ANALÝZA DAT V R

5. ZÁKLADNÍ STATISTICKÉ TESTY

Mgr. Markéta Pavlíková
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

PRINCIPY STATISTICKÉ INFERENCE

- identifikace závisle proměnné
  - typ (spojitá, kategoriální, předpoklad rozdělení ...)
- identifikace kontextu / hypotézy
  - rozdíly ve skupinách = závislost na příslušnosti ke skupině?
  - závislost na hodnotě jiné proměnné? jakého typu je?
  - závislost na čase? opakovaná měření? ...
- identifikace parametru/principu
  - parametr polohy? rozptylu? kovariance (nezávislost)?
- výběr vhodné techniky

PŘEHLED TESTŮ

rozdělení	**normální**	**spojité**	**alternativní / diskrétní**
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2-test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2-test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

PŘEHLED TESTŮ V R

rozdělení	normální	spojité	alternativní / diskrétní
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	`t.test(x)`	`wilcox.test(x)` `prop.test(x,n,p = 0.5)`	`prop.test(x,n)`
výběr dvojic	`t.test(x,y, paired = T)`	`wilcox.test(x,y, paired = T)`	`mcnemar.test(x,y)`
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	`t.test(x,y)`	`wilcox.test(x,y)`	`fisher.test(x,y)` `chisq.test(x,y)`
k nezávislých výběrů (závislost na kategoriální proměnné)	`lm(y~x)`	`kruskal.test(x,group)`	`fisher.test(x,y)` `chisq.test(x,y)`
závislost na spojité proměnné	`lm(y~x)`	`glm(y~x,family= binomial,gaussian,Gamma, inverse.gaussian,poisson, quasi,...)` `ksmooth(x,y,...)`	`glm(y~x,family=binomial)`
opakovaná měření	`lme(y~x,random = ~1|ID)`	`glmer(y~x+(1|ID),family=...)`	

PŘEHLED TESTŮ V R: Y ZÁVISLÁ, X NEZÁVISLÁ

rozdělení	normální	spojité	alternativní / diskrétní
populační parametr (o čem je hypotéza)	**populační průměr**	**populační medián (distribuční funkce)**	**pravděpodobnost jevu / (ne)závislost / poměr šancí**
jeden výběr	`t.test(y)`	`wilcox.test(y)` `prop.test(k,n,p = 0.5)`	`prop.test(k,n)`
výběr dvojic	`t.test(y1,y2,paired = T)`	`wilcox.test(y1,y2, paired = T)`	`mcnemar.test(y1,y2)`
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	`t.test(y1,y2)`	`wilcox.test(y1,y2)`	`fisher.test(x,y)` `chisq.test(x,y)`
k nezávislých výběrů (závislost na kategoriální proměnné)	`lm(y~x)`	`kruskal.test(y,x)`	`fisher.test(x,y)` `chisq.test(x,y)`

závislost na spojité proměnné	`lm(y~x)`	`glm(y~x,family=`
		`binomial,gaussian,Gamma,`
		`inverse.gaussian,poisson,`
		`quasi,...)`
		`ksmooth(x,y,...)`
			`glm(y~x,family=binomial)`

opakovaná měření	`lme(y~x,random = ~1|ID)`	`glmer(y~x+(1|ID),family=...)`	

JEDNOVÝBĚROVÉ TESTY

Modelový příklad: výška desetiletých chlapců v roce 1961

- 130, 140, 136, 141, 139, 133, 149, 151, 139, 136, 138, 142, 127, 139, 147
- μ_{51} = 136.1

- t-test umíme z minula:

$$T = \frac{\bar{X} - \mu_0}{\text{S.E.}(\bar{X})} = \frac{\bar{X} - \mu_0}{S_x}\sqrt{n}$$

- znaménkový test (sign test)
  - kolik chlapců v roce 1961 je menších než populační průměr 1951?
  - pokud průměr stejný, mělo by to být půl napůl, pokud vyšší, bude malých méně
  - **130**, 140, **136**, 141, 139, **133**, 149, 151, 139, **136**, 138, 142, **127**, 139, 147
  - totéž jako udělat pro každého rozdíl mezi jeho výškou a μ_{51} a sečíst počet minusů
  - 5 z 15 = 33%, p = 0.20 (jednostranná), nezamítáme
  - oproti t-testu hodně hrubé, ale zase nepotřebuje jiný předpoklad než spojitost
  - pokud je někde 0, vyhodíme měření úplně

