Pokročilé statistické metody

Alena Černíková

alena.cernikova@ujep.cz

28. května 2024

Realizace výuky

- **Výuku** realizují 3 vyučující
  - Alena Černíková
  - doc. Viktor Maškov
  - prof. Sergii Babichev
- **Zkoušku** realizuje jeden vyučující
  - Alena Černíková

Podmínky zápočtu a zkoušky

- **Zápočet**
  - dva domácí úkoly od Černíkové
  - jeden domácí úkol od prof. Babicheva
  - seminární práce – zaštiťuje Černíková
    – zadání bude v kombinaci s prof. Babichevem
- **Zkouška** – ústní u Černíkové
  - nejspíš bude sestávat z
  - tři příklady u počítače
  - jedna teoretická otázka

Obsah kurzu

- Teorie testování hypotéz – AČ
- Věcná významnost a metaanalýza – AČ
- Mnohonásobná lineární regrese – SB
- Zobecněné lineární modely –SB
- Nelineární modely – SB
- Mnohorozměrná statistika – úvod – AČ
- Metoda hlavních komponent, faktorová analýza – AČ
- Shluková analýza – AČ
- Fuzzy logika a fuzzy modelování – VM & SB
- Bayesovské metody – VM & SB

Základy testování hypotéz

Testování stanoveného tvrzení. Např.

- Nový lék je lepší než stávající.
- Náhodná veličina má normální rozdělení.
- Průměrná výška lidí se za posledních 50 let zvýšila.
- Výnosy z jednotlivých druhů jabloní se liší.
- Krevní tlak závisí na hmotnosti.

Testy mohou být jak grafické, tak číselné. Nyní se budeme zabývat číselnými testy.
Vždy se testují **populační charakteristiky**. Jejich výběrové ekvivalenty se používají jen pro sestrojení testových kritérií.

Základy testování hypotéz

Při statistickém rozhodování testujeme proti sobě 2 hypotézy

- Nulovou hypotézu – značíme H_0
  - obsahuje vždy jen jednu možnost
  - v případě testu nezávislosti sem patří **NEZÁVISLOST**
  - př. nový lék je stejný jako ten stávající, výnosy druhů jabloní jsou stejné.
- Alternativní hypotézu – značíme H_1
  - obsahuje více možností (např. interval)
  - patří sem to, co chci prokázat
  - v případě testu nezávislosti sem patří **ZÁVISLOST**
  - př. nový lék je lepší než ten stávající, výnosy druhů jabloní se liší

Základy testování hypotéz

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

- Zamítneme nulovou hypotézu
  – tím jsme prokázali platnost alternativy
- Nezamítneme nulovou hypotézu
  – tím jsme neprokázali nic – interpretace závisí na formulaci testovaných hypotéz

Jiný závěr udělat nemohu!
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 platí H_1
  – značí se β a hodnota $1 - \beta$ se nazývá **síla testu**
  – za dané hladiny významnosti chceme test co nejsilnější

Základy testování hypotéz

	Nezamítáme H_0	Zamítáme H_0
Skutečně platí H_0	OK	Chyba I. druhu α
Skutečně platí H_1	Chyba II. druhu β	OK síla testu

Základy testování hypotéz

Podle toho, co testujeme a podle typu dat vybereme vhodný statistický test, kterým budeme o platnosti testovaných hypotéz rozhodovat. Rozhodnutí můžeme udělat buď na základě

- porovnání **testové statistiky** (T) a kritické hodnoty (c, jsou tabelovány)
- porovnání p-**hodnoty** a hladiny významnosti (α)

Platí, že

- absolutní hodnota testové statistiky $|T| \geq c$ nebo p-hodnota $\leq \alpha$ potom ZAMÍTÁME H_0
- absolutní hodnota testové statistiky $|T| < c$ nebo p-hodnota $> \alpha$ potom NEZAMÍTÁME H_0

S testovou statistikou se většinou pracuje při ručním výpočtu.
Statistické softwary vrací jako výsledek testu ***p*-hodnotu**.

Co je *p*-hodnota

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti H_0 nastal výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
- definice *p*-hodnoty se týká testové statistiky

(Ne)zamítnout H_0 nestačí, tento výsledek je třeba interpretovat vzhledem k položené otázce.

Vybrané testy

Testy dělíme podle toho, co testují

- **Testy rozdělení**
  – nejčastěji testujeme normalitu
  – př. Shapiro-Wilkův test, Andersonův-Darlingův, Kolmogorovův-Smirnovův, χ^2-test dobré shody atd.
- **Testy o hodnotě parametru**
  – nejčastěji testujeme tvrzení o střední hodnotě
  – př. dvouvýběrový t-test, ANOVA, test o hodnotě korelačního koeficientu, Waldův test, Bartlettův test atd.
- **Testy nezávislosti**
  – většinou se jedná o testy vybraných parametrů
  – test o hodnotě korelačního koeficientu, test o hodnotě regresního koeficientu, ANOVA, χ^2-test nezávislosti atd.

Dále testy dělíme podle typu dat

- **Parametrické testy**
  – testy o hodnotě parametru, nejčastěji o střední hodnotě
  – určené pro data, která mají přibližně normální rozdělení
  – př. t-testy, klasická analýza rozptylu, test o hodnotě Pearsonova korelačního koeficientu atd.
- **Neparametrické testy**
  – testy založené na pořadích, nepracují s odhadem parametru
  – odpadá požadavek na normalitu dat, ale i zde jsou určité předpoklady
  – př. Wilcoxonovy testy, Kruskal-Wallisův test, test o hodnotě korelačního koeficientu atd.
- **Testy pro kategorická data**
  – většinou χ^2-testy, i ty mají své předpoklady
  – př. χ^2-test nezávislosti, Fisherův test, test o hodnotě Kendallova korelačního koeficientu atd.

Schéma některých často používaných testů

Číselné proměnné

- Test o střední hodnotě jednoho výběru
  – normální data – jednovýběrový t-test
  – nenormální data – znaménkový test, jednovýběrový Wilcoxonův test
- Test o stř. hodnotě rozdílu dvou závislých výběrů
  – normální data – párový t-test
  – nenormální data – párový Wilcoxonův test
- Test o stř. hodnotě rozdílu dvou nezávislých výběrů
  – normální rozdělení, shodné rozptyly – dvouvýběrový t-test pro shodné rozptyly
  – normální rozdělení, různé rozpytyly – dvouvýběrový Welchův test (t-test pro různé rozptyly)
  – nenormální rozdělení – dvouvýběrový Wilcoxonův test
- Porovnání stř. hodnot více závislých výběrů
  – normální data – ANOVA pro opakovaná měření
  – nenormální data – Friedmanův test
- Porovnání stř. hodnot více nezávislých výběrů
  – normální rozdělení, shodné rozptyly – klasická ANOVA pro shodné rozptyly
  – normální rozdělení, různé rozpytyly – klasická ANOVA pro různé rozptyly
  – nenormální rozdělení – Kruskall-Wallisův test

Schéma některých často používaných testů

Vztah dvou proměnných

- **Číselná vs. kategorická proměnná**
  – používají se dvouvýběrové testy nebo analýza rozptylu (ANOVA)
- **Dvě číselné proměnné** – Korelační koeficient – spojité obě normálně rozdělené proměnné – Pearsonův korelační koeficient
  – spojité proměnné, alespoň jedna nenormálně rozdělena – Spearmanův korelační koeficient
  – kategorické uspořádané proměnné – Kendallův korelační koeficient
- **Dvě kategorické proměnné**
  – χ^2-kvadrát test, Fisherův test, poměr šancí

Porovnání **rozptylů** ve výběrech

- Dva výběry – F-test pro 2 rozptyly
- Více výběrů – Bartlettův test, Levenův test

Testy používané v regresních modelech budeme řešit později.

