

Pokročilé statistické metody

Alena Černíková

alena.cernikova@ujep.cz

28. dubna 2025

Realizace výuky

- **Výuku** realizují 2 vyučující
 - Alena Černíková
 - prof. Sergii Babichev
- **Zkoušku a zápočty** má na starosti jeden vyučující
 - Alena Černíková :(

Podmínky zápočtu a zkoušky

● Zápočet

- dva až tři domácí úkoly – vyzkoušíte si příklady ze cvičení
- seminární práce – vypracování komplexního statistického úkolu, kde výstupem je souvislý text

● Zkouška – ústní u počítačů

- vylosujete si metodu, kterou předvedete na příkladu a vysvětlíte
- jedna otázka na mnohorozměrnou statistiku
- jedna otázka na regresní modely
- v případě nerozhodnosti známky doplňující otázka jedno ze tří témat:
 - fuzzy modely
 - Bayesovské sítě
 - věcná významnost

Obsah kurzu

- Teorie testování hypotéz – AČ
- Věcná významnost a metaanalýza – AČ
- Mnohonásobná lineární regrese – SB
- Zobecněné lineární modely – SB
- Nelineární modely – SB
- Fuzzy logika a fuzzy modelování – SB
- Bayesovské metody – SB
- Mnohorozměrné statistické metody – AČ
- Dodatky k regresním modelům – AČ
- Praktické úkoly na výše zmíněné – AČ

Základy testování hypotéz

Testuje se platnost tvrzení

- 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.

Vždy se testují **populační charakteristiky**. Jejich výběrové ekvivalenty se používají jen pro sestrojení testových kritérií.

Testované hypotézy

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**
 - v případě porovnání výběrů sem patří konkrétní velikost rozdílu (většinou nulová)
výběry 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**
 - v případě porovnání výběrů sem patří obecný popis rozdílu
výběry se liší

Výsledek testu

Na základě statistického testu uděláme jedno ze dvou rozhodnutí

- **Zamítneme nulovou hypotézu**
 - tím jsme prokázali platnost alternativy
- **Nezamítneme nulovou hypotézu**
 - tím jsme neprokázali nic
 - interpretace závisí na formulaci testovaných hypotéz
 - *neprokázala se platnost alternativy*
 - *nulová hypotéza může platit*

Jiný závěr udělat nemohu!

Chyby testu

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

	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

Rozhodnutí

Výsledek testu získáme

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

Platí, že

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

P-hodnota

Co je to p -hodnota

- aktuální dosažená hladina testu
- pravděpodobnost, že za platnosti H_0 nastane 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 rozdělení

- nejčastěji testujeme normalitu
- př. Shapiro-Wilkův test, χ^2 -test dobré shody atd.

● Parametrické testy

- testová statistika se počítá přímo z naměřených hodnot
- testuje se hodnota parametru, nejčastěji střední hodnoty
- předpokladem bývá konkrétní rozdělení, většinou normální
- př. dvouvýběrový t-test, ANOVA, Waldův test, Bartlettův test atd.

● Neparametrické testy

- testová statistika je založena většinou na pořadích, ne přímo na naměřených hodnotách
- jedná se o robustní metody nevyžadující konkrétní rozdělení dat
- př. Wilcoxonův test, Spearmanův korelační koeficient atd.

● Simulační testy

- nutnost využití počítačů
- na základě daného výběru se simulují další a počítá se p-hodnota
- př. permutační test, atd.

Vybrané testy

- **Test o střední hodnotě jednoho výběru**
 - normální data – jednovýběrový t-test
 - nenormální data – znaménkový test, jednovýběrový Wilcoxonův test
- **Test o střední 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 rozptyly – 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 rozptyly – klasická ANOVA pro různé rozptyly
 - nenormální rozdělení – Kruskal-Wallisův test
- **Test o nezávislosti dvou číselných proměnných**
 - normální rozdělení – Pearsonův korelační koeficient
 - nenormální rozdělení – Spearmanův korelační koeficient
- **Test o vztahu dvou kategorických proměnných**
 - závislé proměnné, test symetrie – McNemarův test
 - test nezávislosti pro velká data – Chí-kvadrát test
 - test nezávislosti pro malá data – Fisherův test
 - test nezávislosti pro ordinální proměnné – Kendallův korelační koeficient

Dvouvýběrový test

Porovnává střední hodnotu dvou **nezávislých** výběrů

Testované hypotézy

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

Kontrolují se zde 2 předpoklady

- normalitu dat
- shodu rozptylů

A vybíráme jeden ze tří testů

- **Dvouvýběrový t-test** pro normální data a shodné rozptyly
- **Welchův dvouvýběrový test** pro normální data a různé rozptyly
- **Wilcoxonův dvouvýběrový test** pro data, která nemají normální rozdělení

Test shody dvou rozptylů

Test shody rozptylů se vyhodnocuje i u nenormálních dat.

