření před a po podání nějakého léku, atd.*

Dvouvýběrový test

Porovnává dva nezávislé výběry (pozorování nemohu napárovat).
Testované hypotézy:

- H_0 : rozdíl středních hodnot $= \mu_0$
- H_1 : rozdíl středních hodnot $\neq \mu_0$, nebo $< \mu_0$, nebo $> \mu_0$

Podle typu dat:

- normální data a shodné rozptyly: dvouvýběrový t-test pro shodné rozptyly
- normální data a různé rozptyly: dvouvýběrový Welchův t-test
- data, která nemají normální rozdělení: dvouvýběrový Wilcoxonův (Mann-Whitneyův) test

Wilcoxonův dvouvýběrový test

Postup dvouvýběrového Wilcoxonova testu

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

Wilcoxonův dvouvýběrový test

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

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

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

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

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

Wilcoxonův dvouvýběrový test

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

Testované hypotézy

- H_0 : *Studenti v Ústí a v Liberci jsou stejní*
- H_1 : *Studenti v Ústí a v Liberci se liší.*

Řešení:

- *srovnáme všechny hodnoty do řady*
  41, 45, 53, 56, 62, 66, 76, 77, 79, 79, 80, 81, 84
- *hodnotám přiřadíme pořadí*
  1, 2, 3, 4, 5, 6, 7, 8, 9.5, 9.5, 11, 12, 13
- *vypočteme statistiky* $T_1 = 38.5$, $T_2 = 52.5$, $U_1 = 24.5$, $U_2 = 17.5$ *a* $U_0 = 0.5$
- *p-hodnota* $= 0.6678 > \alpha$ *a* H_0 ***nezamítáme***
- *neprokázal se rozdíl mezi studenty v Ústí a v Liberci*

Rozsah výběru

Kolik pozorování je potřeba, aby dvouvýběrový test splňoval

- hladina významnosti α
- síla testu $1 - \beta$
- očekávaný rozdíl mezi výběry $\mu_1 - \mu_2$
- očekávaná smíšená směrodatná odchylka σ
- q značí kvantil standardního normálního rozdělení

Budeme potřebovat **rozsah výběru** n

$$n \geq 2 \left(\frac{q(1-\alpha/2) + q(1-\beta)}{\mu_1 - \mu_2} \sigma \right)^2$$

Příklad. *Pro dvouvýběrový t-test na hladině významnosti 0.05, jehož síla by byla 0.9 při rozdílu průměrů mezi skupinami 4 a očekávané směrodatné odchylce 7, potřebujeme n hodnot, kde n je*

$$n \geq \left(\frac{q(1-0.05/2) + q(0.9)}{4} 7 \right)^2 = 64.3$$

Pro Wilcoxonův test potřebujeme o 15% pozorování více.

Analýza rozptylu – ANOVA

Porovnáváme-li střední hodnotu ve více než dvou nezávislých výběrech.

Testované hypotézy:

- 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.

Podle typu dat:

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

Příklad. *Byla měřena koncentrace mědi v těle ryb. Porovnáváno bylo 5 rybníků, kde z každého byl vyloven vzorek alespoň 10-ti ryb. Liší se od sebe tyto rybníky?*

Kruskal-Wallisův test

Postup Kruskal-Wallisova testu

- obdoba s dvouvýběrovým Wilcoxonovým testem
- srovnáme všechny naměřené hodnoty do řady
- určíme jejich pořadí
- pro každý výběr sečteme pořadí a součet označíme $T_i, i = 1, \ldots, k$, kde k je počet výběrů
- testová statistika

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

má za platnosti H_0 χ^2-rozdělení

Dunnův test

Párové srovnání pro data, která nemají normální rozdělení

- porovnání všech dvojic výběrů při udržení celkové hladiny významnosti

Testová statistika porovnávající i-tý a j-tý výběr je

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

- 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

Dunnův test

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

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

kde S_l je počet l-té shodné hodnoty.

- platí pro ni stejná pravidla jako na slidu výše

ANOVA pro opakovaná měření

Porovnává několik závislých výběrů.
Příklady

- **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 jaterních testů.

Testované hypotézy

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

ANOVA pro opakovaná měření

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

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

Test je pak založen na porovnání SSA a SSz.

Friedmanův test

Porovnává závislé výběry, které nemají normální rozdělení.

Postup Friedmanova testu:

- určíme pořadí hodnot v rámci každého jedince
- tato pořadí se sečtou pro každý výběr zvlášť
- označme součet těchto pořadí $T_i, i = 1, \ldots, k$
- testová statistika

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

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

Pearsonův korelační koeficient

Je-li cílem výzkumu zjistit, zda spolu lineárně souvisí dvě číselné proměnné, používá se **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 koeficient $= 0$
- H_1 : korelační koeficient $\neq 0, > 0, < 0$

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.

Pearsonův korelační koeficient

V případě, že chceme testovat konkrétní hodnotu korelačního koeficientu, tedy

- H_0 : korelační koeficient $= \rho_0$
- H_1 : korelační koeficient $\neq \rho_0, > \rho_0, < \rho_0$

pak se využívá tzv. Fisherovy Z-transformace, která říká, že

$$Z = \frac{1}{2} \ln \left\{ \frac{1 + \mathrm{Cor}(X, Y)}{1 - \mathrm{Cor}(X, Y)} \right\} \sim N \left(\frac{1}{2} \ln \left\{ \frac{1 + \rho}{1 - \rho} \right\}, \frac{1}{n - 3} \right)$$

kde ρ je skutečná/ teoretická hodnota korelačního koeficientu.