Jednovýběrový t-test

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

Testované hypotézy

- Nulová hypotéza H_0 : střední hodnota $= \mu_0$
- Alternativní hypotéza H_1 : jedna ze tří možností
  – střední hodnota $\neq \mu_0$
  – střední hodnota $< \mu_0$
  – střední hodnota $> \mu_0$

Není-li řečeno jinak, testujeme na hladině významnosti $\alpha = 0.05$.

Jednovýběrový t-test

Testová statistika jednovýběrového t-testu je

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

a za platnosti nulové hypotézy má tato statistika *t*-rozdělení o $n-1$ stupních volnosti.
Testovou statistiku T porovnáváme s kritickými hodnotami *t*-rozdělení (tj. kvantily), na základě čehož buď můžeme přímo rozhodnot o zamítutí nebo nezamítnutí nulové hypotézy, nebo můžeme spočítat *p*-hodnotu a test vyhodnocovat na základě ní.

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

Souvislost s intervalem spolehlivosti

Víme

$$P(|T| < t_{n-1}(1-\alpha/2)) = 1-\alpha$$

Tedy

$$|T| = \frac{|\overline{X}-\mu|}{\mathrm{sd(X)}}\sqrt{n} < t_{n-1}(1-\alpha/2)$$

Checme interval, kde se nachází μ – skutečná střední hodnota

$$|\overline{X}-\mu| < t_{n-1}(1-\alpha/2))\frac{\mathrm{sd(X)}}{\sqrt{n}}$$

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

Když testovaná hodnota je uvnitř intervalu spolehlivosti, pak H_0 nezamítáme, když je testovaná hodnota vně intervalu spolehlivosti, pak H_0 zamítáme.

Jednovýběrový t-test

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

Testované hypotézy

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

Testujeme na hladině významnosti $\alpha = 0.05$.

Pokračování příkladu.
Testová statistika vyšla

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

Tuto hodnotu porovnám s kvantilem *t*-rozdělení $t_{221}(1 - 0.05) = 1.65$. Jelikož testová statistika je v absolutní hodnotě větší než kritická hodnota, zamítám nulovou hypotézu.
P-hodnota vyšla $p = 0.005 < 0.05$, což také vede na zamítnutí nulové hypotézy.
Závěr: Prokázala jsem, že průměrná výška jedenáctiletých dětí je menší než 150 cm.

Wilcoxonův jednovýběrový test

V případě, že jsou v datech velké odchylky od normality, používá se neparametrický **Wilcoxonův test**, který je založen na pořadích. Tento test netestuje populační průměr, ale medián.

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. *Uvažujme naměřené věky otců 30, 28, 36, 38, 28, 26, 29, 37, 25, 50 a testujme hypotézu, že medián věku otců je 33 let, tj. testujeme*

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

Spočtěme rozdíly $X_i - m_0$*: -3, -5, 3, 5, -5, -7, -4, 4, -8, 17 a jejich absolutním hodnotám přiřaďme pořadí 1.5, 6, 1.5, 6, 6, 8, 3.5, 3.5, 9, 10. Sečtěme kladné (modré) pořadí* $S^+ = 21$ *a záporné (červené) pořadí* $S^- = 34$*. Testová statistika vychází* $U = -0.66$ *a p-hodnota 0.51* $> \alpha (= 0.05)$ *a* H_0 *tedy nezamítáme. Střední hodnota věku otců může být 33.*

Dvouvýběrový t-test

Porovnáváme-li střední hodnotu dvou **nezávislých** výběrů s normálním rozdělením (X, Y), používá se **dvouvýběrový t-test.**

Existují dva typy dvouvýběrového t-testu:

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

Testované hypotézy obou testů jsou

- H_0 : střední hodnota X – střední hodnota $Y = 0$
- H_1 : střední hodnota X – střední hodnota $Y \neq 0, < 0$ nebo > 0

F-test shody rozptylů

K tomu, abychom mohli vybrat správnou verzi testu, je potřeba otestovat shodu rozptylů v obou výběrech. Používá se **F-test shody rozptylů**. Testuje se

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

Testová statistika je

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

a za platnosti H_0 má F-rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti, kde n_1 je rozsah výběru X a n_2 je rozsah výběru Y.

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika dvouvýběrového t-testu pro shodné rozptyly má tvar

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

kde

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

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

Welchův test

Testová statistika dvouvýběrového Welchova testu pro různé rozptyly má tvar

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

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

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

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

Příklad. *Ve výběru mám 222 jedenáctiletých dětí, z toho 159 hochů a 63 dívek. Průměrná hmotnost hochů vyšla 38.1 kg a u dívek 39.1. Směrodatná odchylka pro hochy vyšla 6.7 kg a pro dívky 7.1.*
Je hmotnost jedenáctiletých dětí v průměru stejná pro hochy jako pro dívky?
Nejprve otestujeme shodu rozptylů, testová statistika vychází

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

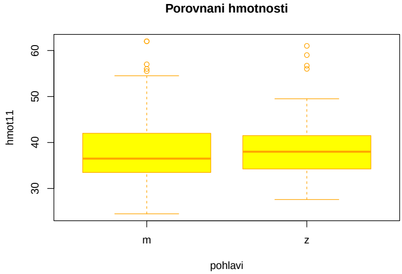
P-hodnota testu vyšla 0,56, což je více než $\alpha = 0.05$. Nulovou hypotézu tudíž nezamítáme, rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly.