JEDNOVÝBĚROVÉ TESTY

- Wilcoxonův pořadový test (Wilcoxon signed rank test)
  - předpoklad spojité a **symetrické**
  - stejně jako u znaménkového: uděláme rozdíl mezi výškou každého chlapce a μ_{51}, ignorujeme nuly, zbude N_r pozorování
  - pamatujeme si znaménko (sgn)
  - seřadíme si hodnoty vzestupně a nahradíme je pořadím
  - tím přestane záležet na rozdělení jako takovém!!
  - testová statistika $$W = \sum_{i=1}^{N_r} [\mathrm{sgn}(x_i) \cdot R_i]$$
  - rozdělení poloha 0, kritická hodnota v tabulkách / počítač
  - pro $N_r \geq 10$ má Z-skór přibližně normální rozdělení, případně Yatesova korekce

$$z = \frac{W}{\sigma_W}, \sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$

PÁROVÉ TESTY

- Máme dvojice provázaných pozorování, jednotlivé dvojice na sobě nezávislé
  - výška otce a syna v dospělosti
  - měření jednoho metabolitu dvěma různými způsoby
  - měření před a po terapii
- Testujeme, zda se parametr polohy měření ve dvojici liší.
- Použijeme jednovýběrový test na rozdíl ve dvojici, testujeme, zda je poloha rozdílu 0.
  - t-test (průměr rozdílů je nula, normální nemusí být měření, stačí rozdíly)
  - znaménkový test (polovina rozdílů je záporných)
  - Wilcoxonův párový test (medián rozdílů je 0)

DVOUVÝBĚROVÉ TESTY

- N_X nezávislých pozorování X, N_Y nezávislých pozorování Y,
- výběry jsou také navzájem nezávislé (zajistí způsob pořízení dat)
- chceme
  - vědět, zda X a Y mají stejné rozdělení (ne nutně konkrétní)
  - vědět, zda je jedno rozdělení „větší“ než druhé (tzv. stochasticky dominantní, speciální případ je stejné s vyšším parametrem polohy) nebo stejné
  - má za předpokladu normálního rozdělení se stejným rozptylem různou (nebo stejnou) střední hodnotu

DVOUVÝBĚROVÉ TESTY

- dvouvýběrový t-test
  - předpoklad normálního rozdělení a stejného rozptylu
  - (pro velká N_X a N_Y není normalita tak podstatná)
  - testujeme shodnost $\mu_X = \mu_Y$
  - společný odhad rozptylu (vážený průměr odhadů z jednotlivých výběrů)

$$S^2 = \frac{n_X - 1}{n_X + n_Y - 2} S_X^2 + \frac{n_Y - 1}{n_X + n_Y - 2} S_Y^2$$

  - statistika (pro test hypotézy, že rozdělení X a Y jsou stejná)

$$T = \frac{\bar{X} - \bar{Y}}{\text{S.E.}(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_X n_Y}{n_X + n_Y}}$$

DVOUVÝBĚROVÉ TESTY

- dvouvýběrový t-test
  - $H_0 : \mu_X = \mu_Y$
    zamítnout ve prospěch alternativy
    - $H_1 : \mu_X \neq \mu_Y$ když $|T| \geq t_{n_X+n_Y-2}(1-\alpha/2)$
    - $H_1 : \mu_X > \mu_Y$ když $T \geq t_{n_X+n_Y-2}(1-\alpha)$
    - $H_1 : \mu_X < \mu_Y$ když $T \leq t_{n_X+n_Y-2}(\alpha) = -t_{n_X+n_Y-2}(1-\alpha)$

    [t.test(hosi,divky,var.equal=TRUE)] nebo
    [t.test(vyska∼Hoch,data=Vysky,var.equal=TRUE)]
  - zamítáme-li H_0, říkáme, že rozdíl výběrových průměrů **je (statisticky) významný**
  - pochyby o shodě rozptylů: Welchův test (modifikace t-testu)
    [t.test(hosi,divky,var.equal=FALSE)] (pro $\sigma_X \neq \sigma_Y$)
    [t.test(hosi,divky)] resp. [t.test(vyska∼Hoch)] (pro $\sigma_X \neq \sigma_Y$)
  - shodu rozptylů lze ověřit např. F-testem ($H_0 : \sigma_X = \sigma_Y$)
    [var.test(hosi,divky)]
  - ověření normality nutně pro každý výběr zvlášť!