Pearsonův korelační koeficient

Pomocí Z-transformace je možné porovnávat i dva korelační koeficienty mezi sebou.
Testují se hypotézy

- H_0 : korelační koeficienty jsou stejné $\rho_1 = \rho_2$
- H_1 : korelační koeficienty se liší $\rho_1 \neq \rho_2$

Testová statistika

$$U = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

má při shodě porovnávaných korelačních koeficientů $N(0, 1)$ rozdělení.

Spearmanův korelační koeficient

Pokud chceme otestovat, zda spolu souvisí dvě číselné proměnné, které nemají normální rozdělení (ale stále se jeví jako spojité), používá se **Spearmanův korelační koeficient**. Stejně jako další neparametrické testy je založen na pořadích.
Postup

- Hodnoty každé proměnné převedu na pořadí.
- Spočítá se Pearsonův korelační koeficient pro tato pořadí.

Spearmanův korelační koeficient měří monotónní vztah dvou veličin. Je tedy obecnější než Pearsonův korelační koeificient, který měřil jen lineární závislost.

Kendallův korelační koeficient

Pokud chceme zjistit, zda je lineární vztah mezi dvěma uspořádanými kategorickými proměnnými, používá se **Kendallův korelační koeficient** (Kendallovo τ).

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

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

Kendallův korelační koeficient

Konkrétně je **Kendallovo** τ definováno jako

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

Rozptyl tohoto koeficientu je

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

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

Kendallovo τ

Výše uvedený koeficient funguje dobře, pokud v datech nejsou stejné hodnoty. Pokud se stejné hodnoty vyskytnou, používají se následující obdoby tohoto koeficientu.
Pro proměnné se **stejným počtem možných hodnot**

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

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

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

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

Lineární regrese

Lineární regrese zkoumá příčinnou závislost jak jedna proměnná závisí na jiné/ jiných.
Označme

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

Předpokládáme lineární model ve tvaru

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

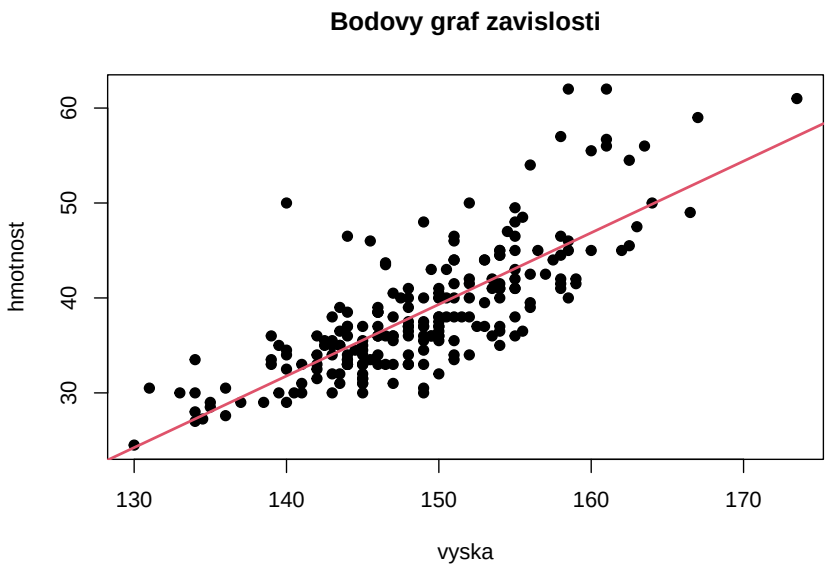
kde

- 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



Lineární regrese

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

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

- R_i jsou residua (rozdíl skutečné a predikované hodnoty)
- $\hat{Y}_i$ jsou odhady, nebo též predikce,
- b_0, b_1 jsou odhady regresních koeficientů (absolutní a lineární člen)
- predikci hodnoty Y v bodě x_0 získáme jako

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

např. ze známé výšky můžeme predikovat očekávanou hmotnost

Lineární regrese

Koeficient determinace

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

- kolik procent variability závisle proměnné se modelem vysvětlí
- tedy z kolika procent závisle proměnná závisí na X a z kolika na něčem jiném

Testované hypotézy **testu nezávislosti**.

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

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

Lineární regrese

Příklad. *Počítejme závislost hmotnosti na výšce u jedenáctiletých dětí.*

- odhadnutá regresní závislost $Y_i = -73.81 + 0.75X_i$
- střední chyba odhadu lineárního členu 0.04
- testovou statistiku 18.76 jsme porovnali s kvantilem t-rozdělení $t_{220}(1 - 0.975) = 1.97$
- p-hodnota testu vyšla $< 2.2 \times 10^{-16}$, což je menší než $\alpha = 0.05$
- zamítáme nulovou hypotézu nezávislosti
- Koeficient determinace vyšel 0.6153.

Závěr: U mužů s jedním rizikovým faktorem ischemické choroby srdeční závisí hmotnost na výšce. Závislost je přímá a vysvětlí se jí 62% variability závisle proměnné.