Testované hypotézy

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

Testová statistika testu je

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

a za platnosti H_0 má F -rozdělení o $n_1 - 1$ a $n_2 - 1$ stupních volnosti.

Dvouvýběrový t-test pro shodné rozptyly

Testová statistika tohoto testu má tvar

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

kde

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

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 tohoto testu má tvar

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

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

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

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

Dvouvýběrový t-test

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? Předpokládejme přibližně normální rozdělení dat.*

Test shody rozptylů

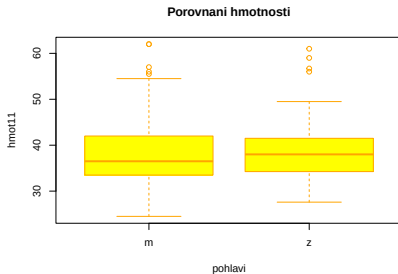
- testová statistika $F = \frac{\text{Var}(X)}{\text{Var}(Y)} = \frac{45.1}{50.6} = 0.89$
- p-hodnota = 0.56 > $\alpha = 0.05$
- nulovou hypotézu nezamítáme
- rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly

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

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

- porovnáváme s kvantilem t-rozdělení
 $t_{220}(1 - 0.025) = 1.97$ (kvantil pro oboustrannou alternativu)
- testová statistika je v absolutní hodnotě menší než tento kvantil, tak **nulovou hypotézu nezamítám.**
- p-hodnota = $0.3151 > \alpha = 0.05$
- **Závěr:** Na hladině významnosti 5% jsem neprokázala, že by se hmotnost jedenáctiletých hochů a dívek lišila.

Wilcoxonův dvouvýběrový test

Používá se pro porovnání dvou nezávislých výběrů, které nesplňují předpoklad normality.

Test je založen na pořadích hodnot sdruženého výběru.

Postup

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

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**.

Chí-kvadrát test nezávislosti

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

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

χ^2 -test nezávislosti

Kontingenční tabulka absolutních četností má tvar

	Y_1	$\dots$	Y_I	
X_1	$n_{1,1}$	$\dots$	$n_{1,I}$	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
X_k	$n_{k,1}$	$\dots$	$n_{k,I}$	$n_{k.}$
	$n_{.1}$	$\dots$	$n_{.I}$	n

χ^2 -test nezávislosti

Testované hypotézy

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

Test je založen na porovnání

- pozorovaných četností n_{ij}
- očekávaných četností $n_{i.}n_{.j}/n$

vychází z definice nezávislosti $P(A \cap B) = P(A)P(B)$

Testová statistika

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(\text{pozorovane}_{i,j} - \text{ocekavane}_{i,j})^2}{\text{ocekavane}_{i,j}} = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{i,j} - n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n}$$

za platnosti H_0 má χ^2 -rozdělení o $(k-1)(l-1)$ stupních volnosti.

Fisherův exaktní test

Test nezávislosti pro malá data

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

Pro čtyřpolní tabulku

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

se p-hodnota vypočítá následujícím způsobem

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

Pro větší tabulky je test složitější.

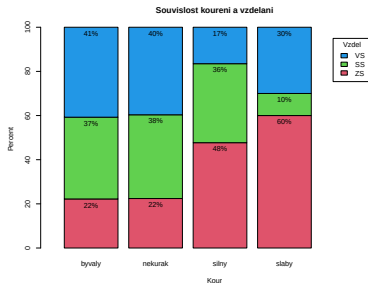
χ^2 -test nezávislosti

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

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

χ^2 -test nezávislosti

Vztah dvou kategoričkových proměnných se zobrazuje pomocí sloupcového grafu



Můžeme zobrazovat pomocí řádkových nebo sloupcových procent.