DVOUVÝBĚROVÉ TESTY

- Wilcoxonův dvouvýběrový test (Mann-Whitneyův)

  (stačí **spojité rozdělení**)

  - dva nezávislé výběry rozsahu n_X, n_Y
  - spojitá rozdělení
  - H_0: rozdělení jsou stejná, tedy i **populační mediány** stejné
  - za H_0 jsou výběry „dobře promíchané“
  - určíme pořadí v rámci spojených výběrů
  - kritický obor: průměrná pořadí se příliš liší
  - W_X součet pořadí hodnot X

$$Z = \frac{W_X - n_X(n_X + n_Y + 1)/2}{\sqrt{n_X n_Y (n_X + n_Y + 1)/12}}$$

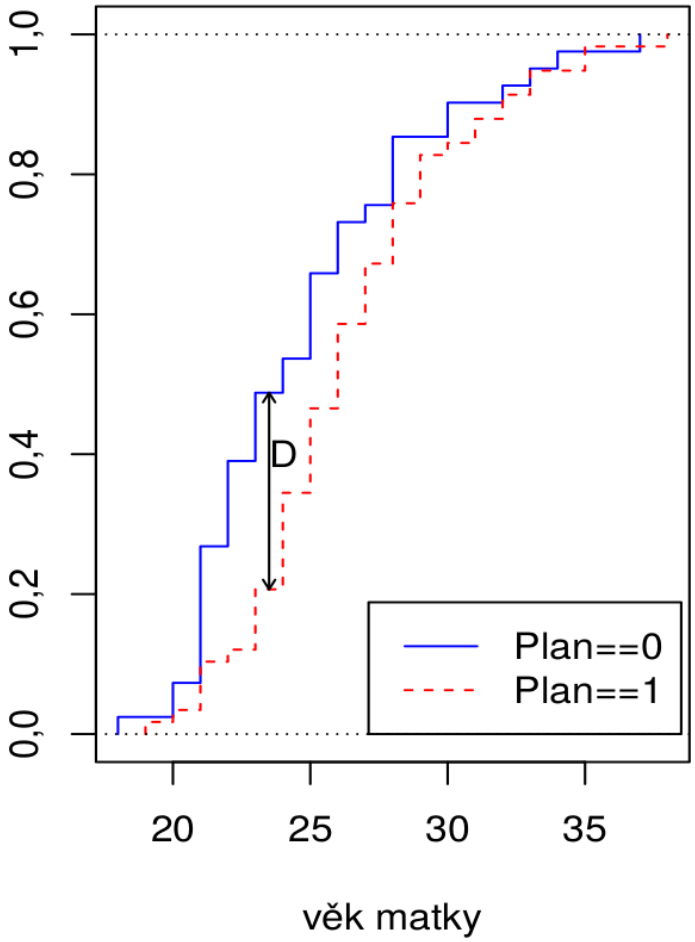
  - shodu zamítni, pokud $|Z| \geq z(1 - \alpha/2)$ (přibližný test)
  - citlivý vůči posunutí, méně vůči nestejné variabilitě

DVOUVÝBĚROVÉ TESTY

- Kolmogorov-Smirnovův test

(stačí **spojité rozdělení**)

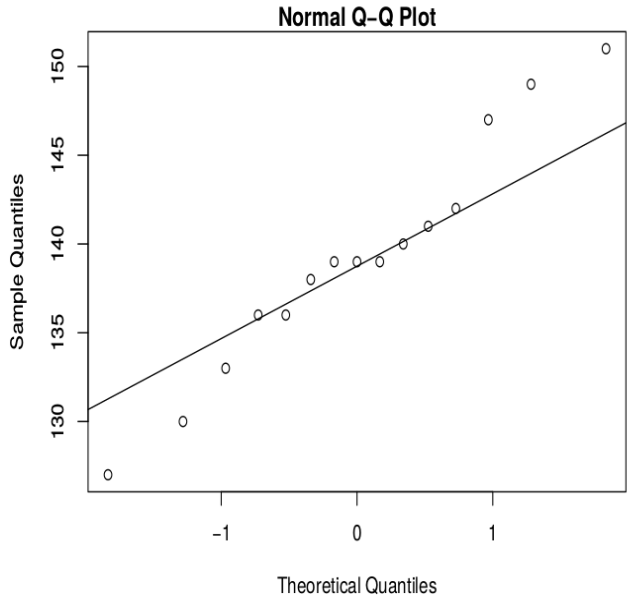
- porovná empirické distribuční funkce dvou **nezávislých** výběrů
- určí jejich největší „svislou“ vzdálenost
- citlivý vůči všem neshodám (nejen co do populačního průměru či populačního mediánu)
- porovnání věku matek podle plánovaného těhotenství
- $D = \frac{20}{41} - \frac{12}{58} = 0{,}2808$
  $p = 4{,}5\ \%$