Lineární regrese

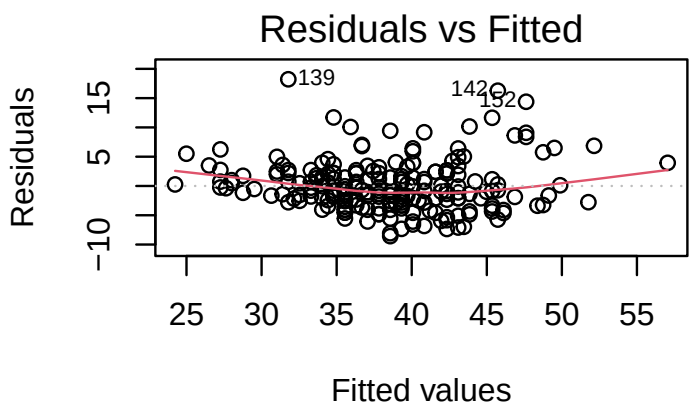
I lineární regrese má své **předpoklady**

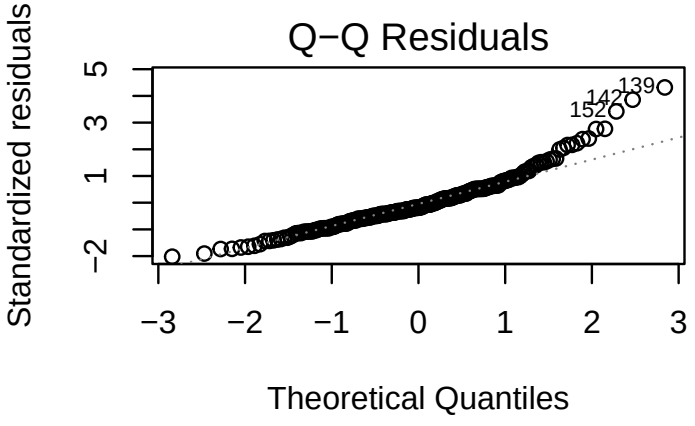
- Mezi proměnnými je skutečně lineární vztah
- Residua jsou nezávislá
- Residua mají normální rozdělení
- Stabilita rozptylu
- V datech nejsou vlivná pozorování

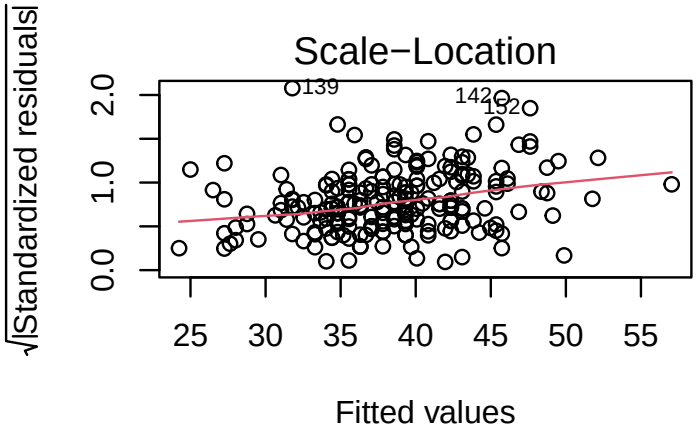
Jednotlivé předpoklady můžeme hodnotit buď na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy.

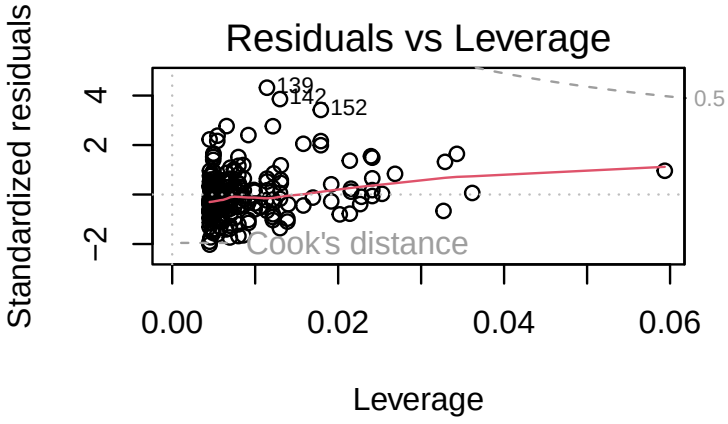
Lineární regrese

Ukázka grafických testů předpokladů









Lineární regrese

Ukázka grafických testů předpokladů

- **1. graf**: lineární vztah – červená čára nemá mít trend
  na grafu je do oblouku, lineární závislost není dostatečná
- **2. graf**: normalita residuí – body mají ležet na přímce
  na grafu se v horní části odchylují, normalita není splněna
- **3. graf**: stabilita rozptylu – červená čára nemá mít trend
  na grafu roste, rozptyl není stabilní
- **4. graf**: body nemají překročit meze (čárkované křivky)
  na grafu vše v pořádku

Lineární regrese

Model **mnohonásobné lineární regrese** má tvar

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

Příklad. *Zkoumáme, o kolik stoupne/klesne voda v řece v závislosti na srážkách, na teplotě, na typu půdy, na nasycenosti půdy, na nadmořské výšce, atd.*
Některé proměnné mají na závisle proměnnou větší vliv, jiné menší.

Lineární regrese

Optimální model, ve kterém budou jen proměnné s významným vlivem, hledáme pomocí **krokové regrese**.