Dvouvýběrový t-test

Testujeme

- H_0 : hmotnost hochů a hmotnost dívek se neliší
  hmotnost hochů – hmotnost dívek = 0
- H_1 : hmotnost hochů a dívek se liší
  hmotnost hochů – hmotnost dívek $\neq 0$

Grafické porovnání



Dvouvýběrový t-test

Testová statistika testu vychází

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

Tuto testovou statistiku porovnáváme s kvantilem t-rozdělení $t_{220}(1 - 0.025) = 1.97$ (kvantil pro oboustrannou alternativu). Jelikož testová statistika je v absolutní hodnotě menší než tento kvantil, tak nulovou hypotézu nezamítám.
P-hodnota testu vyšla 0.3151, tedy číslo větší než $\alpha = 0.05$

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

Wilcoxonův dvouvýběrový test

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

- H_0 : střední hodnota X – střední hodnota $Y = 0$
- H_1 : střední hodnota X – střední hodnota $Y \neq 0, < 0$ nebo > 0

Test je založen na pořadích hodnot sdruženého výběru. Postup

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

Wilcoxonův dvouvýběrový test

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

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

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

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

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

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 ve dvou nezávislých výběrech, používá se **analýza rozptylu**. Testujeme

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

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

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

Analýza rozptylu – ANOVA

Označme X_{ij} i-té pozorování z j-tého výběru, $\overline{X}_{i.}$ 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) SSA a variabilitu nevysvětlenou (zbytkovou, v rámci výběrů) SSe. Platí

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

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

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

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

Tabulku analýzy rozptylu si mohu nechat vypsat i pro modely lineární regrese. Hodí se u modelů vícenásobné regrese, když počítám závislost (i) na kategorických proměnných.

Bartlettův test

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

Testujeme

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

Testová statistika je založena na výběrových rozptylech v každém výběru zvlášť. Označme $\mathrm{Var(X)_i}$ 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

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

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

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

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

Testová statistika má tvar

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

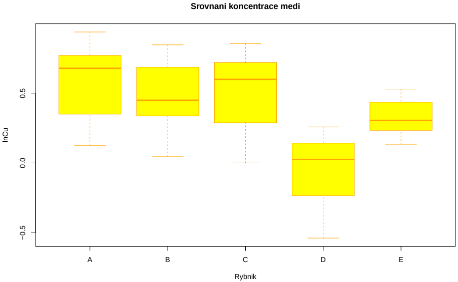
Rozdělení těchto statistik se jmenuje studentizované rozpětí a má své vlastní tabelované kritické hodnoty.

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

Testujeme

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

Grafické porovnání



Analýza rozptylu – ANOVA

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

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

Testová statistika Bartlettova testu vyšla 3.67 při čtyřech stupních volnosti, což dává p-hodnotu 0.45. Jelikož je p-hodnota větší než $\alpha = 0.05$, nulovou hypotézu nezamítáme a můžeme použít klasickou ANOVU pro shodné rozptyly.

Analýza rozptylu – ANOVA

Tabulka analýzy rozptylu

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

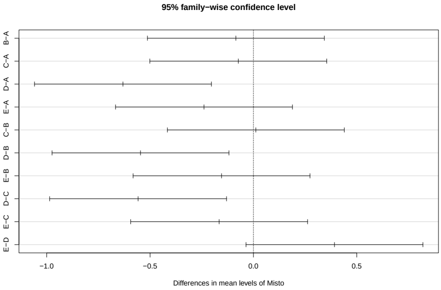
P-hodnota vyšla menší než $\alpha = 0.05$, což znamená, že nulovou hypotézu zamítáme a rybníky se mezi sebou významně liší.

Párové srovnání vrátí následující tabulku

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

Analýza rozptylu – ANOVA

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



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

Kruskal-Wallisův test

V případě, že není splněn předpoklad normality při porovnání více než dvou nezávislých výběrů, používá se **Kruskal-Wallisova ANOVA**. Kruskal-Wallisova ANOVA je přímým zobecněním Wilcoxonova dvouvýběrového testu. Testujeme

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

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

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

má za platnosti H_0 χ^2-rozdělení.

Dunnův test

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

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

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.

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.

χ^2-test nezávislosti

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

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

Kontingenční tabulka absolutních četností pak má tvar

$$
\begin{array}{c|ccc|c}
 & Y_1 & \ldots & Y_l & \\
\hline
X_1 & n_{1,1} & \ldots & n_{1,l} & n_{1.} \\
\vdots & & \ddots & & \vdots \\
X_k & n_{k,1} & \ldots & n_{k,l} & n_{k.} \\
\hline
 & n_{.1} & \ldots & n_{.l} & n
\end{array}
$$

χ^2-test nezávislosti

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

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

Testová statistika má tvar

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

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

Fisherův exaktní test

Předpokladem χ^2-testu je, že všechny očekávané četnosti jsou větší než 5. Pokud předpoklad není splněn, používá se **Fisherův exaktní test**, známý též jako **Fisherův faktoriálový test**. Tento test počítá přímo p-hodnotu, tj. pravděpodobnost, že za platnosti H_0 bude pozorována právě naše tabulka četností. Pro čtyřpolní tabulku

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

se p-hodnota vypočítá následujícím způsobem

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

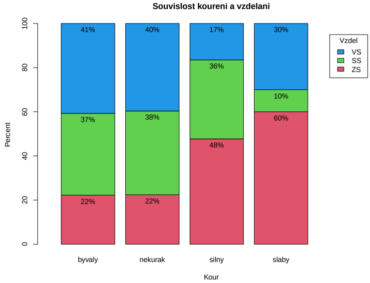
Pro větší tabulky je test složitější.

Příklad. *U 204 mužů s jedním rizikovým faktorem ischemické choroby srdeční bylo zjišťováno i vzdělání a kategorie kouření. Výsledky jsou shrnuty v následující tabulce absolutních četností. Souvisí spolu tyto dvě veličiny?*

	ZŠ	SŠ	VŠ
bývalý kuřák	6	10	11
nekuřák	13	22	23
slabý kuřák	52	39	18
silný kuřák	6	1	3

χ^2-test nezávislosti

Vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

χ^2-test nezávislosti

Testem nezávislosti jsme zjišťovali

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

Testová statistika vyšla 21.286. Porovnáváme ji s kvantilem χ^2-rozdělení $\chi^2_6 = 12.59$. Jelikož testová statistika vyšla vyšší, tak zamítáme nulovou hypotézu. P-hodnota testu vyšla 0.00163, což je menší než $\alpha = 0.05$.
Jelikož však nejsou splněny předpoklady testu, měli bychom vypočítat ještě p-hodnotu Fisherova exaktního testu. Ta vychází 0.00084.
Závěr: Prokázali jsme, že kouření se vzděláním souvisí.

Poměr šancí

Uvažujme dvouhodnotovou veličinu ve dvou populacích. Např. sledujeme výskyt chřipky ve městě a na venkově. Výsledky je možné zapsat do čtyřpolní tabulky

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

Rozdíl mezi populacemi je možné popsat poměrem šancí. Nejprve definujme **šanci** "mít chřipku proti nemít chřipku"jako

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

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

Poměr šancí

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

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

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

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

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

Věcná významnost

Výše uvedené testy měří statistickou významnost. Je ale tato významnost i skutečně zajímavá?