GOODNES-OF-FIT TESTY

- Kolmogorov-Smirnov je typ tzv. „goodnes-of-fit“ testu
- Jak dobře odpovídají naše data nějakému jinému typu dat?
  - mají dvě sady dat stejné rozdělení? (tedy nejen třeba polohu)
  - mají data normální rozdělení?

    → Shapiro-Wilkův test



  - jsou pozorované četnosti stejné jako kdyby pocházely z nějakého konkrétního diskrétního rozdělení? → χ^2-test

GOODNES-OF-FIT TESTY

- χ^2-test
- principem testu je porovnání vzdálenosti mezi pozorovanou hodnotou a očekávanou hodnotou
- čím větší vzdálenost, tím nepravděpodobnější je, že pochází z očekávaného rozdělení

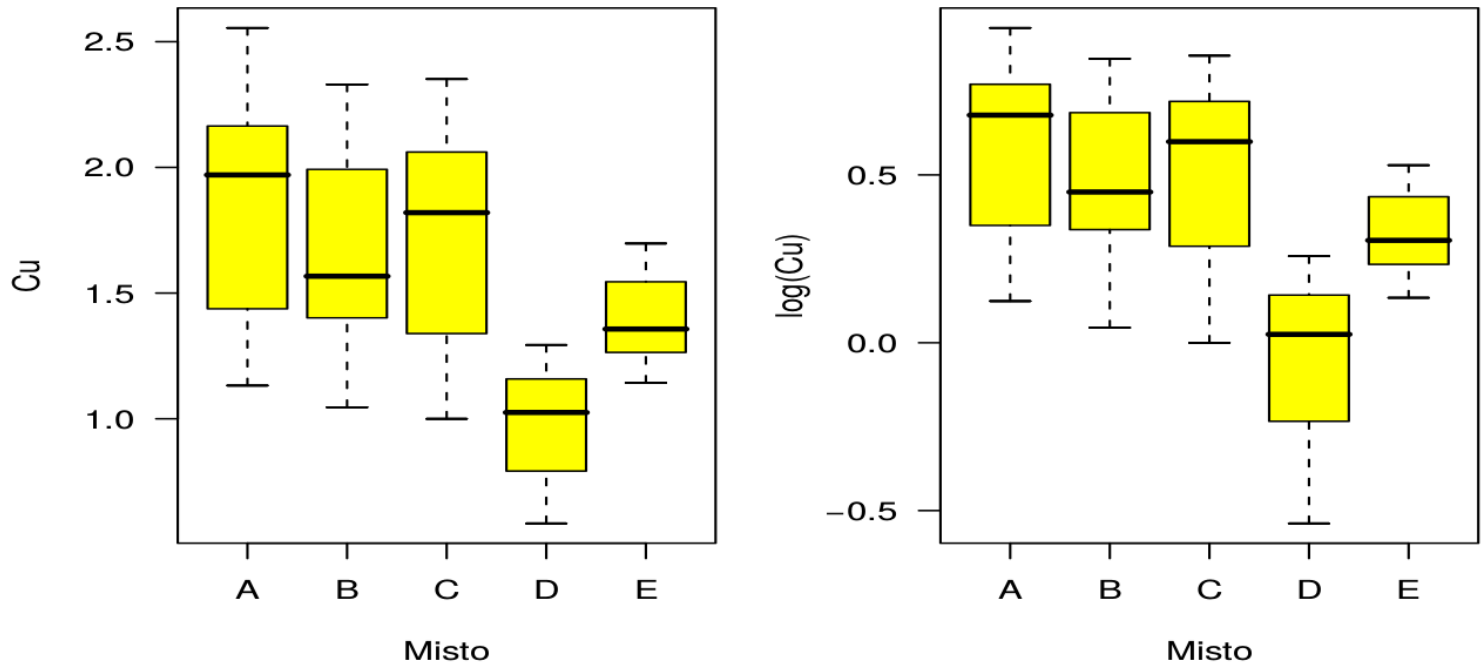
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- O = observed, E = expected, nahoře vzdálenost, dole „měřítko“
- má χ^2 rozdělení se stupni volnosti podle typu úlohy
- příklady:
  - test Hardy-Weinbergova ekvilibria
  - porovnání dvou křivek přežívání
  - test v kontingenční tabulce (probereme později)