- **Dopředná (forward)**: začíná s modelem bez nezávisle proměnných a v každém kroku přidá jednu s největším, statisticky významným vlivem
- **Zpětná (backward)**: začíná s úplným modelem a v každém kroku vynechá jednu proměnnou s nejmenším, statisticky nevýznamným vlivem
- **Kombinace obou předchozích (both sided)**: začíná s prázdným modelem bez nezávisle proměnných a v každém kroku přidá jednu proměnnou s největším, statisticky významným vlivem a poté zkontroluje, zda nelze jinou proměnnou vynechat.

Cílem je získat model, kde budou nezávisle proměnné pouze se statisticky významným vlivem.

Lineární regrese

Nezávislá **kategorická proměnná** v regresním modelu.

- kategorickou proměnnou s k kategoriemi, reprezentujeme pomocí $k-1$ pomocných *dummy* proměnných
- **dummy proměnné** jsou 0-1 proměnné

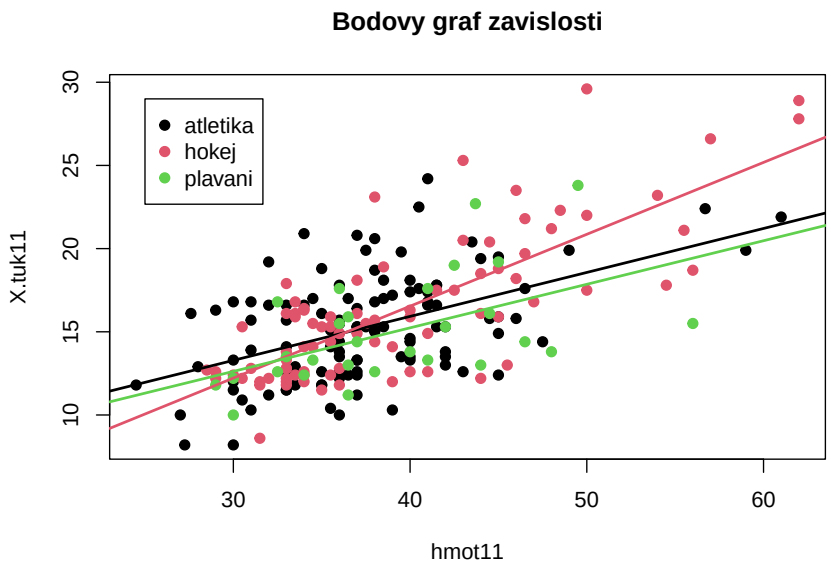
$$\begin{aligned} X_i &= 1 \ldots \text{nastala i-ta kategorie} \\ &= 0 \ldots \text{jinak} \end{aligned}$$

- k-tá kategorie nastane, pokud $X_1 = \ldots = X_{k-1} = 0$
- každá z *dummy* proměnných má svou p-hodnotu
- významnost vlivu celé kategorické proměnné se řeší přes **tabulku analýzy rozptylu**

Lineární regrese

Interakce popisují způsob, jímž se dvě nezávisle proměnné ovlivňují při jejich současném vlivu na proměnnou závislou.

Příklad. *Do výběru bylo zařazeno 222 jedenáctiletých dětí a bylo u nich zjišťováno, jak závisí procento tuku v těle na jejich váze a na sportu, kterému se věnují.*



Interakce jsou vidět již z grafu, do kterého se vykreslí závislost číselných proměnných zvlášť pro každou kategorii proměnné kategorické. Pokud interakce v datech jsou, pak se zobrazí různoběžné přímky. Pokud interakce v modelu nejsou, ale skupiny se od sebe liší, zobrazí se rovnoběžné přímky. Pokud rozdíl mezi skupinami není, přímky splývají.

Příklad. *V modelu interakce existují - závislost procenta tuku na hmotnosti je jiná pro lední hokejisty a pro ostatní.*

Lineární regrese

Informační kritéria hodnotící model.
Každé z níže uvedených kritérií je založeno na věrohodnosti modelu (L - *likelihood*), tj. na ukazateli, jak dobře model kopíruje data. Tato věrohodnost se dále penalizuje počtem parametrů použitých v modelu, k. Platí, že čím menší hodnota kritéria, tím lepší je model.

- **Akaikeho informační kritérium (AIC)**:

$$AIC = 2k - 2\ln(L)$$

- **Upravené Akaikeho informační kritérium (AICc)** pro malé vzorky:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- **Bayesovské informační kritérium (BIC)**:

$$BIC = \ln(n)k - 2\ln(L)$$

Zobecněná lineární regrese

Co dělat, když nejsou splněny předpoklady na rozdělení náhodné chyby modelu?

- Závisle proměnná je **spojitá** – použijeme pro závisle proměnnou transformaci, která ji posune k normálnímu rozdělení. Nejčastěji se používá přirozený logaritmus, nebo Box-Coxova transformace.
- Závisle proměnná je **dvouhodnotová** (0-1) – použije se logistická regrese.
- Závisle proměnnou tvoří **počty** – použije se Poissonova regrese.
- Závisle proměnná je **ordinální** – použije se ordinální regrese.