χ^2 -test nezávislosti

Testované hypotézy

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

Výsledky testu

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

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

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

Rozdíl mezi populacemi je možné popsat **poměrem šancí**.

- šance** "mít chřipku proti nemít chřipku"

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

- poměr šancí je podíl šancí v obou populacích.

Poměr šancí

Definice **poměru šancí**

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

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

Testované hypotézy

- $H_0 : OR = 1$, šance jsou stejné
- $H_1 : OR \neq 1$, šance se v populacích liší

Testová statistika je rovna

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

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

Poměr šancí

Příklad. *Uvažujme následující čtyřpolní tabulku*

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

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

Závěr: *Ve městě je významně větší šance dostat chřipku než na venkově.*

Statistický test

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

Odhad počtu pozorování

Existuje vztah mezi počtem pozorování, hladinou významnosti a silou testu.

Zvolme

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

Optimální počet pozorování v každé skupině

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

Odhad počtu pozorování

- optimální je stejný počet pozorování v obou skupinách
- pokud očekáváte, že budete používat Wilcoxonův dvouvýběrový test, přidejte navíc 15% pozorování
- počet pozorování je možné odhadnout i pro požadovanou délku intervalu spolehlivosti

Tabulka analýzy rozptylu

- používá se pro porovnání variability **vysvětlené** a variability **nevysvětlené**
- nejčastěji v ANOVě (porovnání střední hodnoty v několika nezávislých výběrech)
- označme

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

$$SSA = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2$$

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

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.

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.

Porovnání dvou výběrů

- Cohenovo d**

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

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 jsou residuální "průměrné čtverce" 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

Porovnání více výběrů

- **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

Vztah dvou kategorických proměnných

- **Cramerovo ϕ**

$$\phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\sum_{i=1}^k \frac{(p_i - p_{0i})^2}{p_{0i}}}$$

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

Vztah dvou číselných proměnných

- **korelační koeficient r**

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

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

- **koeficient determinace R^2**

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

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

procento variability vysvětlené modelem

Základy mnohorozměrné statistiky

Nepracuje se s jednou proměnnou X , ale s vektorem proměnných $\mathbf{X} = (X_1, X_2, \dots, X_k)^T$.

Příklady

- 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.
- Zkoumáme chemické složení roztoku a máme zastoupení jednotlivých prvků: železo, sodík, draslík, atd.
- Klient banky vyplní dotazník, tj. odpoví na skupinu otázek

Základní charakteristiky

- střední hodnota

$$\mu = (\mu_1, \dots, \mu_k)^T$$

- varianční matice

$$\Sigma = (\sigma_{ij}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2k} \\ \vdots & & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_k^2 \end{pmatrix}$$

kde $\sigma_{ij} = \sigma_{ji}$ je kovariance i -té a j -té složky

- odhad střední hodnoty je vektor průměrů

$$\bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_k)^T$$

- odhad varianční matice je matice výběrových kovariancí a rozptylů $\mathbf{S} = (\mathbf{s}_{ij})$, kde $s_{ij} = \text{cov}(X_i, X_j)$ pro $i \neq j$ a $s_{ii} = \text{Var}(X_i)$

Měření vzdálenosti

- **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 $\mathbf{S} = \text{cov}(\mathbf{X}, \mathbf{Y})$ je kovarianční matice vektorů $\mathbf{X}$ a $\mathbf{Y}$

Zobecnění jednorozměrných 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

- zobecnění dvouvýběrového testu
- porovnávám střední hodnotu ve dvou nezávislých výběrech
- testované hypotézy
 - H_0 : vektory středních hodnot se rovnají
 - H_1 : vektory středních hodnot se liší
- testová statistika

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})$$
$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

- za platnosti H_0 má Hotellingovo T^2 -rozdělení s k a $n_1 + n_2 - 2$ stupni volnosti
- odpovídá F -rozdělení $T^2 \sim \frac{(n_1 + n_2 - 2)k}{n_1 + n_2 - k - 1} F_{k, n_1 + n_2 - k - 1}$

MANOVA

- zobecnění analýzy rozptylu
- porovnávám střední hodnotu ve více nezávislých výběrech
- testované hypotézy
 - H_0 : vektory středních hodnot se rovnají
 - H_1 : vektory středních hodnot se liší
- porovnává se variabilita vysvětlená a nevysvětlená

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

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

- kde p značí počet výběrů a $\bar{\mathbf{Y}}_i$ 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**

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

- **Hotellingovo lambda**

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

při porovnání dvou výběrů se všechny zjednoduší na Hotellingův dvouvýběrový test.