POROVNÁNÍ VÍCE VÝBĚRŮ

motivační příklad pro analýzu rozptylu (játra):

- pět míst na řece, vždy vyloveno po 7 rybách
- zjišťována koncentrace mědi v játrech
- liší se tato místa svým znečištěním?
- logaritmování na pravé straně stabilizuje rozptyl



ANALÝZA ROZPTYLU

analýza rozptylu jednoduchého třídění (ANOVA)

- $Y_{11}, \ldots, Y_{1n_1} \sim \mathrm{N}(\mu_1, \sigma^2)$ (první výběr, průměr $\bar{Y}_{1\bullet}$)
  $Y_{21}, \ldots, Y_{2n_2} \sim \mathrm{N}(\mu_2, \sigma^2)$ (druhý výběr, průměr $\bar{Y}_{2\bullet}$)
  $\ldots$
  $Y_{k1}, \ldots, Y_{kn_k} \sim \mathrm{N}(\mu_k, \sigma^2)$ (k-tý výběr, průměr $\bar{Y}_{k\bullet}$)
- **nezávislé** výběry (shodné rozptyly, normální rozdělení)
- $\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad (= \mu)$ $\quad \mathrm{H}_1$: neplatí H_0
- celkový průměr $\bar{Y}_{\bullet\bullet}$, celkový rozsah $n = \sum_{i=1}^{k} n_i$
- rozklad součtu čtverců

$$\sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2$$

$$(\text{celková variabilita}) = (\text{variabilita mezi}) + (\text{variabilita uvnitř})$$

$$S_T = S_A + S_e$$

$$f_T = f_A + f_e$$

$$(n-1) = (k-1) + (n-k)$$

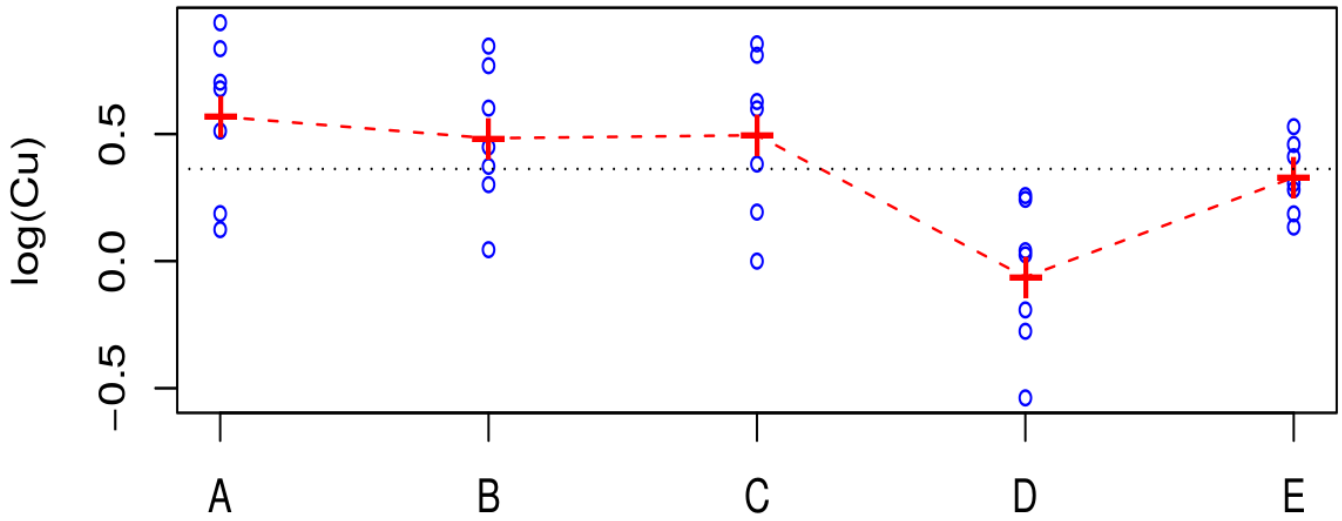
ANALÝZA ROZPTYLU

rozklad součtu čtverců

příklad játra (celkový průměr $\bar{y}_{\bullet\bullet} = 0{,}36$)

$$(\text{celková variabilita}) = (\text{variabilita mezi}) + (\text{variabilita uvnitř})$$

$$\sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{t=1}^{n_i}(Y_{it} - \bar{Y}_{i\bullet})^2$$



