Pokročilé statistické metody pro biology

Alena Černíková

alena.cernikova@ujep.cz

16. prosince 2024

Podmínky zápočtu

- **tři domácí úkoly**
  jednoduché opakování příkladů ze cvičení
  odevzdávat na univerzitní OneDrive – bude upřesněno později
  důraz je kladen na interpretaci výsledků
- **seminární práce**
  zpracování několika proměnných
  od zadání *výzkumu* až po interpretaci

Obsah kurzu

1. Opakování – popisné statistiky
2. Opakování – Pravděpodobnostní rozdělení
3. Opakování – bodové a intervalové odhady
4. Testování hypotéz
5. Regresní modely
6. Mnohorozměrné statistické metody

Základní pojmy

- **Nahodná veličina** – jakákoliv veličina, kterou měříme, 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, jde o nezávislé, stejně rozdělené náhodné veličiny, např. výběr 200 lidí
- **Statistická jednotka** – objekt, na kterém měříme, např. člověk
- **Populační charakteristika** – charakteristika popisující populaci, např. populační průměr
- **Výberová charakteristika** – charakteristika spočítaná na výběru, pomocí níž odhadujeme populační ekvivalent, např. výběrový průměr.

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.
  - a pro malé počty kategorií i absolutní a relativní četnosti

Popisné statistiky polohy

- **průměr** – $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, kde n je počet pozorování a $X_1, X_2, X_3, \ldots, X_n$ jsou jednotlivá měření
- **medián** – hodnota prostřední podle velikosti, nebo průměr prostředních dvou hodnot
- vybrané percentily, především **kvartily** – hodnoty v jedné a ve třech čtvrtinách podle velikosti

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

Popisné statistiky variability

- **Rozptyl** – $\mathrm{Var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$
- **Směrodatná odchylka** – $\mathrm{sd}(X) = \sqrt{\mathrm{Var}X}$
- **Mezikvartilové rozpětí** – $IQR(X) = Q_3 - Q_1$, kde Q_3 je třetí kvartil a Q_1 je první kvartil
- **Variační koeficient** – $V(X) = \frac{\mathrm{sd}(X)}{\overline{X}}$
- **Rozpětí** – $\max(X) - \min(X)$
- **Střední absolutní odchylka** – $\mathrm{MAE}(X) = \frac{\sum_{i=1}^{n}|X_i - \widetilde{X}|}{n}$

Popisné statistiky tvaru rozdělení

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

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

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

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

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

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

Popisné statistiky pro nominální proměnnou

- **absolutní četnosti** – kolik hodnot se naměřilo v dané kategorii, n_i
- **relativní četnosti** – udávají se buď v desetinných číslech nebo v procentech $p_i = n_i/n$
- **grafické znázornění** – sloupcový a koláčový graf

Náhodná veličina

Každá náhodná veličina je definována hodnotami, jichž může nabývat, a jejich pravděpodobnostmi.

Základní pravidla pro počítání s pravděpodobnostmi

- pravděpodobnost jevu nemožného $P(\emptyset) = 0$
- pravděpodobnost jevu jistého $P(\Omega) = 1$
- pravděpodobnost jevu opačného $P(A^c) = 1 - P(A)$
- pravděpodobnost sjednocení
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- podmíněná pravděpodobnost $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Senzitivita a specificita testu

Z podmíněné pravděpodobnosti vychází i definice charakteristik medicínských testů

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

Nezávislost jevů a podmíněná pravděpodobnost

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** – $P(\text{test je negativní}|\text{osoba je zdravá}) = 1750/1950 = 0.897$

Bayesova věta

Bayesova věta udává pravidla pro počítání s podmíněnými pravděpodobnostmi. Říká, jak vypočítat $P(A|B)$ ze znalosti $P(B|A)$

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

Příklad. *Jsem nemocný, když mám pozitivní test? Pomocí Bayesovy věty*

$$\begin{aligned} P(Nemoc|PozT) &= \frac{P(Nemoc \cap PozT)}{P(PozT)} = \\ &= \frac{P(PozT|Nemoc)P(Nemoc)}{P(PozT|Nemoc)P(Nemoc) + P(PozT|Zdravi)P(Zdravi)} = \\ &= \frac{Senzitivita \times podíl\ nemocných}{Senzitivita \times podíl\ nemocných + (1 - Specificita) \times podíl\ zdravých} = \\ &= \frac{0.9 \times 0.025}{0.9 \times 0.025 + 0.102 \times 0.975} = 0.184 \end{aligned}$$

Pravděpodobnostní rozdělení

Pravděpodobnostní rozdělení dělíme podle typu proměnné na

- **Spojitá** – pro číselné proměnné, které teoreticky mohou nabývat libovolné reálné hodnoty z nějakého intervalu, př. normální, exponenciální, chí-kvadrát, . . .
- **Diskrétní** – pro kategorické, nebo i číselné proměnné s jasně oddělitelnými hodnotami,
může být i nekonečně mnoho různých hodnot
př. binomické, poissonovo, alternativní, . . .

