PSM

POKROČILÉ STATISTICKÉ METODY

2024 / 2025

Obsah

1. Testování Hypotéz

1.1 Úvod

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 $\boldsymbol{H_0}$
  – 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 $\boldsymbol{H_1}$
  – 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

Chyby testu

– při rozhodování můžeme udělat chybu

- **Chyba prvního druhu**
  – zamítneme $\boldsymbol{H_0}$, přestože platí
  – značí se α, a jmenuje se hladina významnosti
  – závažnější z obou chyb

- **Chyba druhého druhu**
  – nezamítneme $\boldsymbol{H_0}$, přestože platí $\boldsymbol{H_1}$
  – značí se β a hodnota 1 − β 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 $|\boldsymbol{T}| \geq \boldsymbol{c}$ nebo ***p-hodnota*** $\leq \boldsymbol{\alpha}$ potom *ZAMÍTÁME* H_0
- absolutní hodnota testové statistiky $|\boldsymbol{T}| < \boldsymbol{c}$ nebo ***p-hodnota*** $> \boldsymbol{\alpha}$ potom *NEZAMÍTÁME* H_0

P-hodnota

– aktuální dosažená hladina testu
– pravděpodobnost, že za platnosti $\boldsymbol{H_0}$ nastane výsledek, jaký nastal, nebo jakýkoliv jiný, který ještě více odpovídá alternativě
– definice p-hodnoty se týká testové statistiky
– (ne)zamítnout $\boldsymbol{H_0}$ nestačí, tento výsledek je třeba interpretovat vzhledem k položené otázce

Vybrané testy – přehled

- **Testy rozdělení**
  – nejčastěji testujeme normalitu
  – př. Shapiro-Wilkův test, x2-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.

- **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 střední 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 středních hodnot více závislých výběrů**
  – normální data – ANOVA pro opakovaná měření
  – nenormální data – Friedmanův test

- **Porovnání středních hodnot více nezávislých výběrů**
  – normální rozdělení, shodné rozptyly – klasická ANOVA pro shodné rozptyly
  – normální rozdělení, různé rozptyly – klasická ANOVA pro různé rozptyly
  – nenormální rozdělení – Kruskall-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 – Mc Nemmarův test
  – test nezávislosti pro velká data – Chí-kvadrát test
  – test nezávislosti pro malá data – Fisherův test
  – test nezávislosti pro ordinální proměnné – Kendallův korelační koeficient

1.2 Dvouvýběrové testy

– 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 ≠ 0, < 0, > 0

– kontrolují se zde 2 předpoklady
- normalita dat
- shoda rozptylů

– 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

– vyhodnocuje se i u nenormálních dat
– testované hypotézy:
- H_0 : rozptyly se ve výběrech neliší
- H_1 : rozptyly se ve výběrech liší

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

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

1.2.1 Dvouvýběrový t-test (pro shodné rozptyly)

– testová statistika má tvar: $T = \frac{\overline{X} - \overline{Y} - \mu_0}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$

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

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

1.2.2 Welchův test

– testová statistika má tvar: $T = \frac{\overline{X} - \overline{Y} - \mu_0}{\sqrt{\frac{\mathrm{Var}(X)}{n_1} + \frac{\mathrm{Var}(Y)}{n_2}}}$

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

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

kritické hodnoty je možno odvodit, přestože ν není celé číslo

1.2.3 Wilcoxonův dvouvýběrový test

– používá se pro porovnání dvou nezávislých výběrů, které nesplňují předpoklad normality
– test 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

– výpočet:
- označme T_1, T_2 součet pořadí v prvním (respektive druhém) výběru

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

- kde n_1, n_2 jsou rozsahy jednotlivých výběrů. Přesný test porovnává hodnotu *min(U1, U2)* s kritickou hodnotou
- asymptoticky platí: $U_0 = \frac{U_1 - \frac{1}{2}n_1 n_2}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}$