- p-hodnota statistického testu závisí na počtu pozorování
- málo pozorování dává "velkou"p-hodnotu
- hodně pozorování dává "malou"p-hodnotu
- statistické testy dobře fungují pro počet pozorování kolem 100 hodnot

Věcná významnost

Existuje vztah mezi počtem pozorování, hladinou významnosti a sílou testu. Pro dané parametry pak můžeme dopočítat potřebný počet pozorování. Zvolme

- hladinu významnosti $\alpha = 0.05$
- sílu testu $1 - \beta = 0.9$
- typ testu: dvouvýběrový t-test
- jak velký rozdíl mezi skupinami nás opravdu zajímá $|\mu_1 - \mu_2| = 2$
- očekávanou variabilitu $\sigma = 5$

Požadovaný počet pozorování v každé skupině je

$$n_1 = 2\left(\frac{z(1-\alpha) + z(1-\beta)}{\frac{|\mu_1 - \mu_2|}{\sigma}}\right)^2 = 2 * \left(\frac{1.96 + 1.28}{2/5}\right)^2 = 131.4$$

Věcná významnost

Pro posouzení věcné významnosti jsou vytvořeny ukazatele, které pomohou určit, zda zjištěná statistická významnost je skutečně zajímavá. Tyto ukazatele se převážně používají u velkých vzorků dat.

Velké vzorky můžeme získat např. v rámci metaanalýzy, tj. kombinace několika výzkumů na stejné téma.

Věcná významnost

Porovnání dvou skupin (dvouvýběrový test)

- Cohenovo d

$$d = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2}}, \qquad S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$$

  do 0.5 malý efekt, 0.5-0.8 střední efekt, nad 0.8 velký efekt

- Hedgesovo g

$$g = \frac{\bar{X} - \bar{Y}}{\sqrt{MSe}},$$

  MSe je zbytkový průměrný součet čtverců z tabulky analýzy rozptylu
  do 0.5 malý efekt, 0.5-0.8 střední efekt, nad 0.8 velký efekt

- Glassovo δ

$$\delta = \frac{\bar{X} - \bar{Y}}{\sqrt{S_k^2}}$$

  S_k^2 je rozptyl kontrolní skupiny
  do 0.5 malý efekt, 0.5-0.8 střední efekt, nad 0.8 velký efekt

Věcná významnost

Porovnání více než dvou skupin (ANOVA)

- Fisherovo η^2

$$\eta^2 = \frac{SSA}{SST}$$

  kde SSA a SST jsou součty čtverců z tabulky analýzy rozptylu
  procento vysvětlené variability

- Haysova ω^2

$$\omega^2 = \frac{SSA - (k-1)MSe}{SST + MSe}$$

  kde SSA, SST a MSe jsou součty čtverců/průměrné čtverce z tabulky analýzy rozptylu
  procento vysvětlené variability

Věcná významnost

Porovnání dvou kategorických proměnných (χ^2-test)

- Cramerovo ϕ

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

kde χ^2 je testová statistika χ^2-testu
do 0.29 malý efekt, 0.3-0.49 střední efekt, nad 0.5 velký efekt

- Cramerovo V

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

hodnota od 0 do 1 chovající se přibližně jako korelační koeficient

Věcná významnost

Porovnání dvou číselných proměnných (korelační koeficient, lineární regrese)

- korelační koeficient r

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

  do 0.3 malý efekt, 0.3-0.7 střední efekt, nad 0.7 velký efekt

- koeficient determinace R^2

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

  do 0.01 malý efekt, 0.01-0.25 střední efekt, nad 0.25 velký efekt

  procento variability vysvětlené modelem

Základy mnohorozměrné statistiky

Většinu základních statistických metod lze zobecnit na mnohorozměrnou situaci.
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

Základem většiny metod mnohorozměrné statistiky je měření vzdálenosti mezi dvěma body

- **Eukleidovská vzdálenost:**

$$d(\mathbf{X},\mathbf{Y}) = ||\mathbf{X}_i - \mathbf{Y}_i|| = \sqrt{(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})} = \sqrt{\sum_{i=1}^{k}(X_i - Y_i)^2}$$

  nevýhoda: všechny složky přispívají do vzdálenosti stejnou měrou a není zohledněn jejich vzájemný vztah

- **Mahalanobisova vzdálenost:**

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})^T\mathbf{S}^{-1}(\mathbf{X}-\mathbf{Y})}$$

  pro nezávislé vektory dostáváme

$$d(\mathbf{X},\mathbf{Y}) = \sqrt{\sum_{i=1}^{k}\frac{(X_i - Y_i)^2}{s_{ii}^2}}$$

  kde $\mathbf{S} = \mathrm{cov}(\mathbf{X},\mathbf{Y})$ je kovarianční matice vektorů $\mathbf{X}$ a $\mathbf{Y}$

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 liší

Testová statistika má tvar

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^T \mathbf{S}^{-1}(\overline{\mathbf{X}} - \overline{\mathbf{Y}})$$

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

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: $T^2 \sim \frac{(n-1)k}{n-k} F_{k,n-k}$.
Obdobně lze zkonstruovat i testovou statistiku pro jednovýběrový test.

Při srovnání více nezávislých výběrů se opět testují hypotézy

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

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

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

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

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.

Metoda hlavních komponent (PCA)

Při různých výzkumech bývá často zjišťováno velké množství proměnných, ze kterých má být následně zjištěna nějaká informace. Často bývají mnohé z nich vzájemně korelované a dávají tedy informaci podobnou, ne-li totožnou. Aby bylo možné nějakou informaci z proměnných získat, je často potřeba snížit jejich počet a zabývat se jen těmi skutečně zásadními.

Metoda hlavních komponent (PCA)

Metoda hlavních komponent transformuje vstupní data tak, aby bylo možné snížit jejich dimenzi / počet. Využívá se přepočet

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

kde **X** je centrovavá matice vstupních hodnot (centrování = odečet průměru), **Y** je výstupní - cílová matice a **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 pak matice vlastních čísel matice **C** a matice **P** pak obsahuje vlastní vektory matice **P**.

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

- její vektory jsou vzájemně kolmé (nezávislé)
- součet koeficientů lineárních transformací u každé komponenty je 1
- řadí se podle variability: od vektoru s největší variabilitou k vektoru s nejnižší variabilitou
- obsahuje veškerou informaci, kterou obsahovala původní data

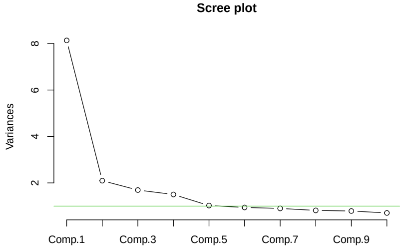
Metoda hlavních komponent (PCA)

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 korelační matice 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

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 a ε jsou náhodné chyby.