ANALÝZA ROZPTYLU

tabulka analýzy rozptylu

$$\mathrm{H}_0 \text{ zamítnout, je-li } F_A = \frac{S_A/f_A}{S_e/f_e} \geq F_{f_A,f_e}(1-\alpha)$$

variabilita	S	f	S/f	F	p
výběry	S_A	$f_A = k-1$	S_A/f_A	F_A	p_A
reziduální	S_e	$f_e = n-k$	S_e/f_e		
celková	S_T	$f_T = n-1$			

- S – součty čtverců, jejich rozklad
- f – počty stupňů volnosti
- S/f – průměrné čtverce
- F – F-statistika
- p – p-hodnota

ANALÝZA ROZPTYLU

příklad játra

variab.	S	f	S/f	F	p
místa	1,796	4	0,4490	5,862	0,0013
rezid.	2,285	30	0,0762		
celk.	4,081	34			

$F = 5{,}862 > F_{4,30}(0{,}95) = 2{,}690$

na 5% hladině jsme **prokázali rozdíl**

[summary(aov(lnCu∼Misto,data=Med))]

nebo také

[anova(lm(lnCu∼Misto,data=Med))]

ANALÝZA ROZPTYLU

ověření předpokladů

- **nezávislost**: dáno organizací (plánem) pokusu
  předpoklad nelze vynechat či nahradit
- **shoda rozptylů**: (vyvážený model je málo citlivý na neshodu populačních rozptylů)
  - Leveneův test
    (vlastně jednoduché třídění s $|Y_{it} - \mathrm{med}_t\, Y_{it}|$)
    $p = 64{,}8$ % [leveneTest(lnCu,Misto)]
  - Bartlettův test
    (citlivý na splnění předpokladu o normálním rozdělení)
    $p = 45{,}3$ % [bartlett.test(lnCu,Misto)]
- **normální rozdělení**: (vyvážený model je málo citlivý na nenormalitu), test normality nutno uplatnit na rezidua $Y_{it} - \bar{Y}_{i\bullet}$ (na všech n reziduí najednou) $p = 6{,}8$ %
  [shapiro.test(resid(aov(lnCu∼Misto)))]
  nebo [shapiro.test(resid(lm(lnCu∼Misto)))]

ANALÝZA ROZPTYLU

varianty zápisu modelu AR jednoduchého třídění

- **model** – idealizovaná představa o vzniku pozorované hodnoty
- měření = úroveň + náhodná „chyba“
  měření = systematická složka + náhodná složka

$$\begin{aligned} Y_{it} &= \mu_i + E_{it} \qquad 1 \le t \le n_i, \quad 1 \le i \le k \\ &= \mu + (\mu_i - \mu) + E_{it} \qquad E_{it} \text{ nezávislé} \\ &= \mu + \alpha_i + E_{it} \qquad E_{it} \sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

- **reparametrizace** (α_i – **efekty** faktoru A):

$$\sum_{i=1}^{k} \alpha_i = 0$$

- $\mathrm{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$ (totéž jako $\mu_1 = \mu_2 = \ldots = \mu_k$)
- pro $k = 2$ je $F_A = T^2$ (vztah s dvouvýběrovým t-testem)

ANALÝZA ROZPTYLU

mnohonásobná srovnání

(Tukeyův test, Kramerova verze, HSD – honest significance test)

- nutnost zachovat zvolenou hladinu testu i při současném rozhodování o řadě hypotéz
  (např. že $\mu_1 = \mu_2, \quad \mu_1 = \mu_3, \quad \mu_2 = \mu_3, \quad \ldots$)
- které dvojice úrovní faktoru (stř. hodnoty μ_i resp. efekty α_i) se liší?