Logistická regrese

Závisle proměnná je dvouhodnotová

- pravděpodobnost hodnoty 1 označme π
- modelujeme, jak π závisí na nezávisle proměnných

$$\frac{\pi}{1-\pi} = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon\}$$

- parametr $\frac{\pi}{1-\pi}$ se jmenuje **šance** a počítá se jako pravděpodobnost, že jev nastal, vs. pravděpodobnost, že jev nenastal
- v praxi se modeluje logaritmus šancí pomocí klasické lineární regrese

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \epsilon$$

- funkce na levé straně rovnosti se jmenuje ***logit link***

Logistická regrese

Pokud z odhadnutého modelu pak chceme zpětně získat vztah pro pravděpodobnost π, použijeme

$$\pi = \frac{\exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k\}}{1 + \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k\}}$$

Interpretace koeficientu β_1:
"Šance vlastnost mít mi při nárůstu X_1 o 1 vzroste průměrně $\exp \beta_1$ krát při stejných hodnotách ostatních nezávisle proměnných."

Logistická regrese

Příklad. *Uvažujme 150 cestujících na Titaniku. Ke každému cestujícímu máme uvedeno pohlaví, věk, třídu, ve které cestoval a informaci, zda se zachránil nebo ne. Následujícíc tabulky ukazují, jací cestující se zachránili, a jací se utopili.*

	Muž	*Žena*
Přežil	*23*	*26*
Nepřežil	*89*	*12*

	Dospělý	*Dítě*
Přežil	*46*	*3*
Nepřežil	*97*	*4*

	1. třída	*2. třída*	*3. třída*	*Posádka*
Přežil	*11*	*10*	*11*	*17*
Nepřežil	*2*	*12*	*39*	*48*

Logistická regrese

Příklad. *Označme π pravděpodobnost, že dotyčný přežil. Odhad modelu logistické regrese vyšel*

$$\ln\left(\frac{\pi}{1-\pi}\right) = 1.45 - 1.58(2.tr) - 3.04(3.tr) - 1.72(posadka) + \\ +2.32(zena) - 0.9(dospely)$$

Jako vyznamné vyšly proměnné pohlaví ($p = 2.55 \times 10^{-6}$) a třída (p=0.0014) v níž dotyčný cestoval, konkrétně se významně liší třetí třída od první třídy a na hladině významnosti 0.1 i posádka od první třídy.

Ženy mají $\exp(2.32) = 10.17$ *krát větší šanci na přežití než muži při ostatních parametrech neměnných.*
Cestující v první třídě mají $1/\exp(-3.04) = 20.9$ *krát větší šanci přežít než cestující ve třetí třídě při ostatních parametrech neměnných. Cestující v první třídě mají* $1/\exp(-1.72) = 5.6$ *krát větší šanci přežít než posádka při ostatních parametrech neměnných.*

Základy mnohorozměrné statistiky

Předpokládejme, že nemáme jednu proměnnou X, ale vektor proměnných $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$.

Příklad. *Měříme několik fyzických parametrů jedince: výška, váha, krevní tlak, vitální kapacitu plic, atd. Každý žák na vysvědčení dostane známku z několika předmětů: čeština, matematika, zeměpis, přírodopis, atd.*

- Namísto jedné střední hodnoty μ a jednoho rozptylu σ^2 máme vektor středních hodnot $\mu = (\mu_1, \ldots, \mu_k)^T$ a varianční matici $\Sigma = (\sigma_{ij})$
- odhadujeme je pomocí vektoru průměrů $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_k)^T$ a maticí $\mathbf{S} = (\mathbf{s_{ij}})$, kde $s_{ij} = \mathrm{cov}(X_i, X_j)$ pro $i \neq j$ a $s_{ii} = \mathrm{Var}(X_i)$

Základy mnohorozměrné statistiky

Zobecnění základních statistických metod.

- Dvouvýběrový test $\Rightarrow$ **Hotellingův test**
- Analýza rozptylu (ANOVA) $\Rightarrow$ **MANOVA**
- Korelační koeficient $\Rightarrow$ **Kanonické korelace**
- Lineární regrese $\Rightarrow$ **Mnohorozměrná lineární regrese**, kde závisle proměnná má více složek.

Hotellingův test

Porovnávám střední hodnotu náhodného vektoru ve dvou populacích. Předpokládám nezávislá měření. Testuji

- H_0 : vektory středních hodnot se rovnají
- H_1 : vektory středních hodnot se nerovnají

Testová statistika má tvar

$$\begin{aligned} T^2 &= \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^T \Sigma^{-1} (\overline{\mathbf{X}} - \overline{\mathbf{Y}}) \\ \Sigma &= \frac{(n_1 - 1)\Sigma_1 + (n_2 - 1)\Sigma_2}{n_1 + n_2 - 2} \end{aligned}$$

Testová statistika má za platnosti H_0 Hotellingovo T^2-rozdělení s k a $n_1 + n_2 - 2$ stupni volnosti. Toto lze převést na F-rozdělení.

Obdobně lze zkonstruovat i testovou statistiku pro jednovýběrový test.

MANOVA

Při srovnání více nezávislých výběrů se opět testují hypotézy

- H_0 : vektory středních hodnot se rovnají
- H_1 : vektory středních hodnot se nerovnají

Stejně jako u jednorozměrné analýzy rozptylu, i ve vícerozměrné verzi je vyhodnocení hypotéz založeno na porovnání variability vysvětlené a nevysvětlené. Existuje několik testových statistik, kde všechny pracují s maticemi

$$\mathbf{W} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_i)^T (\mathbf{Y}_{ij} - \overline{\mathbf{Y}}_i)$$

$$\mathbf{B} = \sum_{i=1}^{p} n_i (\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^T (\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})$$

kde p značí počet výběrů a $\overline{\mathbf{Y}}_i$ průměr i-tého výběru.