Metoda hlavních komponent (PCA)

Principal component analysis

- zjednodušení dat
- využívá se při práci s velkým množstvím proměnných
- proměnné se transformují tak, aby se většina informace soustředila do malého počtu nových veličin
- zjednodušení je podmíněné tím, že vstupní proměnné nejsou nezávislé
- hlavní využití při grafickém zobrazení výstupů

Transformace proměnných

Nově vzniklé proměnné jsou lineární kombinací původních

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

kde

- $\mathbf{X}$ je centrovavá matice vstupních hodnot (centrování = odečet průměru)
- $\mathbf{Y}$ je matice cílových/ výstupních proměnných
- $\mathbf{P}$ je matice transformačních vektorů

Matici $\mathbf{P}$ získáme pomocí rozkladu **korelační matice** vstupních dat $\mathbf{C}$

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

kde

- $\mathbf{\Lambda}$ je matice vlastních čísel matice $\mathbf{C}$
- $\mathbf{P}$ je matice vlastních vektorů matice $\mathbf{C}$

Vsuvka

Vlastní čísla a vlastní vektory matice Bud' $\mathbf{A}$ čtvercová matice řádu n

- vlastní, neboli charakteristické číslo matice $\mathbf{A}$ je takové, pro které platí

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

- matice řádu n má n vlastních čísel, některé mohou být vícenásobné
- vlastní vektor $\mathbf{v}$ matice $\mathbf{A}$ příslušný k vlastnímu číslo λ splňuje

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = 0$$

Vlastnosti hlavních komponent

Výsledná **matice hlavních komponent $\mathbf{Y}$** má následující vlastnosti

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

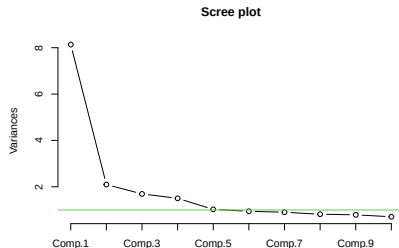
Intuitivní představa PCA

Celý **postup** si můžeme představit následovně

- mějme mnohorozměrná data v prostoru
- daty proložíme vektor ve směru s největší variabilitou
- tak získáme první hlavní komponentu
- 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

Optimální počet hlavních komponent

- **počet** hlavních komponent dostačující k reprezentaci informace obsažené ve vstupních datech odpovídá počtu vlastních čísel korelační matice větších než 1
- grafické znázornění tohoto počtu: *Scree plot*



Faktorová analýza

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

Hlavní myšlenka faktorové analýzy pochází z psychologie:

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

Faktorová analýza

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

Faktorová analýza

Vycházíme z rovnice obdobné jako u analýzy hlavních komponent

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

kde

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

Předpoklady faktorové analýzy

Aby bylo možné faktory odhadnout, musí platit

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

Předpoklady faktorové analýzy

- pokud platí výše uvedené vztahy, pak lze matici loadingů $\mathbf{L}$ určit jednoznačně až na přenásobení ortogonální maticí $\mathbf{T}$
- přenásobení se dá dále využít jako *rotace* k hledání nejlépe interpretovatelných faktorů
- nejčastěji používaná *rotace* je rotace **varimax** - ortogonální rotační matice $\mathbf{T}$
- výsledné loadings jsou buď hodně velké nebo hodně malé, podle toho, jak sytí daný faktor

Faktorová analýza

- hodnoty loadingů hledáme obdobně jako hlavní komponenty, tedy rozkladem korelační matice naměřených proměnných $\mathbf{X}$
- **faktorové skóry** jsou odhadnuté hodnoty faktorů přiřazené jednotlivým pozorováním

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

- faktorové skóry dále používáme jako nové "zjednodušující" proměnné

Diskriminační analýza

Příklad. *Uvažujme pacienty s různými nemocemi a mějme ke každému skupinu lékařských testů. Chceme pak najít způsob, jak určit, kterou nemocí pacient trpí jen na základě výsledků testů*

- mějme mnohorozměrná data z několika různých populací
- chceme najít nejlepší možný způsob, jak na základě dat rozlišit populace mezi sebou
- výsledkem mají být pravděpodobnosti příslušnosti k jednotlivým skupinám