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

1.2.4 Příklady

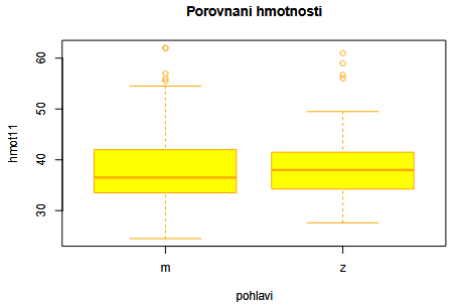
– ***Příklad 1 – dvouvýběrový t-test***:

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 = 45.1/50.6 = 0.89
  - p-hodnota = 0.56 > α = 0.005
  - nulovou hypotézu **ne**zamítáme
  - rozptyly ve skupinách jsou přibližně stejné a můžeme použít dvouvýběrový t-test pro shodné rozptyly

- 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 ≠ 0

- grafické porovnání:



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

- 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* ***ne****zamítám*
- p-hodnota = 0.3151 > α = 0.05

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

– Příklad 2 – Wilcoxonův dvouvýběrový test:

Chceme porovnat výsledky test ˚u student ˚u 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 srovnáme všechny hodnoty do řady:
  - ***41***, 45, 53, 56, ***62***, ***66***, ***76***, 77, ***79***, 79, ***80***, 81, ***84***

- následně jim přiřadíme pořadí:
  - ***1***, 2, 3, 4, ***5***, ***6***, ***7***, 8, ***9.5***, 9.5, ***11***, 12, ***13***

- pak vypočteme
  - $T_1 = 38.5$, $\boldsymbol{T_2 = 52.5}$, $U_1 = 24.5$, $\boldsymbol{U_2 = 17.5}$, $U_0 = 0.5$, p = 0.6678

- ***Závěr***: p-hodnota > α a tedy **ne**zamítám nulovou hypotézu, neprokázal se rozdíl mezi studenty v Ústí a v Liberci.

1.3 Chí-kvadrát testy – příklady testů

1.3.1 χ^2-test nezávislosti (chí-kvadrát)

– 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 má tvar:

	Y_1	$\ldots$	Y_l	
X_1	$n_{1,1}$	$\ldots$	$n_{1,l}$	$n_{1.}$
$\vdots$		$\ddots$		$\vdots$
X_k	$n_{k,1}$	$\ldots$	$n_{k,l}$	$n_{k.}$
	$n_{.1}$	$\ldots$	$n_{.l}$	n

– testované hypotézy:

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

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

- za platnosti H_0 má χ2-rozdělení o $(k-1)(l-1)$ stupních volnosti

1.3.2 Fisherův exaktní test

– test nezávislosti pro malá data

- když není splněn předpoklad χ2-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: $$p = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

– pro větší tabulky je test složitější

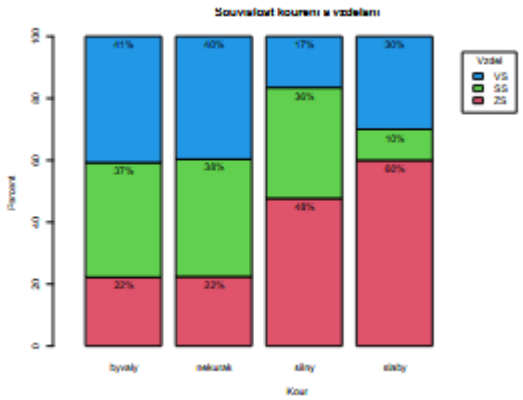
1.3.3 Příklady

– Příklad 1 – χ2-test nezávislosti:

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

– vztah dvou kategorických proměnných se zobrazuje pomocí sloupcového grafu (můžeme zobrazovat pomocí řádkových nebo sloupcových procent) :



– testované hypotézy

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

– výsledky testu

- testová statistika χ2 testu 21.286 > 12.59, kvantil χ2-rozdělení s 6 stupni volnosti
- p-hodnota 0.00163 < α = 0.05
- p-hodnotu Fisherova exaktního testu 0.00084 < α = 0.05
- některé očekávané četnosti jsou menší než 5 (není splněn předpoklad χ2 testu)
- na základě Fisherova testu ***zamítáme*** *nulovou hypotézu*

– ***Závěr***: Prokázali jsme, že kouření se vzděláním souvisí.