MANOVA

Testové statistiky pro MANOVu.

- **Wilkovo lambda**

$$\Lambda_W = \det\left(\frac{\mathbf{W}}{\mathbf{W}+\mathbf{B}}\right)$$

- **Pillayova stopa**

$$\Lambda_P = \mathbf{tr}\left(\frac{\mathbf{B}}{\mathbf{W}+\mathbf{B}}\right)$$

- **Hotellingovo lambda**

$$\Lambda_H = \mathbf{tr}\left(\frac{\mathbf{B}}{\mathbf{W}}\right)$$

při porovnání dvou výběrů se všechny tyto statistiky smrští na Hotellingův dvouvýběrový test.

Kanonické korelace

Máme dvě skupiny proměnných **X** a **Y** měřených na stejných jedincích a chceme zjistit, zda mezi těmito skupinami je nějaký vztah, případně jaký.

Příklad. *Uvažujme dvě různé skupiny lékařských vyšetření a hodnotíme, zda obě tyto skupiny měří to samé, nebo ne.*

Kanonickou korelaci získáme jako maximální korelaci mezi dvěma proměnnými reprezentujícími skupiny **X** a **Y**. Tyto proměnné se nazývají **kanonické proměnné**.

Kanonické korelace

Hledání párů **kanonických proměnných**

- kanonická proměnná je lineární kombinace proměnných ve skupině
- získáme k párů kanonických proměnných, kde k je počet proměnných v menší skupině
- první pár kanonických proměnných

$$K_{11} = \mathbf{a}^T\mathbf{X}, K_{21} = \mathbf{b}^T\mathbf{Y}$$

pro něž platí

$$\mathrm{cor}(K_{11}, K_{21}) = \max\{\mathrm{cor}(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y})\}$$

- druhý pár kanonických proměnných je kolmý k prvnímu a opět je pro něj korelace maximální,
- kanonické korelace jsou korelace mezi páry kanonických proměnných

Metoda hlavních komponent (PCA)

Snížení počtu proměnných v mnohorozměrném prostoru

- v mnohorozměrném prostoru bývají proměnné vzájemně korelované
- tyto proměnné dávají podobnou / stejnou informaci
- proměnné podle podobnosti sloučíme do skupin a každou reprezentujeme jednou proměnnou
- použijeme jen malý počet nových proměnných s velkým množstvím informace

Metoda hlavních komponent (PCA)

Transformace původních proměnných do nových

$$\mathbf{Y} = \mathbf{X}^T\mathbf{P}$$

kde

- **X** je centrovavá matice vstupních hodnot (centrování = odečet průměru),
- **Y** je výstupní - cílová matice
- **P** je matice transformačních vektorů. Matici **P** získáme pomocí rozkladu korelační matice vstupních dat **C**

$$\mathbf{C} = \mathbf{P}\Lambda\mathbf{P}^T$$

- Λ je matice vlastních čísel korelační matice **C**
- matice **P** obsahuje vlastní vektory korelační matice **C**.

Metoda hlavních komponent (PCA)

Výsledná matice hlavních komponent **Y** má následující vlastnosti

- její vektory jsou vzájemně kolmé (nezávislé)
- řadí se podle variability: od vektoru s největší variabilitou k vektoru s nejnižší variabilitou
- obsahuje veškerou informaci, kterou obsahovala původní data

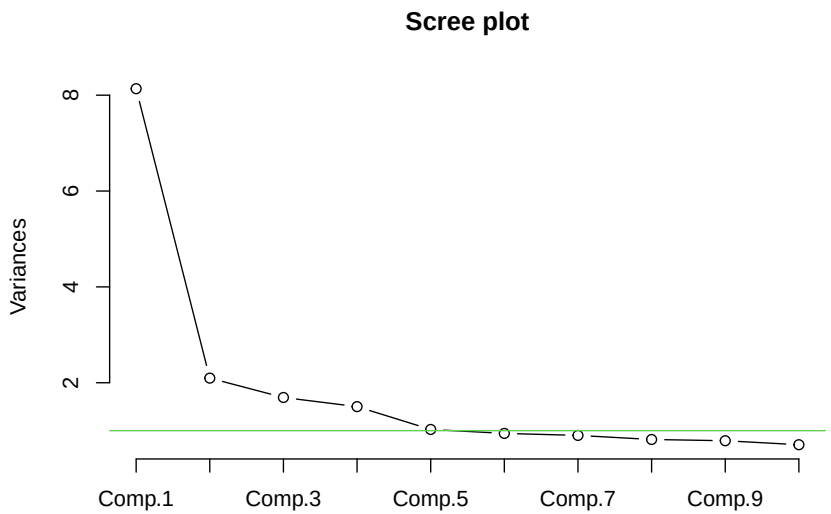
Metoda hlavních komponent (PCA)

Celý postup si můžeme představit následovně

- představíme si mnohozměrná data v prostoru
- daty proložíme vektor ve směru s největší variabilitou
- tak získáme první hlavní komponentu (PC)
- hledáme vektor, který by byl k prvnímu kolmý a opět byl ve směru s největší variabilitou
- získáme druhou hlavní komponentu
- hledáme vektor, který by byl kolmý k prvním dvěma a byl ve směru s největší variabilitou
- získáme třetí hlavní komponentu
- poslední dva kroky opakujeme, dokud máme body ve volném prostoru

Metoda hlavních komponent (PCA)

Vstupní data poté reprezentujeme menším množstvím nových proměnných (hlavních komponent) tak, abychom ztratili co nejméně informace / variability. Jejich optimální počet je počet vlastních čísel větších než 1. Graficky znázorněno pomocí tzv. "*Scree plot*".