Pro faktory 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, tedy jednotlivé faktory mají nulovou střední hodnotu, jednotkový rozptyl a jsou nezávislé
- $\mathrm{E}(\varepsilon) = 0$ a $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}$, tedy náhodné chyby jsou nezávislé, stejně rozdělené s nulovou střednín hodnotou a konstantním rozptylem σ^2

Dále musí platit

- $\mathrm{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \sigma^2\mathbf{I}$, tedy
    - $\mathrm{Var}(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \sigma^2$
    - $\mathrm{Cov}(X_i, X_j) = \ell_{i1}\ell_{j1} + \ell_{i2}\ell_{j2} + \ldots + \ell_{im}\ell_{jm}$
- $\mathrm{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$, tedy $\mathrm{Cov}(X_i, F_j) = \ell_{ij}$

kde ℓ_{ij} jsou prvky matice **L**.
Na základě výše uvedených vztahů lze matici loadingů **L** určit jednoznačně až na přenásobení ortogonální maticí **T**. Toto přenásobení se dá dále využít jako *rotace* k hledání nejlépe interpretovatelných faktorů.

Hodnoty loadingů hledáme obdobně jako hlavní komponenty, tedy rozkladem korelační matice naměřených proměnných **X**. Abychom dostali interpretovatelné faktory, využívá se varimax rotace, což je taková ortogonální rotační matice **T** která dá jednotlivým proměnným co možná nejrozdílnější loadings. Pro další zpracování se používají i tzv. faktorové skóry, což jsou odhadnuté hodnoty faktorů přiřazené jednotlivým pozorováním. Ty můžeme spočítat např. pomocí násedujícího vztahu

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}(\hat{\sigma}^2\mathbf{I})^{-1}\hat{\mathbf{L}})^{-1}\hat{\mathbf{L}}'(\hat{\sigma}^2\mathbf{I})^{-1}(x_j - \overline{x})$$

Diskriminační analýza

Máme mnohorozměrná data z několika různých populací a chceme najít nejlepší možný způsob, jak na základě dat rozlišit populace mezi sebou.

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*

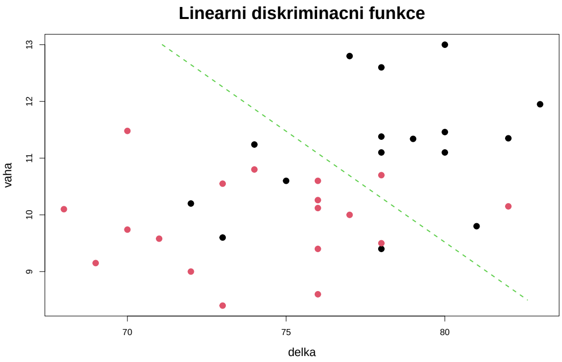
Nabízející se **postup**

- pro každou populaci spočítáme průměrný vektor
- nového jedince zařadíme do populace, která bude mít svůj průměrný vektor nejblíže k jeho hodnotám

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

Výše uvedený "nabízející se"postup vede na **lineární diskriminační analýzu**.

Uvažujme dvě populace ve dvourozměrném případě. Lineádní diskriminační analýza je odděluje přímkou



Diskriminační pravidlo pro dvě populace a obecný počet proměnných.

Označme průměrné vektory v populacích $\overline{\mathbf{X}}_{1,n}, \overline{\mathbf{X}}_{2,n}$. Pro měření vzdáleností využijeme Mahalanobisovu vzdálenost $d^2(\mathbf{X}, \mathbf{Y})$. Rozhodovací pravidlo pak zní. Pokud

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

přiřadíme pozorování k první populaci, v opačném případě ke druhé. Aritmetickými operacemi lze získat vektor

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

kde $\mathbf{S}$ je kombinovaná výběrová varianční matice obou populací

$$\mathbf{S} = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\mathbf{S}_2$$

a n_1, n_2 jsou velikosti výběrů z obou populací a $\mathbf{S}_1, \mathbf{S}_2$ jsou výběrové varianční matice obou populací.

Rozhodovací pravidlo potom zní: pokud

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

pak pozorování patří do první populace, v opačném případě do druhé. Toto pravidlo je možné také přepsat v nevektorové podobě jako

$$\sum_{i=1}^{k} c_i X_i - c_0 > 0$$

kde koeficienty c_0, c_i lze jednoznačně odvodit z vektoru **b**. Z tohoto zápisu je také zřejmé, že rozhodovací pravidlo je v tomto případě přímka.

Poznámka: Výše uvedené rozhodovací pravidlo je možné odvodit také metodou maximální věrohodnosti z hustoty mnohorozměrného normálního rozdělení

Vzniklou přímku je možné dále "posouvat" přidáním dalších podmínek:

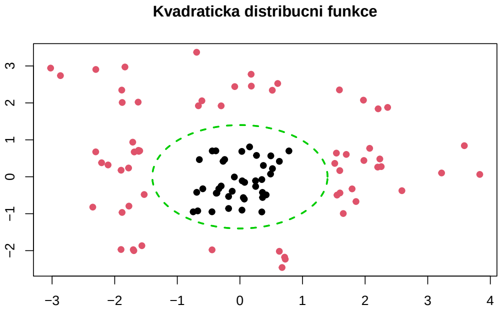
- podmínky na apriorní pravděpodobnosti obou populací, označme je π_1 a π_2
  využíváme, když je výskyt jedné populace je výrazně častější než je tomu u populace druhé
- penalizace pro špatné zařazení jednotky, označme
  $c(2|1)$ penalizaci za špatné přířazení jednotky z první populace
  $c(1|2)$ penalizaci za špatné přířazení jednotky z druhé populace

Rozhodovací pravidlo se změní na

$$\mathbf{b}^T\mathbf{X} - \mathbf{b}^T\frac{\overline{\mathbf{X}}_{1,n} + \overline{\mathbf{X}}_{2,n}}{2} + \ln\left(\frac{c(2|1)}{c(1|2)}\frac{\pi_1}{\pi_2}\right) > 0$$

Kvadratická diskriminační analýza

Někdy přímka pro oddělení populací nestačí a je potřeba použít křivku



Diskriminační pravidlo pro dvě populace pak vypadá následovně. Pokud

$$\frac{1}{2}\mathbf{X}'(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\mathbf{X}-(\overline{\mathbf{X}}_{1,n}\mathbf{S}_1^{-1}-\overline{\mathbf{X}}_{2,n}\mathbf{S}_2^{-1})\mathbf{X}+k+\ln\left(\frac{c(1|2)}{c(2|1)}\frac{\pi_2}{\pi_1}\right)\leq 0$$

kde

$$k=\frac{1}{2}\ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right)+\frac{1}{2}(\overline{\mathbf{X}}'_{1,n}\mathbf{S}_1^{-1}\overline{\mathbf{X}}_{1,n}-\overline{\mathbf{X}}'_{2,n}\mathbf{S}_2^{-1}\overline{\mathbf{X}}_{2,n})$$

pak nového jedince přiřadíme k první populaci, v opačném případě ke druhé

Mějme mnohorozměrná data a snažme se v nich najít podobnosti, abychom identifikovali různé skupiny pozorování v datech. Cílem je