1.4 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"
- poměr šancí je podíl šancí v obou populacích

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

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

- interpretace: kolikrát je větší šance na chřipku ve městě než na venkově

– testované hypotézy

- H_0 : OR = 1, šance jsou stejné
- H_1 : OR ≠ 1, šance se v populacích liší

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

- za platnosti nulové hypotézy má *N(0, 1)* rozdělení

– ***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 < α = 0.05
- ***zamítáme nulovou hypotézu***

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

2. Věcná významnost

2.1 Počet pozorování

Statistický test

– doposud 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

– vztah mezi počtem pozorování, hladinou významnosti a sílou 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: $n_1 = 2\left(\frac{z(1-\alpha)+z(1-\beta)}{\frac{|\mu_1-\mu_2|}{\sigma}}\right)^2 = 2\left(\frac{1.96+1.28}{2/5}\right)^2 = 131.4$

- 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

2.2 Tabulka analýzy rozptylu

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

– označme:

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

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

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

	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

2.3 Věcná významnost

– pro posouzení věcné významnosti 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 (kombinace několika výzkumů na stejné téma)

2.3.1 Porovnání dvou výběrů

– ***Cohenovo d***

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

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

– ***Hedgesovo g***

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

- *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***

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

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

2.3.2 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 ω2***

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

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

2.3.3 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 χ2 je testová statistika χ2-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

2.3.4 Vztah dvou číselných proměnných

– 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

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

- do 0.01 malý efekt, 0.01 – 0.25 střední efekt, nad 0.25 velký efekt
- procento variability vysvětlené modelem

3. Mnohorozměrná statistika

3.1 Úvod

3.1.1 Základy mnohorozměrné statistiky

– nepracuje se s jednou proměnnou X , ale s vektorem proměnných $X = (X_1, X_2, \ldots, X_k)^T$

– příklady:
- Měříme několik fyzických parametrů jedince: výška, váha, krevní tlak, vitální kapacitu plic, atd.
- Každý žák na vysvědčení dostane známku z několika předmětů
- Zkoumáme chemické složení roztoku a máme zastoupení jednotlivých prvků
- Klient banky vyplní dotazník, tj. odpoví na skupinu otázek

– základní charakteristiky
- střední hodnota $\mu = (\mu_1, \ldots, \mu_k)^T$
- variační matice $\Sigma = (\sigma_{ij}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2k} \\ \vdots & & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_k^2 \end{pmatrix}$
- odhad střední hodnoty je vektor průměrů $\overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_k)^T$
- odhad varianční matice je matice výběrových kovariancí a rozptylů $\mathbf{S} = (\mathbf{s_{ij}})$ kde $s_{ij} = \mathrm{cov}(X_i, X_j)$ pro $i \neq j$, $s_{ii} = \mathrm{Var}(X_i)$

3.1.2 Měření vzdálenosti

– Euklidovská 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 ***S = cov(X, Y)*** je kovarianční matice vektorů ***X*** a ***Y***

3.1.3 Zobecnění jednorozměrných metod

– dvouvýběrový test	⇒	Hotellingův test
– analýza rozptylu (ANOVA)	⇒	MANOVA
– korelační koeficient	⇒	Kanonické korelace
– lineární regrese	⇒	Mnohorozměrná lineární regrese (kde závisle proměnná má více složek)

3.2 Porovnání výběrů

3.2.1 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:

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

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

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

3.2.2 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

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

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

- kde $\boldsymbol{p}$ značí počet výběrů a $\boldsymbol{Y_i}$ průměr $\boldsymbol{i}$-tého výběru

– testové statistiky pro MANOVu:
- **Wilkovo lambda**

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

- **Pillayova stopa**

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

- **Hotellingovo lambda**

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

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

3.3 Metoda hlavních komponent (PCA)

– 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
  - ***X*** je centrovavá matice vstupních hodnot (centrování = odečet průměru)
  - ***Y*** je matice cílových/ výstupních proměnných
  - ***P*** je matice transformačních vektorů

– matici P získáme pomocí rozkladu korelační matice vstupních dat ***C***

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

- kde
  - ***Λ*** je matice vlastních čísel matice ***C***
  - ***P*** je matice vlastních vektorů matice ***C***