$$|\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}| \geq q_{k,n-k}(1-\alpha)\sqrt{\frac{S^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

kde $q_{k,n-k}(1-\alpha)$ je tabelovaná kritická hodnota

$$S^2 = \frac{S_e}{f_e} = \frac{\sum\sum(Y_{it} - \bar{Y}_{i\bullet})^2}{n-k}$$

ANALÝZA ROZPTYLU

příklad játra

místo	počet	průměr	efekt	směr. odchylka
A	7	0,568	0,206	0,312
B	7	0,484	0,121	0,279
C	7	0,495	0,133	0,318
D	7	-0,063	-0,426	0,290
E	7	0,329	-0,034	0,144
celkem	35	0,363	0,000	0,104

$$q_{5,30}(0{,}95)\sqrt{\frac{0{,}0762}{2}\left(\frac{1}{7}+\frac{1}{7}\right)} = 4{,}10 \cdot 0{,}104 = 0{,}428$$

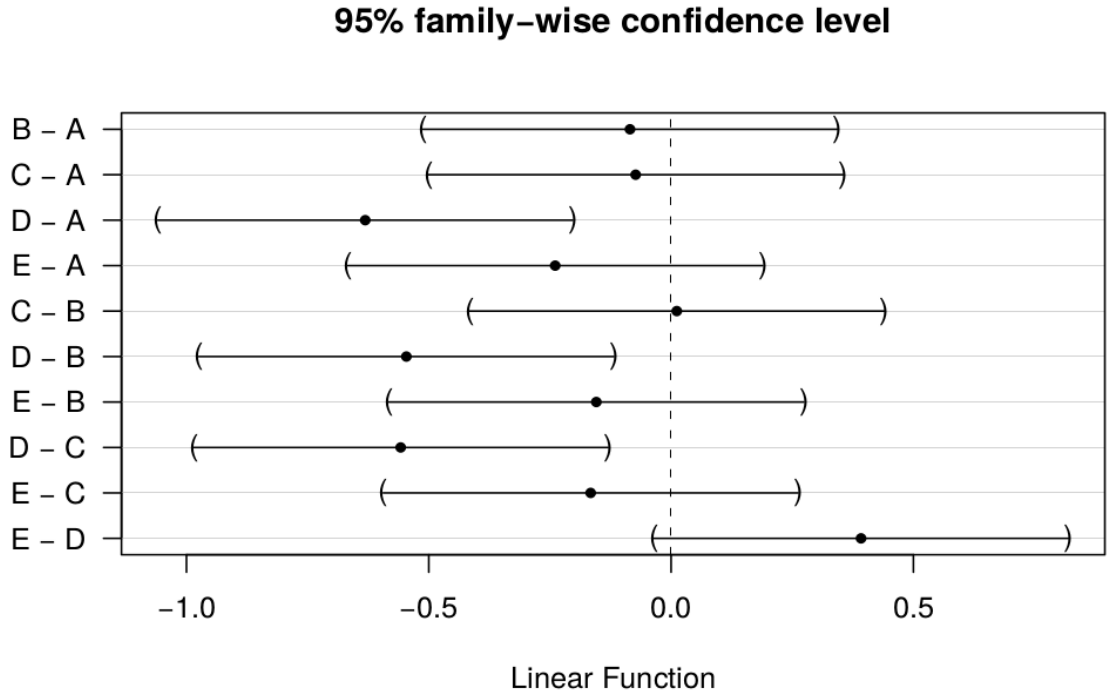
$-0{,}063 + 0{,}428 = 0{,}365 \Rightarrow$ na 5% hladině se místa D s nejmenším průměrem liší všechna místa s průměry aspoň 0,365, tedy místa A, B, C, nikoliv E

[TukeyHSD(aov(lnCu∼Misto,data=Med))]

ANALÝZA ROZPTYLU

příklad játra

funkce [TukeyHSD(aov(lnCu∼Misto,data=Med))]
dá tabulku porovnání všech dvojic
funkce [plot(TukeyHSD(aov(lnCu∼Misto,data=Med)))]
graficky znázorní porovnání všech dvojic
pomocí knihovny Rcmdr dostaneme také graf



ANALÝZA ROZPTYLU

Kruskalův-Wallisův text

(neparametrický test)