Funkce určující rozdělení

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

Binomické rozdělení

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

Značení $Bi(n, p)$, kde

- n – počet pokusů,
- p – pravděpodobnost úspěchu

Hodnoty pravděpodobnostní funkce

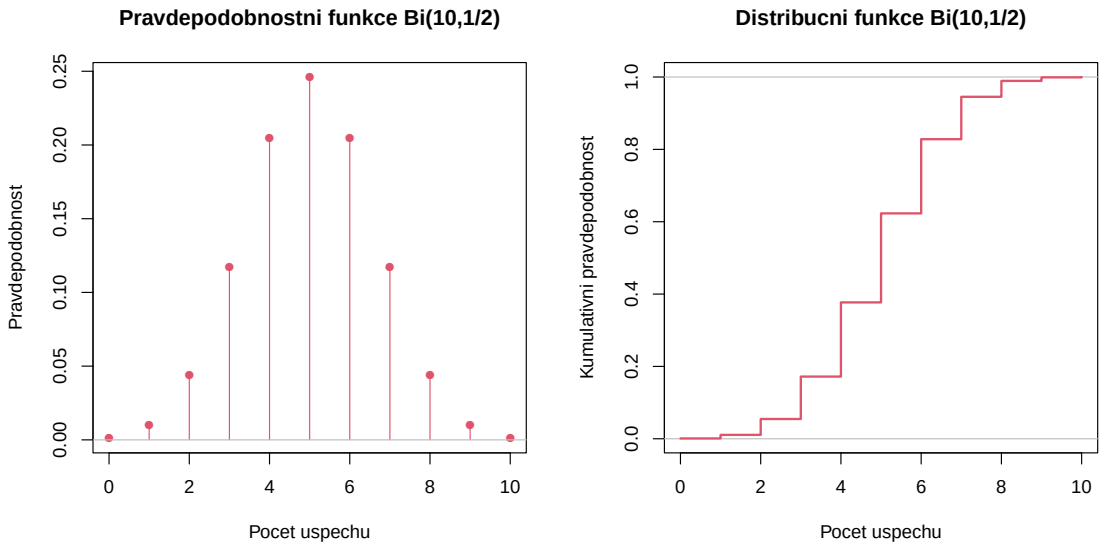
$$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Střední hodnota a rozptyl

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

Binomické rozdělení

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



Střední hodnota a rozptyl

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

Normální rozdělení

Jedná se o "hezké"rozdělení, se kterým se dobře pracuje. Toto rozdělení má výška lidí určitého věku, IQ, ... Ve statistice se nejčastěji používá standardní normální rozdělení $N(0, 1)$