– vlastní čísla a vlastní vektory matice bud' *A* čtvercová matice řádu ***n***

- vlastní, neboli charakteristické číslo matice *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 v matice *A* příslušný k vlastnímu číslo ***λ*** splňuje

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

Vlastnosti hlavních komponent

– výsledná matice hlavních komponent ***Y*** má 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

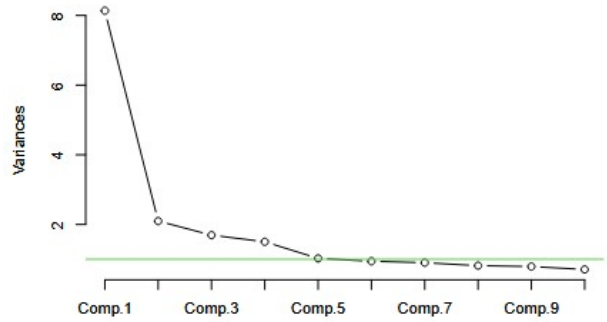
Intuitivní představa PCA

– celý postup si můžeme představit následovně:

- mějme mnohozměrná 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*



3.4 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

– 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.*

– vycházíme z rovnice obdobné jako u analýzy hlavních komponent

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

- kde
  - ***X*** je centrovaná matice naměřených dat
  - ***L*** jsou tzv. *loadings*
  - ***F*** jsou hledané faktory
  - ***ε*** jsou náhodné chyby

Předpoklady faktorové analýzy

– aby bylo možné faktory odhadnout, musí platit
- ***F*** a ***ε*** jsou nezávislé

- $\mathrm{E}(\mathbf{F}) = 0$ a $\mathrm{Cov}(\mathbf{F}) = \mathbf{I}$ kde **I** je jednotková matice
  - tj. faktory mají nulovou střední hodnotu, jednotkový rozptyl a jsou nezávislé

- $\mathrm{E}(\varepsilon) = 0$ a $\mathrm{Cov}(\varepsilon) = \sigma^2\mathbf{I}$
  - tj. náhodné chyby jsou nezávislé, stejně rozdělené s nulovou středních hodnotou a konstantním rozptylem $\boldsymbol{\sigma^2}$

- $\mathrm{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \sigma^2\mathbf{I}$, 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**

– pokud platí výše uvedené vztahy, pak lze matici *loadingů* **L** určit jednoznačně až na přenásobení ortogonální maticí **T**
– přenásobení se dá dále využít jako rotace k hledání nejlépe interpretovatelných faktorů
– nejčastěji používaná *rotace* je rotace ***varimax*** – ortogonální rotační matice **T**
– výsledné *loadingy* jsou bud' 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 X
- faktorové skóry jsou odhadnuté hodnoty faktorů přiřazené jednotlivým pozorováním

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

- faktorové skóry dále používáme jako nové "zjednodušující" proměnné

3.5 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.*

– nabízející se postup:

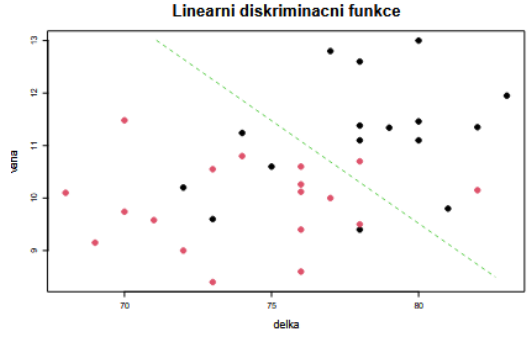
- pro každou populaci spočítáme průměrný vektor
- nového jedince zařadíme do populace, která bude mít svůj průměrný vektor nejblíže k jeho hodnotám

– 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 $\overline{\mathbf{X}}_{1,n}, \overline{\mathbf{X}}_{2,n}$
– 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 ***S*** je kombinovaná výběrová varianční matice obou populací

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

n_1, n_2 jsou velikosti výběrů z obou populací a S_1, 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 $\boldsymbol{c_0}$, $\boldsymbol{c_i}$ lze jednoznačně odvodit z vektoru ***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