Diskriminační analýza

Nabízející se **postup**

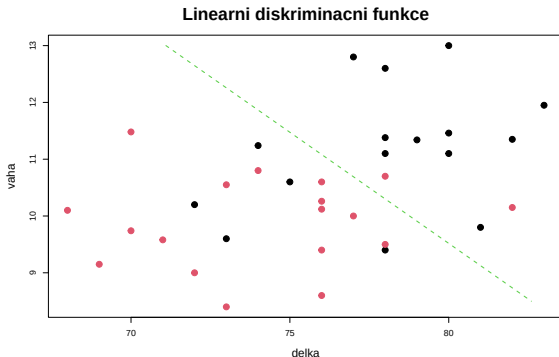
- pro každou populaci spočítáme průměrný vektor
- nového jedince zařadíme do populace, která bude mít svůj průměrný vektor nejbližší k jeho hodnotám

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

- jak dobré je určené rozhodovací pravidlo zjistíme na základě klasifikace
- počet správně přiřazených jednotek a počet špatně přiřazených jednotek

Lineární diskriminační analýza

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



Diskriminační pravidlo pro dvě populace

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

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

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

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

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.

Diskriminační pravidlo pro dvě populace

Rozhodovací pravidlo potom zní: pokud

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

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 $\mathbf{b}$. Z tohoto zápisu je také zřejmé, že rozhodovací pravidlo je v tomto případě přímka.

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

Diskriminační pravidlo pro dvě populace

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

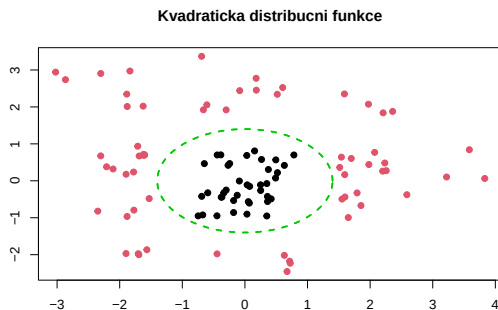
- podmínky na apriorní pravděpodobnosti obou populací, označme je π_1 a π_2
využíváme, když je výskyt jedné populace výrazně častější než je tomu u populace druhé
- penalizace pro špatné zařazení jednotky, označme $c(2|1)$ penalizaci za špatné přiřazení jednotky z první populace
 $c(1|2)$ penalizaci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

Rozhodovací pravidlo se změní na

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

Kvadratická diskriminační analýza

Někdy přímka pro oddělení populací nestačí a je potřeba použít křivku



Diskriminační analýza

Diskriminační pravidlo pro dvě populace pak vypadá následovně. Pokud

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

kde

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

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

Shluková analýza

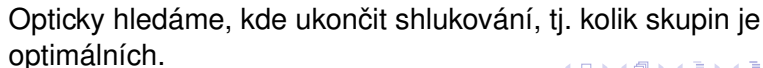
Hledání skupin v mnohorozměrných datech

- chceme optimální počet skupin, tak aby
 - rozdíly mezi skupinami byly co možná největší
 - rozdíly v rámci skupin byly co možná nejmenší (homogenní skupiny)
- výsledné skupiny se snažíme popsat, aby se mezi nimi dalo rozlišovat
- základem analýz je měření vzdáleností mezi body

Hierarchické shlukování

- začínáme s množinou bodů. kde každý tvoří jednu skupinu
- postupně slučujeme body do skupin, až tvoří jednu velkou skupinu
- podle způsobu měření vzdálenosti mezi skupinami rozlišujeme
 - **average linkage** – vzdálenost skupin je vzdálenost jejich středů (průměrů)
 - **single linkage** – vzdálenost skupin je vzdálenost nejbližších bodů
 - **complete linkage** – vzdálenost skupin je vzdálenost nejvzdálenějších bodů
 - **Wardova metoda** – skupiny jsou určeny tak, aby se minimalizovala variabilita v rámci skupin
- Wardova metoda dává většinou "nejlepší výsledky"

Grafické znázornění hierarchického shlukování se nazývá **dendrogram**.