Značení $N(\mu, \sigma^2)$, kde

- μ – střední hodnota
- σ^2 – rozptyl

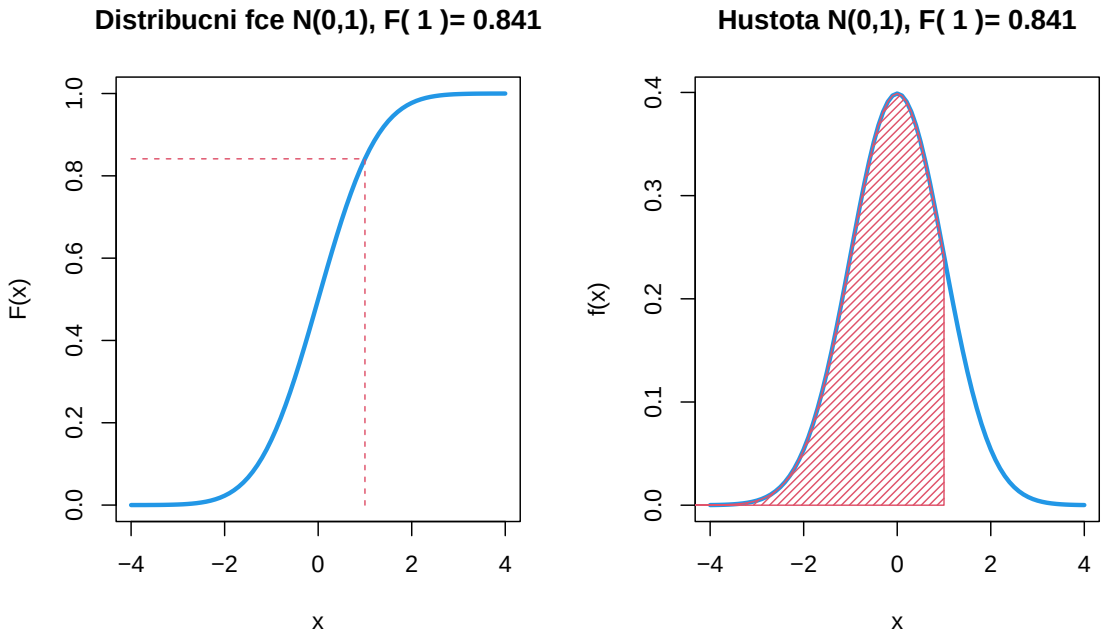
Hustota normálního rozdělení má tvar

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

Je to tak zvaná Gaussova křivka.

Normální rozdělení

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

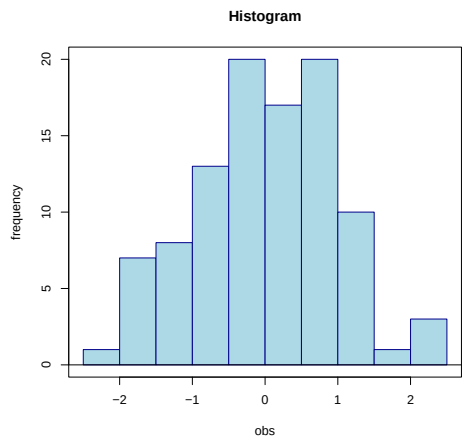


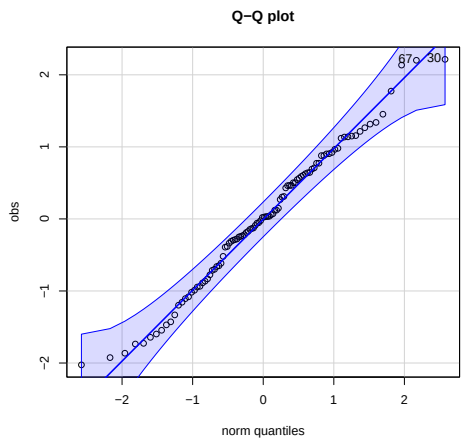
Předpokládejme binomické rozdělení $Bi(n,p)$, kde $0.1 \leq p \leq 0.9$, pak pro $n \to \infty$ toto rozdělení konverguje k normálnímu s parametry $np, np(1-p)$.

Testování normality

Každý statistický test má své předpoklady. Nejčastějším z nich je normální rozdělení dat.
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

Opakování – bodové a intervalové odhady

Statistika se zabývá odhadem/testováním teoretických/ populačních charakteristik. Označme

- X, Y náhodné veličiny
- X_i, Y_i hodnoty, jichž mohou nabývat diskrétní veličiny X, Y
- p_i, q_i pravděpodobnosti hodnot X_i, Y_i
- $f(x)$, $f(y)$ hustoty spojitých veličin X, Y
- $f(x, y)$ sdruženou hustotu veličin X, Y

Nejčastěji odhadujeme následující charakteristiky

- **Pravděpodobnost** náhodného jevu – π_i
- **Střední hodnotu** – $E(X) = \sum_{i=1}^{n} X_i p_i = \int_{-\infty}^{\infty} xf(x)dx$ s vlastnostmi:
  - $E(aX + b) = aE(X) + b$
  - $E(X + Y) = E(X) + E(Y)$

Opakování – bodové a intervalové odhady

Nejčastěji odhadujeme následující charakteristiky

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

$$\begin{aligned}
\mathrm{cor}(X, Y) &= \frac{\mathrm{cov}(X, Y)}{(\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)})} = \\
&= \frac{\sum_{i=1}^{n}(X_i - \mathrm{E}(X))(Y_i - \mathrm{E}(Y))p_i q_i}{(\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)})} = \\
&= \frac{\int_{-\infty}^{\infty}(x - \mathrm{E}(X))(y - \mathrm{E}(Y))f(x, y)dxdy}{(\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)})}
\end{aligned}$$

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

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

Testované hypotézy

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

Vyhodnocení testu

Na základě testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu – 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**
  – za dané hladiny významnosti chceme test co nejsilnější

Rozhodování ve statistickém testu

Testovat mohu porovnáním

- **testové statistiky** a **kritické hodnoty** (kvantil vybraného teoretického rozdělení)
- ***p*-hodnoty** a **hladiny významnosti** (α)

Definice ***p*-hodnoty**

- pravděpodobnost, že za platnosti H_0 nastal výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- jinak se nazývá aktuální dosažená hladina testu

Vyhodnocení testu

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

Statistický test

U testování rozlišujeme dva základní typy testů

- **Parametrické testy**
  - předpokládají normální rozdělení
  - založené na odhadu testovaného parametru
  - t-testy, klasická ANOVA, Pearsonův korelační koeficient
- **Neparametrické testy**
  - normalitu nepředpokládají
  - založené na pořadích
  - Wilcoxonův test, Kruskal-Wallisův test, Spearmanův korelační koeficient, atd.