– vzniklou přímku je možné dále "posouvat" přidáním dalších podmínek:

- podmínky na apriorní pravděpodobnosti obou populací, označme je $\boldsymbol{\pi_1}$ a $\boldsymbol{\pi_2}$
  využíváme, když je výskyt jedné populace výrazně častější než je tomu u populace druhé

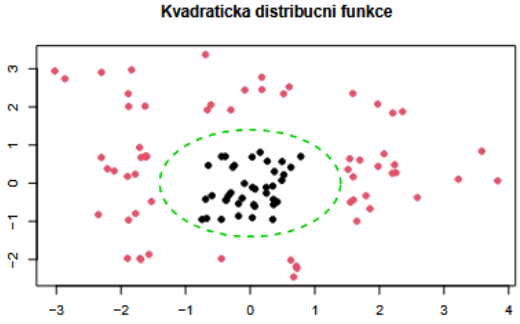
- penalizace pro špatné zařazení jednotky, označme
  c(2\|1) penalizaci za špatné přiřazení jednotky z první populace
  c(1\|2) penalizaci za špatné přiřazení jednotky z druhé populace

– rozhodovací pravidlo se změní na:

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

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

3.6 Shluková analýza

– hledání skupin v mnohorozměrných datech

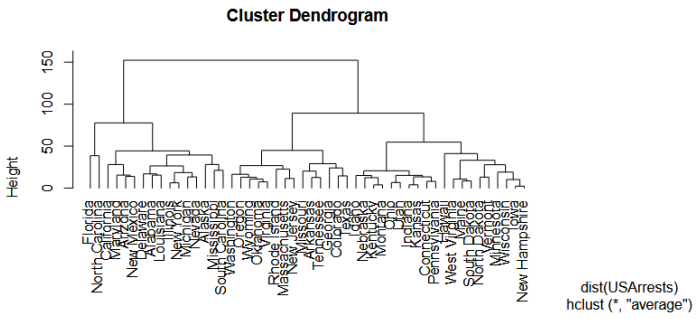
- chceme optimální počet skupin, tak aby
  - rozdíly mezi skupinami byly co možná největší
  - rozdíly v rámci skupin byly co možná nejmenší (homogenní skupiny)
- výsledné skupiny se snažíme popsat, aby se mezi nimi dalo rozlišovat
- základem analýz je měření vzdáleností mezi body

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 určené tak, aby se minimalizovala variabilita v rámci skupin
  – dává většinou "nejlepší výsledky"

– grafické znázornění hierarchického shlukování se nazývá ***dendrogram***



– opticky hledáme, kde ukončit shlukování, tj. kolik skupin je optimálních

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

3.7 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

– označme:

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

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \Sigma_{12} = \Sigma_{21}'$$

– pak víme, že:

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

– hledejme k dvojic proměnných ***Ui*** , ***Vi*** , kde k je počet proměnných v menší skupině

– pro tyto proměnné nechť platí:

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

– takovéto páry proměnných $\boldsymbol{U_i}$, $\boldsymbol{V_i}$ se nazývají kanonické proměnné a jejich vzájemné korelace potom *kanonické korelace*
– platí:

$$\mathrm{Cor}(U_1, V_1) \geq \mathrm{Cor}(U_2, V_2) \geq \cdots \geq \mathrm{Cor}(U_k, V_k)$$

– 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 $\boldsymbol{\Sigma}$

– 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 ***S*** je matice složená z $\boldsymbol{S_{11}}$, $\boldsymbol{S_{12}}$, $\boldsymbol{S_{21}}$, $\boldsymbol{S_{22}}$
- za platnosti H_0 má asymptoticky χ^2 rozdělení o $\boldsymbol{k_p}$ stupních volnosti, kde ***p*** je počet proměnných ve větší skupině

– testová statistika dalších testů mí tvar:

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

- 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

4. Regresní modely

4.1 Metoda maximální věrohodnosti

– způsob odhadu parametru určitého rozdělení

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

– naměřené hodnoty $\boldsymbol{X_1, \ldots, X_n}$ 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

– 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 pak 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:

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

- derivací podle λ dostaneme:

$$= \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \ln(\prod_{i=1}^{n} x_i!)$$

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

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

4.2 Dodatky k regresním modelům

– uvažujme model lineární regrese

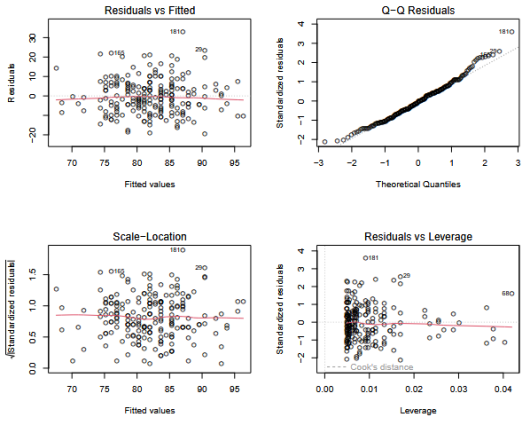
- kde
  - $\boldsymbol{Y_i}$ jsou hodnoty závisle proměnné
  - $\boldsymbol{X_{1i}}$, . . . , $\boldsymbol{X_{ki}}$ jsou hodnoty nezávisle proměnných $\boldsymbol{X_1}$, . . . , $\boldsymbol{X_k}$
  - $\boldsymbol{\beta_0}$, . . . , $\boldsymbol{\beta_k}$ jsou regresní koeficienty
  - $\boldsymbol{e_i}$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

– **předpoklady** modelu lineární regrese:

- $\boldsymbol{ei \sim iid\ N(0, \sigma^2)}$ jsou nezávislé, stejně rozdělené náhodné veličiny s normálním rozdělením, nulovou střední hodnotou a konstantním rozptylem
- $\boldsymbol{X_1}$, . . . , $\boldsymbol{X_k}$ jsou vzájemně nezávislé proměnné
- mezi závisle proměnnou $\boldsymbol{Y}$ a nezávisle proměnnými $\boldsymbol{X}$ 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***

- $Z_1 = 1$ . . . když nastane 1. kategorie a $Z_1 = 0$ . . . jinak
- $Z_2 = 1$ . . . když nastane 2. kategorie a $Z_1 = 0$ . . . jinak
- $Z_{k-1} = 1$ . . . když nastane *(k–1)*. kategorie a $Z_1 = 0$ . . . jinak kde k je počet kategorií

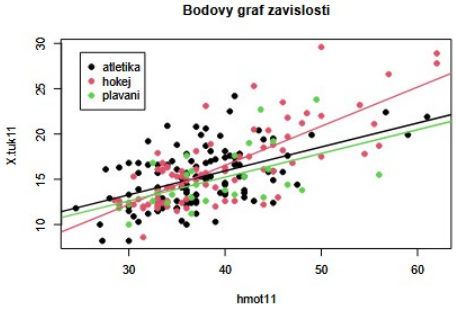
– v modelu se testuje, jak se která kategorie liší od referenční

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*)
  – končíme, když máme 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*)
  – končíme, když nemůžeme 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 × srážky + 3.8 × pole + 5.9 × 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
  - logit(přežití) = −0.07−1.34×2.třída−2.19×3.třída−1×posádka+3.2×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

4.3 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: $Y = 0 \ldots X \leq j$
$= 1 \ldots X > j$
– 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*** ⇒ **Quasi-Poisson**
    - použije se namísto poissonova rozdělení **negativně binomické rozdělení**

5. Simulace ve statistice

5.1 Statistické testy

5.1.1 Typy testů

– statistické testy dělíme na

- **Parametrické**
  – předpokládají určité rozdělení dat, nejčastěji normální

- **Neparametrické**
  – jsou použitelné nezávisle na rozdělení dat
  - založené na **pořadích** namísto původních hodnot pracují s jejich pořadími, nenáročné na výpočet
  - založené na **přeuspořádání** (*resampling*) pracují přímo s naměřenými hodnotami, které různě přeskupují
    - ***Permutační***
    - ***Bootstrap***