Graf zobrazující hodnoty pro prvních 10 hlavních komponent získaných z původních 24 proměnných. Optimální počet hlavních komponent je 5.

Faktorová analýza

Nevýhodou hlavních komponent je, že nemají přirozenou interpretaci. Pokud tedy chceme získat menší počet proměnných, které jsou interpretovatelné, používá se **faktorová analýza**.

Hlavní myšlenka faktorové analýzy pochází z psychologie:

- na každého působí k neměřitelných faktorů
- podle toho, jak na nás působí, my reagujeme
- podle reakcí na p podnětů se snažíme identifikovat původní faktory

Faktorová analýza

Metoda vychází z metody hlavních komponent. Identifikujeme k hlavích komponent, a ty pak "rotujeme", dokud nedostanou nějakou přirozenou interpretaci. K rotaci je možné použít několik metod, nejčastěji se používá **varimax**.

Příklad. *Děti nosí ze školy vysvědčení. Podle známek, pak lze identifikovat dvě skupiny studentů, jedna z nich má dobré známky v předmětech matematika, fyzika, přírodopis, zeměpis, chemie, druhá má dobré známky v předmětech čeština, angličtina, dějepis, občanská výchova. Faktory, které na ně působí jsou pak přírodní vědy a humanitní obory.*

Faktorová analýza

Vycházíme z rovnice obdobné jako u analýzy hlavních komponent

$$\mathbf{X} = \mathbf{LF} + \varepsilon$$

kde

- $\mathbf{X}$ je centrovaná matice naměřených dat
- $\mathbf{L}$ jsou tzv. *loadings*
- $\mathbf{F}$ jsou hledané faktory
- ε jsou náhodné chyby

Předpoklady faktorové analýzy

Aby bylo možné faktory odhadnout, musí platit

- $\mathbf{F}$ a ε jsou nezávislé
- $\mathrm{E}(\mathbf{F}) = 0$ a $\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$, kde $\mathbf{I}$ je jednotková matice, tj. faktory mají nulovou střední hodnotu, jednotkový rozptyl a jsou nezávislé
- $\mathrm{E}(\varepsilon) = 0$ a $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}$, tj. náhodné chyby jsou nezávislé, stejně rozdělené s nulovou střednín hodnotou a konstantním rozptylem σ^2
- $\mathrm{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \sigma^2\mathbf{I}$, tedy
  - $\mathrm{Var}(\mathrm{X_i}) = \ell_{\mathrm{i1}}^2 + \ell_{\mathrm{i2}}^2 + \ldots + \ell_{\mathrm{im}}^2 + \sigma^2$
  - $\mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j}) = \ell_{\mathrm{i1}}\ell_{\mathrm{j1}} + \ell_{\mathrm{i2}}\ell_{\mathrm{j2}} + \ldots + \ell_{\mathrm{im}}\ell_{\mathrm{jm}}$
- $\mathrm{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$, tedy $\mathrm{Cov}(\mathrm{X_i}, \mathrm{F_j}) = \ell_{\mathrm{ij}}$ kde ℓ_{ij} jsou prvky matice $\mathbf{L}$

Diskriminační analýza

Máme mnohorozměrná data z několika různých, známých populací a chceme najít nejlepší možný způsob, jak podle dostupných informací rozlišit skupiny mezi sebou.

Příklad. *Uvažujme pacienty s různými nemocemi a mějme ke každému skupinu lékařských testů. Chceme pak najít způsob, jak zařadit pacienta do do správné skupiny (určit správně nemoc, kterou má) jen na základě výsledků testů*

Diskriminační analýza

Intuitivní postup jak pro pacienta určit správnou skupinu

- pro každou skupinu spočítáme průměrný vektor za všechny proměnné (lékařské testy)
- nového pacienta zařadíme do skupiny, která bude mít průměrný vektor nejblíže k pacientovým výsledkům

Jak dobré je určené rozhodovací pravidlo?

- data rozdělíme na 2 části: trénovací a testovací
- na trénovací části se naučíme diskriminační pravidlo
- na testovací části ho vyzkoušíme
- určíme, v kolika procentech případů jsme se na testovací sadě trefili (**klasifikace**)

Diskriminační analýza

Výběr mezi dvěma skupinami

- uvažujme pouze dvě populace $\mathbf{X}_1$ a $\mathbf{X}_2$ s průměry $\overline{\mathbf{X}}_{1,n}, \overline{\mathbf{X}}_{2,n}$.
- nové pozorování $\mathbf{X}$ chceme přiřadit k jedné z těchto populací
- vydálenost $\mathbf{X}$ od průměrných vektorů měříme Mahalanobisovou vzdáleností

$$D(\mathbf{X}, \overline{\mathbf{Y}}) = \sqrt{(\mathbf{X} - \overline{\mathbf{X}})^T \mathbf{V}^{-1} (\mathbf{X} - \overline{\mathbf{X}})}$$

kde $\mathbf{V}$ značí variační matici dat

- platí-li

$$D^2(\mathbf{X}, \overline{\mathbf{X}}_{1,n}) < D^2(\mathbf{X}, \overline{\mathbf{X}}_{2,n}),$$

přiřadíme pozorování k první populaci, v opačném případě ke druhé

Diskriminační analýza

- aritmetickými operacemi lze získat vektor

$$\mathbf{b} = \mathbf{S}^{-1}(\overline{\mathbf{X}}_{1,n} - \overline{\mathbf{X}}_{2,n}),$$