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 t-test

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

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

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$ **Testová statistika** jednovýběrového t-testu je

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

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

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

Znaménkový test

Test o hodnotě mediánu jednoho výběru. Testujeme

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

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

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

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

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.

Znaménkový test

Příklad. *Pokračujme v příkladu s věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

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

Spočtěme rozdíly $X_i - m_0$: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17. Kladných hodnot je mezi nimi $Z = 4$. P-hodnota testu vychází 0.75, což je hodnota $> \alpha (= 0.05)$ a H_0 tedy nezamítáme. Použitím U-transformace dostaneme $U = -0.632$ a p-hodnotu 0.527.

Wilcoxonův jednovýběrový test

Znaménkový test porovnává pouze počet hodnot ležících pod mediánem a těch, co leží nad ním. Lepším testem je **Wilcoxonův test**, který je založen na pořadích. Testované hypotézy zůstávají stejné.

Postup testu

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

Pro větší n lze užít transformaci

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

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

Wilcoxonův jednovýběrový test

Příklad. *Pokračujme v příkladu s věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a opět testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

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

Spočtěme rozdíly $X_i - m_0$*: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17 a jejich absolutním hodnotám přiřaďme pořadí 1.5, 6, 1.5, 6, 6, 8, 3.5, 3.5, 9, 10. Sečtěme kladné (modré) pořadí* $S^+ = 21$ *a záporné (červené) pořadí* $S^- = 34$*. Testová statistika vychází* $U = -0.66$ *a p-hodnota 0,51* $> \alpha (= 0.05)$ *a* H_0 *tedy nezamítáme.*

Párový test

V případě, že porovnáváme dva závislé výěry, tedy taková data, která tvoří přirozené páry, používá se **párový test**. Testované hypotézy v něm jsou

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

Postup testu je takový, že v prvním kroku spočítám rozdíly mezi všemi páry

$$R_i = X_i - Y_i$$

kde X_i a Y_i jsou párová měření, a ve druhém kroku se testuje střední hodnota/ průměr tohoto rozdílu běžným **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

Pokud porovnávám dva nezávislé výběry (pozorování nemohu napárovat), pak je potřeba použít **dvouvýběrový test**.
Testujeme

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

Vybíráme jeden ze tří testů:

- Dvouvýběrový t-test pro shodné rozptyly
- dvouvýběrový t-test pro různé rozptyly item dvouvýběrový Wilcoxonův (Mann-Whitneyův) test

Test shody rozptylů ve dvou výběrech

Chceme-li rozhodnout, kterou variantu dvouvýběrového t-testu máme použít, je nutné zjistit, zda jsou v obou výběrech stejné rozptyly.
Testujeme

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

Testová statistika **F-testu pro dva rozptyly** má tvar

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

a za platnosti nulové hypotézy má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti.

Dvouvýběrový t-test

Testová statistika dvouvýběrového t-testu pro shodné rozptyly

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

kde

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

a n_1, n_2 je rozsah výběru X, respektive Y. Za platnosti nulové hypotézy má tato statistika t-rozdělení o $n_1 + n_2 - 2$ stupních volnosti.

Dvouvýběrový t-test

Testová statistika dvouvýběrového **Welchova t-testu**

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

Tato statistika má za platnosti nulové hypotézy 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.

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.

Wilcoxonův dvouvýběrový test

Postup výpočtu dvouvýběrového Wilcoxonova testu

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

Wilcoxonův dvouvýběrový test

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

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

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

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

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

Wilcoxonův dvouvýběrový test

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

- H_0 : *Studenti v Ústí a v Liberci jsou stejní*
- H_1 : *Studenti v Ústí a v Liberci se liší.*
- *V prvním kroku srovnám všechny hodnoty do řady* 41, 45, 53, 56, 62, 66, 76, 77, 79, 79, 80, 81, 84
- *následně jim přiřadím pořadí* 1, 2, 3, 4, 5, 6, 7, 8, 9.5, 9.5, 11, 12, 13
- *pak vypočtu* $T_1 = 38.5$, $T_2 = 52.5$, $U_1 = 24.5$, $U_2 = 17.5$, $U_0 = 0.5$, $p = 0.6678$

P-hodnota $> \alpha$ *a tedy nezamítám nulovou hypotézu, neprokázal se rozdíl mezi studenty v Ústí a v Liberci.*

Analýza rozptylu – ANOVA

Porovnáváme-li střední hodnotu ve více než 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.