- najít optimální počet skupin, tak aby mezi nimi byly rozdíly co možná největší, a v rámci skupiny, aby byly hodnoty co nejpodobnější,
- popsat skupiny tak, aby se mezi nimi dalo rozlišovat

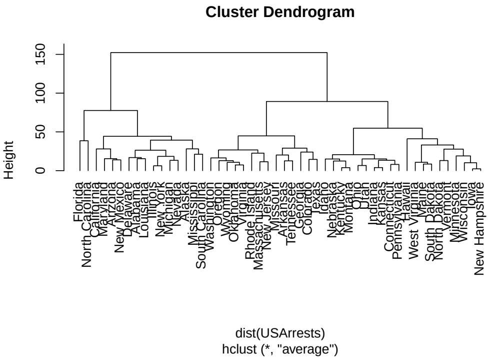
Hierarchické shlukování měří vzdálenosti mezi jednotlivými pozorováními např. euklidovskou vzdáleností a shlukuje k sobě jednotky, co jsou si nejblíže. Vzdálenost skupin se dá měřit trojím způsobem

- vzdálenost středů (průměrů) – **average linkage**
- vzdálenost nejbližších bodů – **single linkage**
- vzdálenost nejvzdálenějších bodů – **complete linkage**

Complete linkage dává většinou nejlepší výsledky.

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 pak hledáme, kde ukončit shlukování, tj. kolik skupin je optimálních.

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

- nejprve se zvolí počet skupin p
- náhodně vybereme p bodů v mnohorozměrném prostoru jako středy těchto skupin
- zařadíme prvek, který je nejbližší nějakému středu k této skupině
- středy se přepočítají
- poslední dva body se opakují, dokud nejsou rozřazeny všechny prvky

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

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.*

Pro každou skupinu proměnných pak hledáme jejich vhodnou lineární kombinaci

$$U = \mathbf{a}^T\mathbf{X}, \qquad V = \mathbf{b}^T\mathbf{Y}$$

takovou, že má mezi sebou maximální korelaci.

Kanonické korelace

Označme

$$\begin{aligned} \mathrm{E}(\mathbf{X}) = \ & \mu_1, \qquad \mathrm{Cov}(\mathbf{X}) &= \Sigma_{11} \\ \mathrm{E}(\mathbf{Y}) = \ & \mu_2, \qquad \mathrm{Cov}(\mathbf{Y}) &= \Sigma_{22} \\ \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \ & \Sigma_{12} = \Sigma_{21}' \end{aligned}$$

Pak víme, že

$$\begin{aligned} \mathrm{Var}(U) &= \mathbf{a}'\Sigma_{11}\mathbf{a} \\ \mathrm{Var}(V) &= \mathbf{b}'\Sigma_{22}\mathbf{b} \\ \mathrm{Cov}(U, V) &= \mathbf{a}'\Sigma_{12}\mathbf{b} \\ \mathrm{Cor}(U, V) &= \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}} \end{aligned}$$

Kanonické korelace

Hledejme k dvojic proměnných U_i, V_i, kde k je počet proměnných v menší skupině. Pro tyto proměnné nechť platí

- proměnné U_1, V_1 mají obě rozptyl roven jedné a maximalizují vzájemnou korelaci
- proměnné U_2, V_2 mají obě rozptyl roven jedné, jsou nekorelované s proměnnými U_1, V_1 a maximalizují vzájemnou korelaci
- $\ldots$
- proměnné U_k, V_k mají obě rozptyl roven jedné, jsou nekorelované s proměnnými $U_1, \ldots, U_{k-1}, V_1, \ldots, V_{k-1}$ a maximalizují vzájemnou korelaci.

Takovéto páry proměnných U_i, V_i se nazývají kanonické proměnné a jejich vzájemné korelace potom **kanonické korelace.**
Platí

$$\mathrm{Cor}(\mathrm{U}_1, \mathrm{V}_)) \geq \mathrm{Cor}(\mathrm{U}_2, \mathrm{V}_2) \geq \cdots \geq \mathrm{Cor}(\mathrm{U}_\mathrm{k}, \mathrm{V}_\mathrm{k})$$

Kanonické korelace

Matematická konstrukce kanonických proměnných. Lineární koeficienty **a** a **b** lze určit jako

- $\mathbf{a} = \mathbf{e}\mathbf{S}_{11}^{-1/2}$, kde **e** jsou vlastní vektory matice $\mathbf{S}_{11}^{-1/2}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1/2}$
- $\mathbf{b} = \mathbf{f}\mathbf{S}_{22}^{-1/2}$, kde **f** jsou vlastní vektory matice $\mathbf{S}_{22}^{-1/2}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1/2}$
- matice **S** jsou odhady matic Σ.

Kanonické korelace

Pokud jsou skupiny proměnných **X** a **Y** nezávislé, pak jejich teoretické kovarianční matice Σ_{12} a Σ_{21} jsou nulové. Jak však pomocí kanonických korelací tuto nezávislost otestovat?
Můžeme testovat několik různých hypotéz

- H_0 : všechny kanonické korelace jsou nulové, tedy $\Sigma_{12} = 0$
- H_0 : druhá a další kanonické korelace jsou nulové a první je nenulová, tedy $\rho_2 = \cdots = \rho_k = 0$
- H_0 : třetí a další kanonické korelace jsou nulové a první dvě jsou nenulové, tedy $\rho_3 = \cdots = \rho_k = 0$
- atd.

kde ρ_i je i-tá kanonická korelace.

Testová statistika první nulové hypotézy má tvar

$$n \ln \left(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|} \right) = -n \ln \prod_{i=1}^{k} (1 - \hat{\rho}_i^2)$$

kde $\mathbf{S}$ je matice složená z $\mathbf{S}_{11}, \mathbf{S}_{12}, \mathbf{S}_{21}, \mathbf{S}_{22}$. Tato statistika má za platnosti H_0 asymptoticky χ^2 rozdělení o kp stupních volnosti, kde p je počet proměnných ve větší skupině.
Testová statistika dalších testů má tvar

$$-(n - 1 - \frac{1}{2}(k + p + 1)) \ln \prod_{i=m+1}^{k} (1 - \hat{\rho}_i^2)$$

a za platnosti H_0 má asymptoticky χ^2 rozdělení o $(k - m)(p - m)$ stupních volnosti. m je zde počet kanonických korelací, které nechceme testovat.

Metoda maximální věrohodnosti

Způsob odhadu parametru určitého rozdělení.

- označme odhadovaný parmetr θ
- mějme naměřené hodnoty $X_1, X_2 \ldots, X_n$ z rozdělení s hustotou $f(x, \theta)$
- chceme takové $\hat{\theta}$, aby pravděpodobnost, že hodnoty X_i pochází z rozdělení $f(x, \hat{\theta})$, byla maximální
- potřebujeme konkrétní specifikaci pravděpodobnostního rozdělení $f(x, \theta)$

Metoda maximální věrohodnosti

Naměřené hodnoty $X_1, \ldots, X_n$ jsou nezávislé. Jejich sdružená hustota je tedy rovna

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

větší hodnota této funkce vyjadřuje větší shodu pozorovaných hodnot s předpokládaným rozdělením.