K-means

Postup shlukování metodou **K-means**

- 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

Srovnání metod

Nevýhody jednotlivých metod

- **hierarchické shlukování** – odlehlé hodnoty zde často tvoří samostatné skupiny
- **K-means** – pokud v datech nejsou jednoznačné skupiny, pak rozřazování dopadne jinak při jiné volbě náhodných středů

Kanonické korelace

Máme dvě skupiny proměnných $\mathbf{X}$ a $\mathbf{Y}$ 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

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

takovou, že má mezi sebou maximální korelaci.

Kanonické korelace

Označme

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

Pak víme, že

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

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

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

Platí

$$\text{Cor}(U_1, V_1) \geq \text{Cor}(U_2, V_2) \geq \dots \geq \text{Cor}(U_k, V_k)$$

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 $\rho_2 = \dots = \rho_k = 0$
- H_0 : třetí a další kanonické korelace jsou nulové a první dvě jsou nenulové, tedy $\rho_3 = \dots = \rho_k = 0$
- atd.

kde ρ_i je i -tá kanonická korelace.

Kanonické 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 $X_1, X_2, \dots, X_n$ 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 $X_1, \dots, X_n$ jsou nezávislé. Jejich sdružená hustota je tedy rovna

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

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

Odhad parametru θ získáme maximalizací této funkce přes θ

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

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

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

Tuto funkci pak derivujeme podle θ a položíme rovnu nule.

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

Metoda maximální věrohodnosti

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

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

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

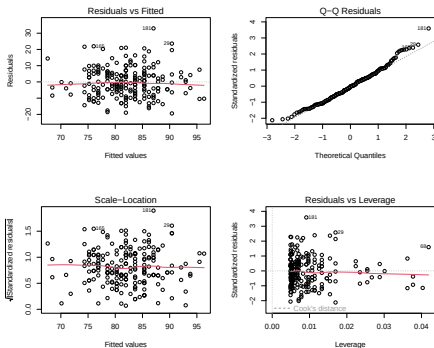
- Y_i jsou hodnoty závisle proměnné
- $X_{1i}, \dots, X_{ki}$ jsou hodnoty nezávisle proměnných $X_1, \dots, X_k$
- $\beta_0, \dots, \beta_k$ jsou regresní koeficienty
- e_i jsou náhodné chyby

Předpoklady modelu lineární regrese

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

Testy předpokladů

V R-ku máme k dispozici diagnostické grafy



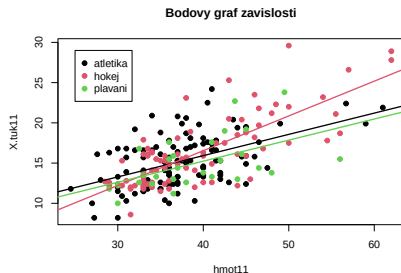
Dále test normality, test homoskedasticity, test nekorelovanosti residuí, test multikolinearity, Cookovu vzdálenost.

Závislost na kategorické proměnné

- do modelu lze vkládat i kategorické regresory
- závislost na nich se modeluje pomocí **dummy variables**
 - $Z_1 = 1 \dots$ když nastane 1. kategorie a $Z_1 = 0 \dots$ jinak
 - $Z_2 = 1 \dots$ když nastane 2. kategorie a $Z_1 = 0 \dots$ jinak
 - $\vdots$
 - $Z_{k-1} = 1 \dots$ když nastane $(k-1)$. kategorie a $Z_1 = 0 \dots$ 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

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

- 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

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

Interpretace regresních koeficientů

Mnohonásobná lineární regrese

Mějme odhadnutý regresní model

průtok vody v řece = $53 + 2.6 \times \text{srážky} + 3.8 \times \text{pole} + 5.9 \times \text{město}$

- na jeden díl srážek stoupne průtok vody v průměru o 2.6 na stejném typu půdy
- v polích stoupne průtok vody o 3.8 více než v lese při stejném množství srážek

Logistická regrese

Mějme odhadnutý regresní model

$\text{logit}(\text{přežití}) = -0.07 - 1.34 \times 2.\text{třída} - 2.19 \times 3.\text{třída} - 1 \times \text{posádka} + 3.2 \times \text{ženy}$

- cestující ve 3. třídě mají $1 / \exp / \{-2.19\} = 8.9$, tj. téměř 9 krát menší šanci na přežití, než cestující v první třídě, při stejném pohlaví
- ženy mají $\exp\{3.2\} = 24.2$ krát větší šanci na přežití než muži cestující ve stejné třídě