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?*

Analýza rozptylu – 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 a k počet výběrů.
Analýza rozptylu rozkládá celkovou variabilitu

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

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

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

Analýza rozptylu – ANOVA

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

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

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

Bartlettův test

Předpokladem analýzy rozptylu je shoda rozptylů ve všech výběrech. Kontrolujeme ji **Bartlettovým testem**.

Testujeme

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

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

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

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

Testová statistika

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

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

Párové srovnání

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

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

pro všechny dvojice i a j.
Testová statistika má tvar

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

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

Welchova analýza rozptylu

Nemají-li všechny porovnávané skupiny stejné rozptyly, používá se tzv. **Welchova ANOVA**. Ta je založena na myšlence vážení skupinových průměrů vahou odpovídající jejich variabilitě. Namísto celkového průměru se pracuje s váženým průměrem

$$\overline{X}_w = \frac{\sum_{i=1}^{k} w_i \overline{X}_{i.}}{\sum_{i=1}^{k} w_i}, \text{ kde } w_i = \frac{n_i}{\mathrm{Var(X)_i}}$$

Variabilita mezi výběry se pak počítá jako

$$SSA_w = \sum_{i=1}^{k} w_i (\overline{X}_{i.} - \overline{Y}_w)^2, MSA_w = \frac{SSA_w}{k-1}$$

Welchova analýza rozptylu

Dále se zavádí parametr

$$\Lambda = \frac{3\sum_{i=1}^{k} \frac{\left(1-\frac{w_i}{\sum_{i=1}^{k} w_i}\right)^2}{n_i-1}}{k^2-1}$$

testová statistika pak má tvar

$$F_w = \frac{SSA_w/(k-1)}{1+\frac{2\Lambda(r-2)}{3}}$$

která má za platnosti H_0 F-rozdělení o $r-1$ a $1/\Lambda$ stupních volnosti.

Kruskal-Wallisův test

Pro porovnání více nezávislých výběrů, které nemají normální rozdělení se používá **Kruskal-Wallisova ANOVA**.
Testujeme

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

Srovnáme všechny naměřené hodnoty do řady, určíme jejich pořadí a spočteme statistiky $T_1, \ldots, T_k$, kde k je počet výběrů. Pak platí, že testová statistika

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

má za platnosti H_0 χ^2-rozdělení.

Dunnův test

V případě, že Kruskal-Wallisova ANOVA určí, že se výběry mezi sebou významně liší, je potřeba zjistit, ktwré konkrétní dvojice výběrů se liší. K tomu může sloužit např. **Dunnův test**.
Testová statistika porovnávající i-tý a j-tý výběr je

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

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

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

kde S_l je počet l-té shodné hodnoty.

Tato statistika má za platnosti H_0 $N(0,1)$-rozdělení. Pro vícenásobné porovnání se pak použijí upravené p-hodnoty, aby byla udržena celková hladina testu.

ANOVA pro opakovaná měření

Pokud se chystáme porovnávat několik závislých výběrů, používá se **ANOVA pro opakovaná měření**.
Příklady takovéto situace mohou být

- **Ochutnávka jogurtů:** 20 lidí ochutnává a hodnotí každý všech 5 porovnávaných vzorků jogurtu.
- **Měření opakovaná v čase:** chceme hodnotit vývoj pacientova zdravotního stavu v čase. Pro 30 pacientů děláme opakovaná měření játrových testů.

Stále se testují hypotézy

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

ANOVA pro opakovaná měření

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

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

Test je pak založen na porovnání SSA a SSz.

Friedmanův test

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

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

Za platnosti nulové hypotézy má 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. Pomocí této Z-transformace je možné porovnávat i dva korelační koeficienty mezi sebou. Platí totiž, že

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

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 zamložen 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 koeicifient, 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