Odhad parametru θ získáme maximalizací této funkce přes θ

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(x_1, \ldots, x_n|\theta)$$

kde Θ je prostor všech možných hodnot parametru.

Metoda maximální věrohodnosti

Uvažujeme-li tuto funkci jako funkci parametru θ, nazýváme ji **věrohodnostní funkce** a $\hat{\theta}$ **maximálně věrohodným odhadem.**
Častěji se pracuje s logaritmickou věrohodnostní funkcí

$$\ell(\theta|x_1,\ldots,x_n) = \ln L(\theta|x_1,\ldots,x_n) = \ln \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

Tuto funkci pad derivujeme podle θ a položíme rovnu nule.

Metoda používaná pro odhad parametrů v zobecněné lineární a nelineární regresi.

Příklad. *Hledejme maximálně věrohodný odhad parametru λ z poissonova rozdělení, které má hustotu $f(x|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$. Logaritmus věrohodnostní funkce pak má tvar*

$$\begin{aligned} \ell(\theta|x_1,\dots,x_n) &= \ln \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \ln \frac{e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \\ &= \sum_{i=1}^{n} x_i \ln\lambda - n\lambda - \ln(\prod_{i=1}^{n} x_i!) \end{aligned}$$

derivací podle λ dostanu

$$\frac{d\ell}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

a tedy $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$

Dodatky k regresním modelům

Uvažujme model lineární regrese

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + e_i$$

kde

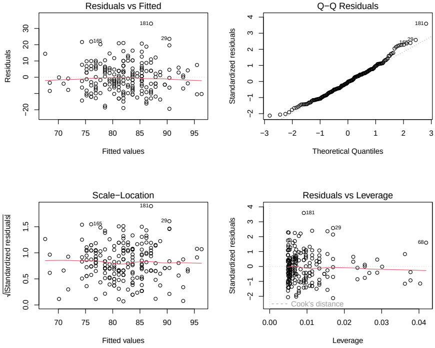
- Y_i jsou hodnoty závisle proměnné
- $X_{1i}, \ldots, X_{ki}$ jsou hodnoty nezávisle proměnných $X_1, \ldots, X_k$
- $\beta_0, \ldots, \beta_k$ jsou regresní koeficienty
- e_i jsou náhodné chyby

Předpoklady modelu lineární regrese

- $e_i \sim iid\ \mathrm{N}(0, \sigma^2)$ jsou nezávislé, stejně rozdělené náhodné veličiny s normálním rozdělením, nulovou střední hodnotou a konstantním rozptylem
- $X_1, \ldots, X_k$ jsou vzájemně nezávislé proměnné
- mezi závisle proměnnou Y a nezávisle proměnnými X je lineární vztah
- v datech nejsou vlivná pozorování

Dodatky k regresním modelům

V R-ku máme k dispozici diagnostické grafy



Dále test normality, test homoskedasticity, test nekorelovanosti residuí, test multikolinearity, Cookovu vzdálenost.

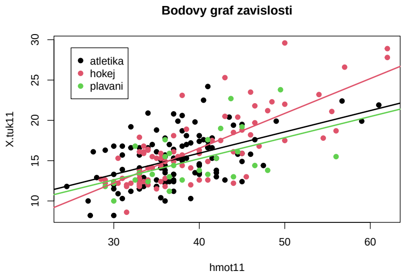
Dodatky k regresním modelům

Závislost na kategorické proměnné

- do modelu lze vkládat i kategorické regresory
- závislost na nich se modeluje pomocí dummy variables
  - $Z_1 = 1 \ldots$ když nastane 1. kategorie a $Z_1 = 0 \ldots$ jinak
  - $Z_2 = 1 \ldots$ když nastane 2. kategorie a $Z_1 = 0 \ldots$ jinak
  - $\vdots$
  - $Z_{k-1} = 1 \ldots$ když nastane (k-1). kategorie a $Z_1 = 0 \ldots$ jinak
    kde k je počet kategorií
  - proč chybí k-tá proměnná?
- v modelu se testuje, jak se která kategorie liší od referenční

Interakce

Jak se nezávisle proměnné ovlivňují při současném vlivu na proměnnou závislou



- závislost procenta tuku na hmotnosti je stejná u atletiky a plavání – není interakce
- závislost procenta tuku na hmotnosti se u hokejistů liší od ostatních sportů – interakce

Kroková regrese

Hledáme optimální regresní model

- backward – udělá se co nejsložitější model a postupně se z něj ubírají nevýznamné proměnné
  vždy se ubere proměnná s nejmenším vlivem (nejvyšší p-hodnotou, která optimalizuje AIC)
  končím, když mám v modelu jen významné proměnné
- forward – do modelu bez nezávislých proměnných se postupně po jedné přidávají
  vždy se přidá proměnná s největším vlivem (nejnižší p-hodnotou, která optimalizuje AIC)
  končím, když nemohu přidat žádnou významnou proměnnou
- both sided – kombinuje obě výše zmíněné
  v každém kroku zkusím jednu proměnnou přidat, ale také ubrat (optimalizace AIC)

Intervaly spolehlivosti

- pro regresní koeficienty

$$b_j \pm \mathrm{s.e.}(b_j)t_{n-k-1}(1-\alpha/2)$$

- pro odhad

$$b_0 + b_1 x_0 \pm sd(x_0)t_{n-k-1}(1-\alpha/2)$$

- pro předpověď

$$b_0 + b_1 x_0 \pm s\sqrt{1+d^2(x_0)}t_{n-k-1}(1-\alpha/2)$$

- kde s je střední chyba residuí a

$$d^2(x_0) = \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

Zkouška

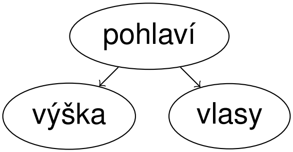
- zkouška bude ústní u počítačů
- dostanete dvě základní otázky
  - jedna analýza z mnohorozměrné statistiky
    úkolem bude připravit si ukázkový příklad a teoreticky popsat myšlenku metody a základní kroky výpočtu
  - jedna úloha na regresní modely
    úkolem bude vyřešit zadanou úlohu a popsat základní teorii popsané metody
- v případě nerozhodné známky dostanete doplňující otázku
- budete mět popsat základní myšlenku jedné ze tří metod
  - věcná významnost
  - fuzzy množiny a fuzzy modelování
  - Bayesovské sítě

Pro kombinovanou formu studia

- grafické vyjádření závislosti mezi jevy
- mírou závislosti je podmíněná pravděpodobnost
- *Příklad.* Uvažujme tři proměnné, každou se dvěma úrovněmi
  - pohlaví: muž $\times$ žena
  - výška: vyšší $\times$ nižší
  - délka vlasů: kratší $\times$ delší

  Mezi těmito proměnnými existují následující vztahy
  - Muž je spíše vyšší a žena spíše nižší
  - Muž má spíše kratší vlasy a žena spíše delší vlasy

  Jaká je pravděpodobnost, že když potkám vyššího člověka s kratšími vlasy, že to bude muž?