- rozhodovací pravidlo pak je říká, že pokud

$$\mathbf{b}^T\mathbf{X} - \mathbf{b}^T\frac{\overline{\mathbf{X}}_{1,n} + \overline{\mathbf{X}}_{2,n}}{2} = \sum_{i=1}^{k} b_i X_i - b_0 > 0$$

  pak pozorování patří do první populace
- přidání **apriorních pravděpodobností**/ velikostí skupin (např. relativní četnosti nemocí v populaci)
- označme π_1 a π_2 apriorní psti skupin, pak

$$\mathbf{b}^T\mathbf{X} - \mathbf{b}^T\frac{\overline{\mathbf{X}}_{1,n} + \overline{\mathbf{X}}_{2,n}}{2} + \ln\frac{\pi_1}{\pi_2} > 0.$$

Shluková analýza – hierarchické metody

Hledáme v datech předem neznámé skupiny tak, aby

- rozdíly mezi skupinami byly co možná největší,
- v rámci skupiny, aby byly hodnoty co nejpodobnější,

Shlukování je založené na měření vzdáleností.

Hierarchické shlukování

- postupné shlukování od jednotek po jednu velkou skupinu
- nejprve každá jednotka tvoří samostatnou skupinu
- spojí se vždy dvě nejbližší skupiny

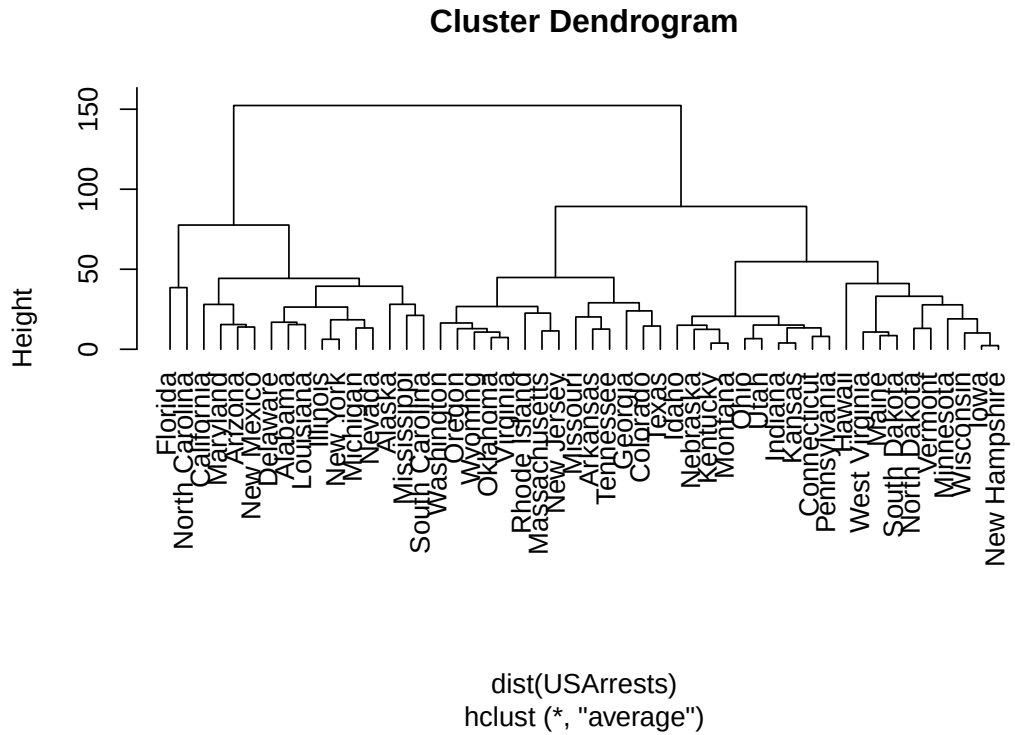
Shluková analýza – hierarchické metody

Hierarchické shlukování využívá euklidovskou vzdálenost
Existují různé typy hierarchického shlukování

- vzdálenost středů (průměrů) – **average linkage**
- vzdálenost nejbližších bodů – **single linkage**
- vzdálenost nejvzdálenějších bodů – **complete linkage**
- minimalizace variability v rámci skupin – **Ward linkage**

Shluková analýza – hierarchické metody

Graficky se proces shlukování znázorňuje pomocí **dendrogramu**.



Opticky hledáme, kde ukončit shlukování, tj. kde je k dalšímu sloučení skupin velká vzdálenost.

Shluková analýza – K-means

Nevýhodou hierarchické metody je, že odlehlé hodnoty v ní často tvoří samostatné skupiny. Alternativou je použít tzv. **K-means** shlukování. Postup je následující

- zvolíme počet skupin p
- náhodně vybereme p bodů v mnohorozměrném prostoru jako středy těchto skupin
- zařadíme prvek ke skupině s nejbližším středem
- středy se přepočítají
- poslední dva body se opakují, dokud nejsou rozřazeny všechny prvky

Nevýhodou tohoto postupu je, že pokud v datech nejsou jednoznačné skupiny, pak rozřazování dopadne jinak při jiné volbě náhodných středů.