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

Předpokládáme lineární model ve tvaru

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

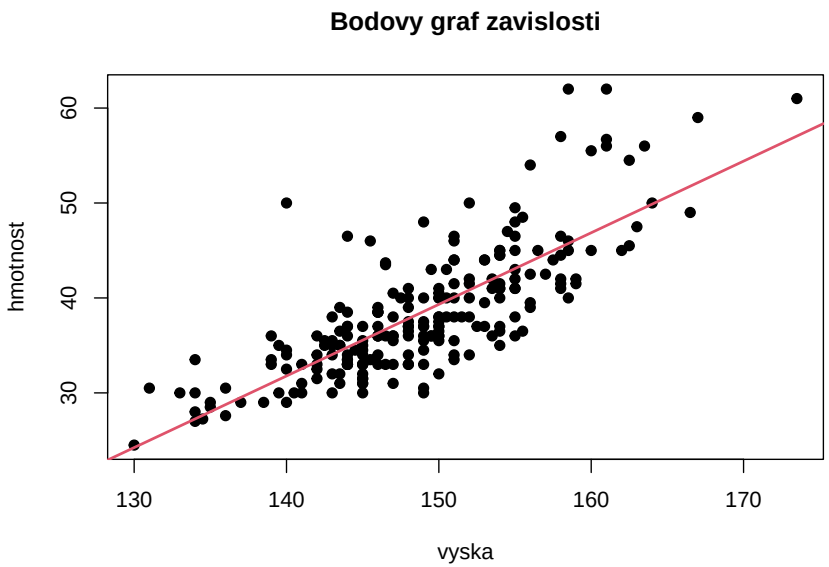
kde

- 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

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

- $\hat{Y}_i$ jsou odhady, nebo též predikce,
- b_0, b_1 jsou pak odhady regresních koeficientů
- 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

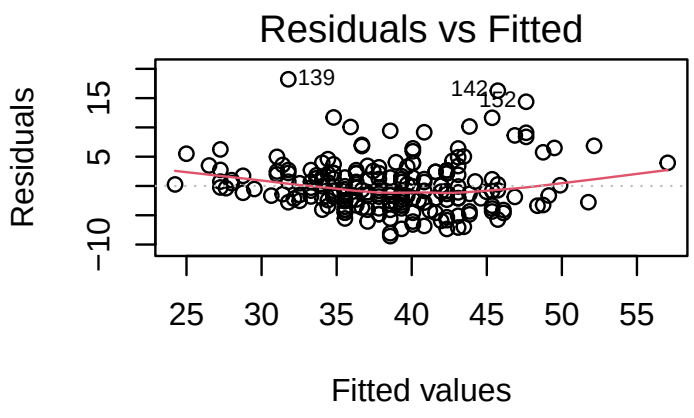
I lineární regrese má své **předpoklady**

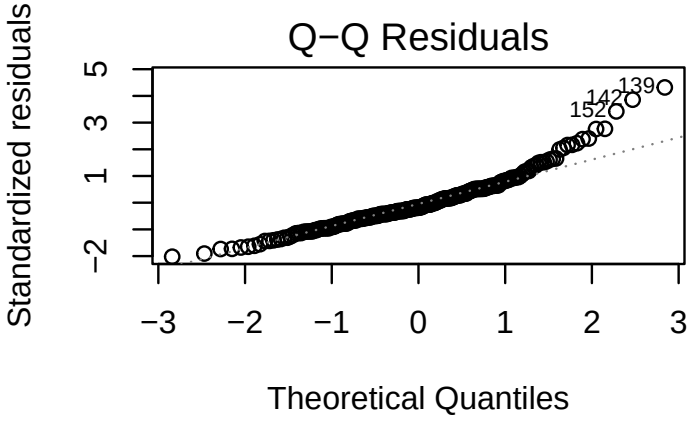
- Mezi proměnnými je skutečně lineární vztah
- Residua jsou nezávislá
- Residua mají normální rozdělení
- Stabilita rozptylu
- V datech nejsou vlivná pozorování

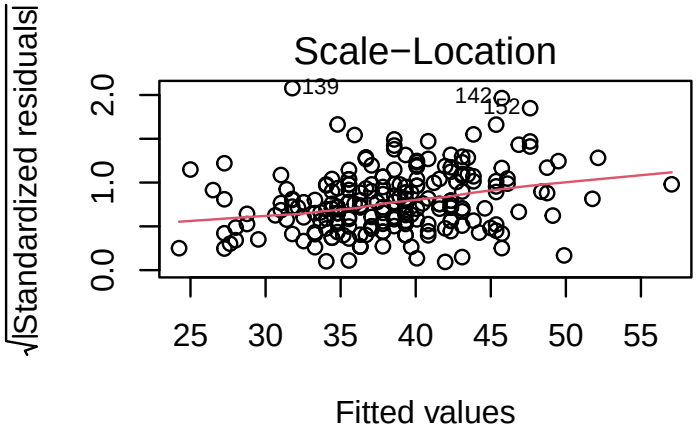
Jednotlivé předpoklady můžeme hodnotit buď na základě znalosti dat (nezávislost), nebo grafickými případně číselnými testy.

