ANALÝZA DAT V R

7. KONTINGENČNÍ TABULKA

Mgr. Markéta Pavlíková
Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

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

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

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	

HODNOCENÍ KVALITATIVNÍCH ZNAKŮ

- znaky v nominálním měřítku (neuspořádané hodnoty)
- existují techniky i pro ordinální měřítko: více struktury = přesněji zacílené testy
- příklady
  - počet osob s krevní skupinou A, B, AB, 0
  - počet dětí narozených v jednotlivých měsících v roce v Praze
  - vzdělání matky novorozence (ZŠ, SŠ, VŠ)
- statistické jednotky třídíme podle hodnoty znaku do k neslučitelných skupin
- výsledkem je k-tice (náhodný vektor) četností
- modelem pro tento vektor je multinomické rozdělení

MULTINOMICKÉ ROZDĚLENÍ

příklad: hod hrací kostkou, $k = 6$

- v dílčím pokusu k možných výsledků (jevů) $A_1, \ldots, A_k$
- $A_1, \ldots, A_k$ jsou neslučitelné jevy, sjednocení všech je jev jistý
- π_j je pst, že vyjde A_j $\qquad (\pi_1 + \pi_2 + \ldots + \pi_k = 1)$
- n **nezávislých** dílčích pokusů (n opakování)
- N_j – počet dílčích pokusů, kdy nastalo A_j
- $(N_1, \ldots, N_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$

důležitý předpoklad!

- platí pro velká n, např. pokud $n\pi_j \geq 5$ pro všechna j,

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j}$$

má přibližně rozdělení χ^2_{k-1}

HODNOCENÍ KVALITATIVNÍCH ZNAKŮ

Co nás zajímá?

- A. u jedné proměnné: jsou pravděpodobnosti výskytu jednotlivých kategorií takové, jaké očekáváme?
  - ekvivalent otázky na polohu spojitého rozdělení
  - test typu goodness-of-fit
- B. u dvou (nebo více) populací: jsou pravděpodobnosti výskytu jednotlivých kategorií u obou skupin stejné?
  - ekvivalent otázky na stejný parametr polohy (dvouvýběrový test)
  - test homogenity
- C. dvou proměnných (znaků): jsou na sobě závislé?
  - ekvivalent korelačního koeficientu nebo regrese
  - test nezávislosti

χ^2-TEST

- ve všech třech případech vede na χ^2-test
- porovnání vzdálenosti mezi pozorovanou hodnotou a očekávanou hodnotou

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

- **chí-kvadrát test dobré shody** $H_0 : \pi_1 = \pi_1^0, \ldots, \pi_k = \pi_k^0$ (pravděpodobnosti jsou hypotézou dány **jednoznačně**)
- platí-li H_0, očekáváme četnosti blízké hodnotám $E\,N_j = n\pi_j^0$:
- H_0 zamítáme, je-li $X^2 \geq \chi^2_{k-1}(1-\alpha)$,

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j^0)^2}{n\pi_j^0}$$

nezapomeň, že očekávané četnosti musí být alespoň 5

- N_j – **empirické** (experimentální) četnosti, $n\pi_j^0$ – **očekávané** (za platnosti H_0, teoretické) četnosti
- statistika X^2 (velké chí-kvadrát) porovnává empirické a očekávané četnosti (měří jejich neshodu, „vzdálenost“)

A. TEST DOBRÉ SHODY

příklad: reprezentativnost výběru

(**porovnat procenta** v populaci a výběru **nestačí**)

- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určuje H_0)
- **v průměru** očekáváme četnosti $200 \cdot 0{,}35 = 70 \qquad (70, 40, 20)$
- lze považovat tento výběr za **reprezentativní vzhledem k výskytu krevních skupin**?

$$\chi^2 = \frac{(56-70)^2}{70} + \frac{(72-70)^2}{70} + \frac{(54-40)^2}{40} + \frac{(18-20)^2}{20}$$
$$= 7{,}96 > 7{,}81 = \chi^2_3(0{,}95) \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní

```
chisq.test(x = c(56,72,54,18),p = c(0.35,0.35,.20,.10))
```

A. TEST DOBRÉ SHODY

příklad: reprezentativnost výběru

(**porovnat procenta** v populaci a výběru **nestačí**)

- ve vzorku pacientů byly počty osob s krevními skupinami 0, A, B a AB po řadě 56, 72, 54, 18 (tedy $n = 200$)
- ve vyšetřované populaci jsou krevní skupiny 0, A, B a AB v poměru 35 %, 35 %, 20 % a 10 % (to určuje H_0)
- **v průměru** očekáváme četnosti $200 \cdot 0{,}35 = 70$ $\quad (70, 40, 20)$
- lze považovat tento výběr za **reprezentativní vzhledem k výskytu krevních skupin**?