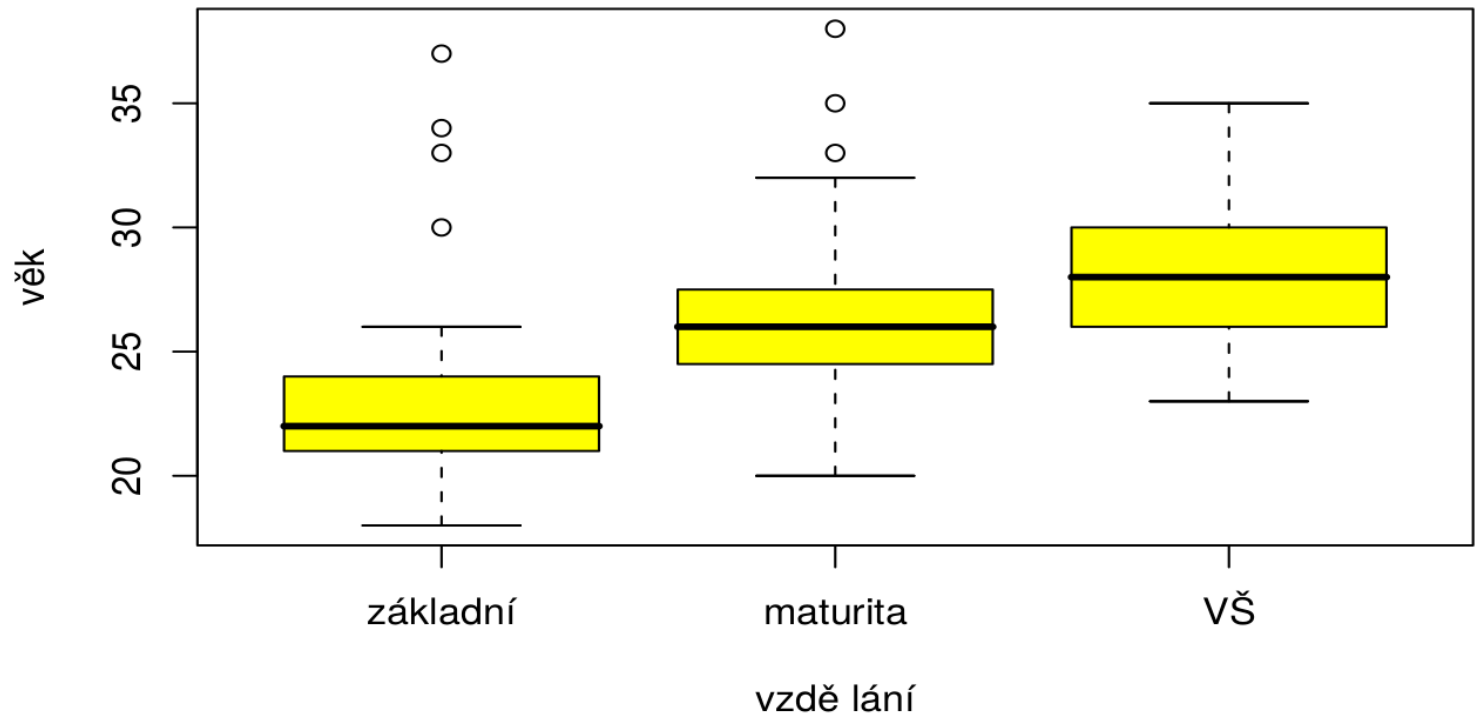
- zobecnění dvouvýběrového Wilcoxonova testu
  (použije opět pořadí místo původních hodnot)
- předpoklady:
  - k nezávislých výběrů
  - spojitá rozdělení
- H_0: rozdělení jsou stejná (tedy i mediány jsou stejné)
- T_i - součet pořadí v i-tém výběru

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

H_0 se zamítá při $Q \geq \chi^2_{k-1}(1-\alpha)$
(velká variabilita průměrných pořadí)

ANALÝZA ROZPTYLU

příklad kojení – věk matek podle vzdělání



je patrná nesymetrie, zejména u základního vzdělání

ANALÝZA ROZPTYLU

příklad kojení – věk matek podle vzdělání

vzdělání	n_i	průměrný věk	střední chyba	součet pořadí	průměrné pořadí
základní	34	23,412	0,638	1025	30,15
maturita	47	26,278	0,543	2618	55,70
VŠ	18	28,500	0,877	1307	72,61
celk.	99	25,697		4950	50,00

$$Q = \frac{12}{99 \cdot 100}\left(\frac{1025^2}{34} + \frac{2618^2}{47} + \frac{1307^2}{18}\right) - 3 \cdot 100 = 29{,}25$$

$$\chi^2_2(0{,}05) = 5{,}99 \qquad p < 0{,}0001$$

[kruskal.test(vek.m∼Vzdelani,data=Kojeni)]

(přesnější hodnocení přihlíží ke shodám při určování pořadí)

DĚKUJI ZA POZORNOST

www.biostatisticka.cz