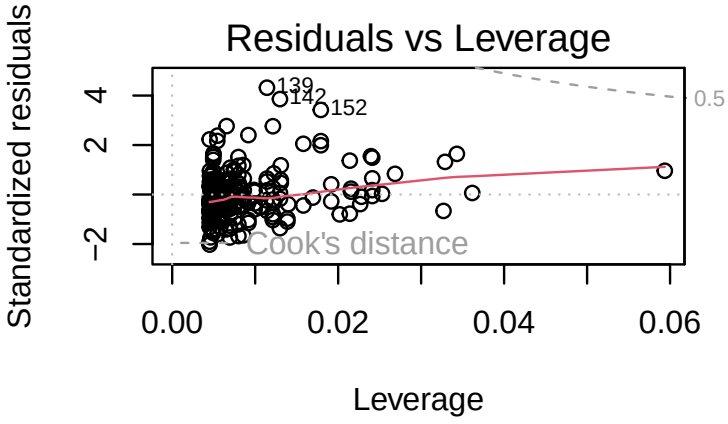
Lineární regrese

Ukázka grafických testů předpokladů









Lineární regrese

Ukázka grafických testů předpokladů

- **1. graf**: lineární vztah – červená čára nemá mít trend
- **2. graf**: normalita residuí – body mají ležet na přímce
- **3. graf**: stabilita rozptylu – červená čára nemá mít trend
- **4. graf**: body nemají překročit meze (čárkované křivky)

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

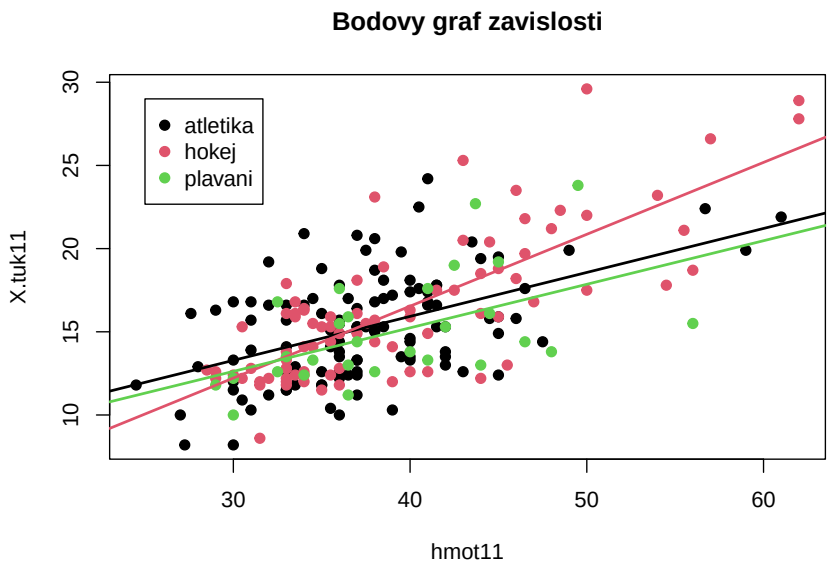
$$\begin{aligned} X_i &= 1 \ldots \text{nastala i-ta kategorie} \\ &= 0 \ldots \text{jinak} \end{aligned}$$

- 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 nejseou, 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 **ortogonální** – 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

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

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.*
Hledání **kanonických proměnných**

- první pár kanonických proměnných

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

- druhý pár kanonických proměnných je kolmý k prvnímu a opět je pro něj korelace maximální,
- získáme k kanonických proměnných, kde k je počet proměnných v menší skupině
- kanonické korelace jsou korelace mezi páry kaninických proměnných

Diskriminační analýza

Máme mnohorozměrná data z několika různých populací a chceme najít nejlepší možný způsob, jak na základě dat rozlišit skupiny mezi sebou. Hledáme postup, jak určit skupinu na základě dat.

Příklad. *Uvažujme pacienty s různými nemocemi a mějme ke každému skupinu lékařských testů. Chceme pak najít způsob, jak zařadit pacienta do skupiny jen na základě výsledků testů*

Postup

- pro každou skupinu spočítáme průměrný vektor
- nového pacienta zařadíme do skupiny, která bude mít svůj průměrný vektor nejblíže k pacientovým výsledkům

Jak dobré je určené rozhodovací pravidlo zjistíme na základě klasifikace, tj. zjištění, kolik jednotek jsme přiřadili správně a kolik chybně.

Diskriminační analýza

Uvažujme pouze dvě populace s průměry $\overline{\mathbf{X}}_{1,n}, \overline{\mathbf{X}}_{2,n}$. Vzdálenosti od těchto průměrů měříme Mahalanobisovou vzdáleností

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

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.
Aritmetickými operacemi lze získat vektor

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

a rozhodovací pravidlo, 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.
K tomuto pravidlu mohu přidat ještě apriorní pravděpodobnosti (třeba relativní četnosti nemocí v populaci.)