5.1.2 Permutační testy

– počítá se testová statistika

– ***princip*** – jak si stojí aktuálně naměřená hodnota testové statistiky mezi statistikami, které je možné získat permutací původních hodnot

– ***p-hodnota*** – podíl testových statistik v absolutní hodnotě větších než aktuální naměřená hodnota

Příklad pro dvouvýběrový test:

- porovnáváme dva nezávislé výběry s počty pozorování n_1 a n_2
- testujeme
  - H_0 : střední hodnoty jsou stejné
  - H_1 : střední hodnoty se liší
- vypočteme testovou statistiku klasického dvouvýběrového t-testu

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

kde

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

- spočítáme velké množství (alespoň $N = 10000$) permutací původních hodnot při neměnných rozsazích výběrů n_1, n_2
- pro každou permutaci spočítáme hodnotu testové statistiky T_i
- vypočteme p-hodnotu:

$$p = \frac{\#T_i : |T_i| \geq T}{N}$$

Vlastnosti permutačních testů

– permutační test lze využít pro libovolnou testovou statistiku
– lze ho využít bez předpokladu na rozdělení dat
– předpokladem je přibližně stejný rozptyl ve výběrech
– je-li příliš malý počet pozorování, je malý i počet různých permutací a tedy i počet možných různých p-hodnot
– pro menší vzorky je možné spočítat "přesnou" p-hodnotu udržují hladinu významnosti, mnohdy lépe než klasické testy
– výpočetně náročné

5.1.3 Bootstrap

– princip je obdobný jako u permutačních testů
– počítá se testová statistika a její aktuální hodnota se porovnává s hodnotami získanými z náhodných výběrů z dat
– náhodné výběry se ale nezískávají permutováním, ale ***náhodným výběrem s vracením***
– bootstrapové náhodné výběry mohou některé původní hodnoty obsahovat víckrát a jiné vůbec
– častěji se používá pro výpočet ***intervalů spolehlivosti***, ale je možné je využít i k testování
– optimální, pokud chceme odhadnout ***rozdělení*** měřeného parametru

5.1.4 Odhady metodou Jackknife

– jednoduché přeuspořádání původních hodnot používané k odhadům libovolných parametrů
  - mějme jeden výběr a potřebujeme odhadnout parametr θ (průměr, rozptyl, medián, šikmost, ...)
  - tvoříme náhodné výběry, pro něž vypočteme odhadovaný parametr a výsledky pak zprůměrujeme
  - náhodné výběry metodou Jackknife vznikají vynecháním jedné hodnoty z původních dat
  - pro n pozorování vytvoříme n přeuspořádaných výběrů o $n - 1$ hodnotách

– **vlastnosti metody**
  - výpočetní jednoduchost
  - deterministická, tj. nenáhodná metody (vždy dostanu stejný výsledek)
  - optimálně pracuje při známém rozdělení dat snižuje vychýlení (*bias*) odhadu
  - nedoporučuje se používat při velké variabilitě v datech nebo při příliš komplexních modelech

5.1.5 Bayesovská statistika

– předpokládejme náhodnou veličinu X s hustotou rozdělení $f(x|\theta)$ závislou na parametru θ
– mějme náhodný výběr $x = (x_1, x_2, \ldots, x_n)$ z rozdělení X
– v klasické (frekvenční) statistice předpokládáme, že θ je konstanta, kterou chceme odhadnout na základě výběru x
– Bayesovská statistika předpokládá, že θ je náhodná veličina s rozdělením/hustotou $g(.)$
– cílem je odhadnout parametry rozdělení g

– k dispozici máme apriorní rozdělení $\pi(\theta)$ – náš předpoklad, jak by se měla náhodná veličina chovat
– dále máme náhodný výběr x
– cílem je získat aposteriorní rozdělení $\pi(\theta|x)$

– pomocí Bayesovy věty lze odvodit, že: $\pi(\theta|x) \approx f(x|\theta)\pi(\theta)$

– nemáme-li žádnou představu o apriorním rozdělení, používá se neurčité apriorní rozdělení, tj. rozdělení konstantní
– odhad aposteriorního rozdělení se realizuje pomocí *Markov Chain Monte Carlo* (MCMC) metod