Zobecněné lineární modely

- 0 – 1 závislá proměnná – **Logistická regrese**
- ordinální závislá proměnná – **Ordinální regrese**
- závislá proměnná počet (count data) – **Poissonova regrese**

Ordinální regrese

- kombinace několika logistických regresí
- závislá proměnná v těchto jednotlivých modelech je

$$\begin{aligned} Y &= 0 \dots X \leq j \\ &= 1 \dots X > j \end{aligned}$$

- modeluje se šance přechodu do vyšší kategorie
- předpokladem je, že všechny dílčí logistické regrese mají stejný "směr", tedy lineární člen
- absolutní člen je pro každou logistickou regresi individuální

Poissonova regrese

- závisle proměnná je **počet**
- pomocí této regrese je možné modelovat i vztahy v kontingenční tabulce
- model má tvar

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

- předpokladem je, že rozptyl závisle proměnné je stejný jako její střední hodnota (vlastnost Poissonova rozdělení)
- není-li splněno
 - použijte se *dispersion parameter* $\Rightarrow$ **Quasi-Poisson**
 - použijte se namísto poissonova rozdělení **negativně binomické rozdělení**

Bayesovské sítě

Pro kombinovanou formu studia

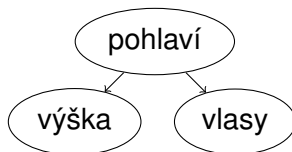
Bayesovské sítě

- 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?



Bayesovské sítě

Bayesovská 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 $P(\text{následovník} \mid \text{rodič})$
- pro rodiče bývají určeny apriorní pravděpodobnosti

Bayesovské sítě

Základní stavební kameny

- Podmíněná pravděpodobnost

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

$$P(A|B, C) = \frac{P(A \cap B|C)}{P(B|C)}$$

- Nezávislost jevů
jevy A a B jsou nezávislé, když

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

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

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

Bayesovské sítě

Základní stavební kameny

- Věta o úplné pravděpodobnosti

Nechť $B_1, \dots, B_n$ 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 $P(A_1, \dots, A_{n-1}) > 0$

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1, \dots, A_{n-1})$$

Bayesovské sítě

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

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

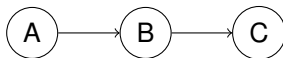
Použití Bayesovy věty

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

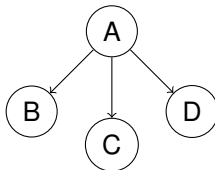
Bayesovské sítě

V Bayesovských sítích kombinujeme **tři typy propojení**

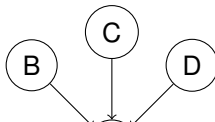
- sériové



- divergentní



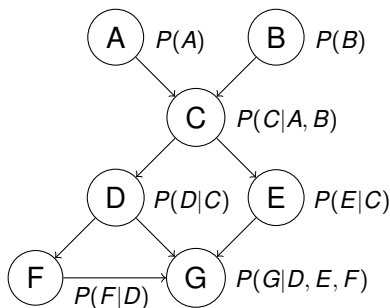
- konvergentní



Bayesovské sítě

Počítání pravděpodobností v Bayesovské síti

Příklad Bayesovské sítě s uvedenými pravděpodobnostmi



Sdružené rozdělení se pak počítá jako součin podmíněných pravděpodobností

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

Bayesovské sítě

Při práci s Bayesovskou sítí musíme

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

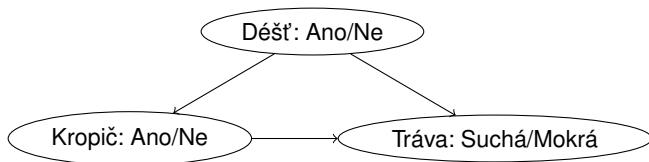
Pozorovaná hodnota určité proměnné se nazývá **evidence**

- provádíme výpočty za podmínky pozorovaných evidencí

Bayesovské sítě

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 travu. Jaká je pravděpodobnost, že prší?

Návrh Bayesovské sítě



Bayesovské sítě

Příklad: známé pravděpodobnosti

Děšť		Kropič		Tráva	
				Mokrá	Suchá
Ano	Ne	Ano	Ne		
0.2	0.8	Ano	0.01	0.99	0.01
		Ne	0.4	0.9	0.1
				0.8	0.2
				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

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

Bayesovské sítě

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

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