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

Shluková analýza – hierarchické metody

Hledáme ve datech předem neznámé skupiny tak, aby

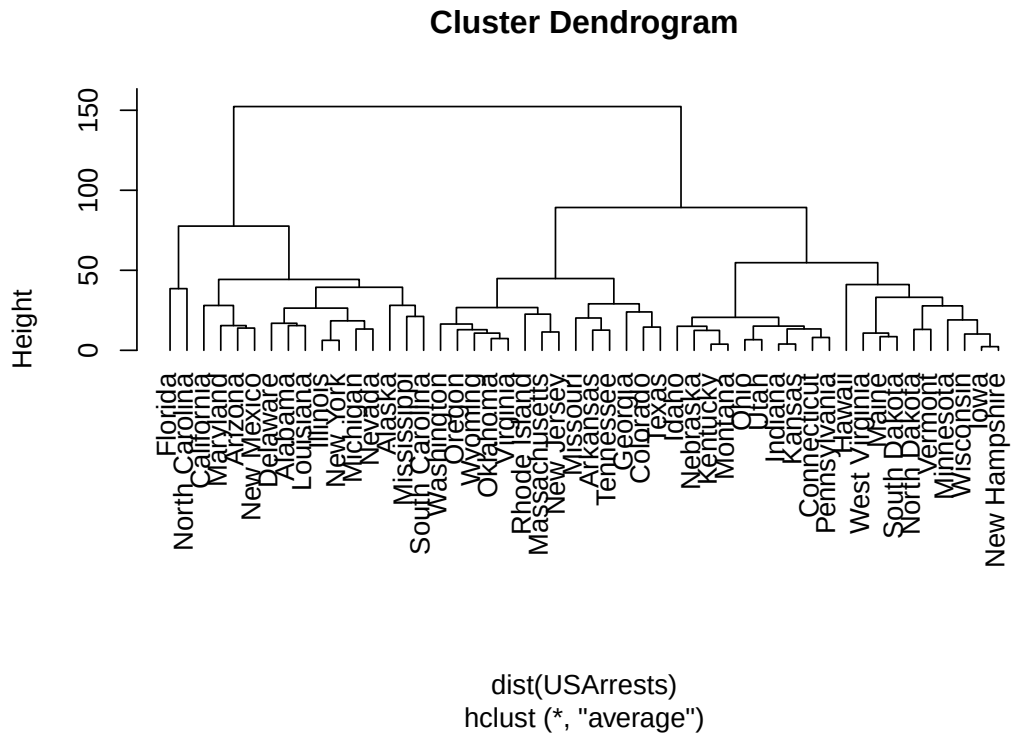
- rozdíly mezi skupinami byly co možná největší,
- v rámci skupiny, aby byly hodnoty co nejpodobnější,

Hierarchické shlukování měří vzdálenosti mezi jednotlivými pozorováními např. euklidovskou vzdáleností a shlukuje k sobě jednotky, co jsou si nejblíže. Spojování skupin

- 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

V této analýze nejprve považujeme každé jedno pozorování za samostatnou skupinu a postupně tyto skupiny spojujeme. Graficky se tento proces znázorňuje pomocí **dendrogramu**.



Opticky hledáme, kde ukončit shlukování, tj. kolik skupin je optimum.

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.

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

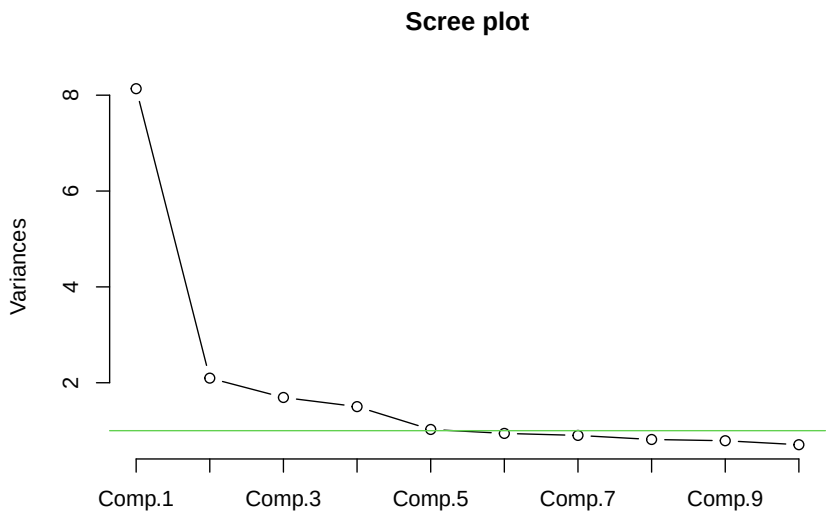
Metoda hlavních komponent (PCA)

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

Vycházíme z rovnice obdobné jako u analýzy hlavních komponent

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

kde $\mathbf{X}$ je centrovaná matice naměřených dat, $\mathbf{L}$ jsou tzv. *loadings* a $\mathbf{F}$ jsou hledané 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.*