Bayesovká síť je orientovaný acyklický graf

- vrcholy jsou náhodné proměnné s konečně mnoha navzájem disjunktními stavy
- hrany/šipky vedou od rodiče k následovníkovi
- hrany jsou ohodnoceny podmíněnými pravděpodobnostmi P(následovník | rodič)
- pro rodiče bývají určeny apriorní pravděpodobnosti

Základní stavební kameny

- Podmíněná pravděpodobnost

$$\begin{aligned} P(A|B) &= \frac{P(A\cap B)}{P(B)} \\ P(A|B,C) &= \frac{P(A\cap B|C)}{P(B|C)} \end{aligned}$$

- Nezávislost jevů
  jevy A a B jsou nezávislé, když

$$\begin{aligned} P(A\cap B) &= P(A)P(B) \\ P(A|B) &= P(A) \end{aligned}$$

- $P(A)$ se nazývá apriorní pravděpodobnost
  $P(A|B)$ se nazývá aposteriorní pravděpodobnost

Základní stavební kameny

- Věta o úplné pravděpodobnosti
  Nechť $B_1, \ldots, B_n$ jsou jevy takové, že $\forall i, P(B_i) > 0, \forall i \neq j, B_i \cap B_j = \emptyset, \sum_{i=1}^n B_i = 1$. Pak
$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$
- Bayesova věta
$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$
- Věta o násobení pravděpodobností
  Nechť $P(A_1, \ldots, A_{n-1}) > 0$
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \ldots P(A_n|A_1, \ldots, A_{n-1})$$

Příklad využití Bayesovy věty

V morseově abecedě jsou znaky *tečka* a *řárka* v poměru 5:3. Pro přenos signálu platí:

- pokud je vyslána *tečka*, je přijata *tečka* s pstí 3/5
- pokud je vyslána *čárka*, je přijata *čárka* s pstí 2/3

Přijata byla tečka, jaká je pravděpodobnost, že byla vyslána tečka?

Řešení:
Označme

- A – jev, že byla přijata tečka
- B_1 – jev, že byla vyslána tečka
- B_2 – jev, že byla vyslána čárka

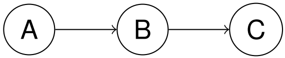
$$P(B_1) = 5/8, \qquad P(B_2) = 3/8, \qquad P(A|B_1) = 3/5, \qquad P(A|B_2) = 1/3$$

Použití Bayesovy věty

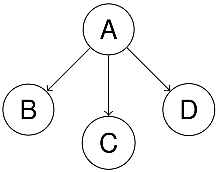
$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2)} = \frac{3/5 \cdot 5/8}{3/5 \cdot 5/8 + 1/3 \cdot 3/5} = \frac{3/8}{4/8} = \frac{3}{4}$$

V Bayesovkých sítích kombinujeme **tři typy propojení**

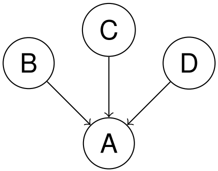
- sériové



- divergentní

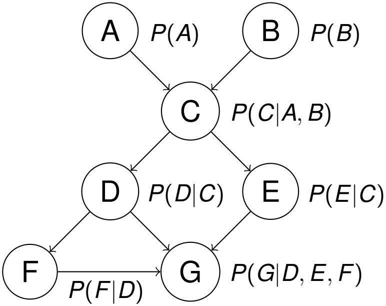


- konvergentní



Počítání pravděpodobností v Bayesovské síti

Příklad Bayesovské sítě s uvedenými pravděpodobnostmi



Sdružené rozdělení se pak počítá jsko součin podmíněných pravděpodobností

$$P(A \cap B \cap \cdots \cap G) = P(A)P(B)P(C|A,B)P(D|C)P(E|C)P(F|D)P(G|D,E,F)$$

Při práci s Bayesovskou sítí musíme

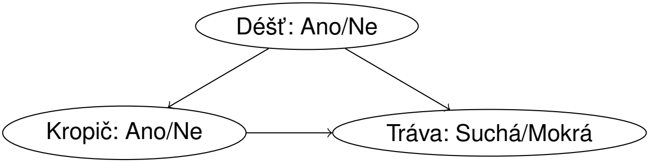
- navrhnout strukturu sítě, tj. uzly a orientované hrany
- navrhnout výchozí pravděpodobnosti
- provést výpočty

Pozorovaná hodnota určité proměnné se nazývá **evidence**

- provádíme výpočty za podmínky pozorovaných evidencí

Příklad: V určité lokalitě prší s pravděpodobností 0.2. V závislosti na tom, zda prší, se spustí kropič v zahradě. Pokud prší, spustí se s pravděpodobností 0.01, pokud neprší, spustí se s pravděpodobností 0.4. V závislosti na dešti a kropení je tráva v zahradě buďto suchá nebo mokrá (viz. tabulka níže). Z okna vidíme mokrou trávu. Jaká je pravděpodobnost, že prší?

Návrh Bayesovské sítě



Příklad: známé pravděpodobnosti

Déšť	
Ano	Ne
0.2	0.8

	Kropič	
Déšť	Ano	Ne
Ano	0.01	0.99
Ne	0.4	0.6

		Tráva	
Déšť	Kropič	Mokrá	Suchá
Ano	Ano	0.99	0.01
Ano	Ne	0.9	0.1
Ne	Ano	0.8	0.2
Ne	Ne	0.0	1.0

Víme, že sdružená pravděpodobnost je

$$P(D \cap K \cap T) = P(D)P(K|D)P(T|D,K)$$

Chceme

$$\begin{aligned} P(D = A|T = M) &= \frac{P(D = A \cap T = M)}{T = M} = \\ &= \frac{\sum_{x \in \{A,N\}} P(T = M \cap K = x \cap D = A)}{\sum_{x,y \in \{A,N\}} P(T - M \cap K = x \cap D = y)} \end{aligned}$$

Příklad: Konkrétní výpočet

$$\begin{aligned} P(T = M \cap K &= A \cap D = A) = \\ &= P(D = A)P(K = A|D = A)P(T = M|D = A, K = A) = \\ &= 0.99 \times 0.01 \times 0.2 = 0.00198 \end{aligned}$$

A celá podmíněná pravděpodobnost vychází

$$\begin{aligned} P(D = A|T = M) &= \frac{0.00198_{MAA} + 0.1584_{MNA}}{0.00198_{MAA} + 0.1584_{MNA} + 0.288_{MAN} + 0.0_{MNN}} = \\ &= \frac{891}{2491} = 0.3577 \end{aligned}$$