$$\chi^2 = \overset{2.8}{\frac{(56-70)^2}{70}} + \frac{(72-70)^2}{70} + \overset{4.9}{\frac{(54-40)^2}{40}} + \frac{(18-20)^2}{20}$$

$$= 7{,}96 > 7{,}81 = \chi^2_3(0{,}95) \qquad p = 4{,}7\ \%$$

- výběr **nelze** považovat za reprezentativní

```
chisq.test(x = c(56,72,54,18),p = c(0.35,0.35,.20,.10))
```

B. TEST HOMOGENITY

též test o shodnosti struktury pravděpodobností

- hodnoty znaku $B_1, \ldots, B_c$
- r **nezávislých** výběrů z různých populací
- H_0 : populace se **neliší** tedy jakoby ve skutečnosti pocházejí z jedné populace

- příklad **krevní skupiny**

populace	skupina				celkem
	0	A	B	AB	
C	121	120	79	33	353
D	118	95	121	30	364
celkem	239	215	200	63	717

B. TEST HOMOGENITY

též test o shodnosti struktury pravděpodobností

- hodnoty znaku $B_1, \ldots, B_c$
- r **nezávislých** výběrů z různých populací
- H_0 : populace se **neliší**

- příklad **krevní skupiny**

populace	skupina				celkem
	0	A	B	AB	
C	121	120	79	33	353
D	118	95	121	30	364
celkem	239	215	200	63	717

teoretická společná populace

$0.33 = \frac{239}{717}$ 0.30 0.28 0.09

četnosti v teoretické společné populaci

B. TEST HOMOGENITY

též test o shodnosti struktury pravděpodobností

- hodnoty znaku $B_1, \ldots, B_c$
- r **nezávislých** výběrů z různých populací
- H_0 : populace se **neliší**

- příklad **krevní skupiny**

populace	skupina				celkem
	0	A	B	AB	
C	121	120	79	33	353
D	118	95	121	30	364
celkem	239	215	200	63	717

$$\chi^2 = \frac{(121 - 353 \cdot 239/717)^2}{353 \cdot 239/717} + \ldots = 11{,}742 > \chi^2_3(0{,}95) = 7{,}815$$

splněn předpoklad použití testu

nejm. teoretická četnost: $353 \cdot 63/717 = 31{,}02 > 5$ $p = 0{,}8\ \%$

```
chisq.test(rbind(c(121,120,79,33),c(118,95,121,30)))
```

A1. HYPOTÉZA O STRUKTUŘE

- očekávané četnosti zatím vycházely z populačních dat nebo přímo z naměřených dat
- co když chceme testovat nějakou vlastnost, strukturu?
- příklad: Hardy-Weinbergova rovnováha
  - fenotypy AA, Aa, aa
  - předpoklad (který testujeme) je, že odpovídají rozložení genotypů za předpokladu nezávislého křížení a absence selekce
  - jaké očekáváme rozložení?
  - pravděpodobnosti parametrizujeme četností jedné z alel θ

$$\mathrm{P}(AA) \equiv \pi_1(\theta) = \theta^2$$
$$\mathrm{P}(Aa) \equiv \pi_2(\theta) = 2\theta(1-\theta)$$
$$\mathrm{P}(aa) \equiv \pi_3(\theta) = (1-\theta)^2$$

násobení pravděpodobností je ta nezávislost

METODA MAXIMÁLNÍ VĚROHODNOSTI

$$\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1! n_2! n_3!} (\theta^2)^{n_1} (2\theta(1-\theta))^{n_2} ((1-\theta)^2)^{n_3}$$

multinomické rozdělení

- najít θ takové, aby pravděpodobnost konkrétního výsledku byla maximální možná (maximálně věrohodná)

pravděpodobnost maximální = výběr nejspíš pochází z rozdělení s tímto θ a ne z jiného

- odhad θ maximalizací *logaritmické věrohodnostní funkce*

$$\ell(\theta) = \ln(\mathrm{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3))$$

maximum likelihood

$$= \ln\left(c_1 \left(\theta^2\right)^{n_1} (2\theta(1-\theta))^{n_2} \left((1-\theta)^2\right)^{n_3}\right)$$

- v našem příkladu vyjde

$$\hat{\theta} = \frac{2 \cdot N_1 + N_2}{2n}$$

2x četnost AA + 1x četnost Aa
děleno počtem všech alel

tedy počet alel A na počet všech „míst“ pro alely

A1. HYPOTÉZA O STRUKTUŘE

- θ má obecně q nezávislých složek, zde $q = 1$
  každý odhadovaný parametr ubere jeden stupeň volnosti
- H_0 zamítá pokud

$$X^2 = \sum_{j=1}^{k} \frac{(N_j - n\pi_j(\hat{\theta}))^2}{n\pi_j(\hat{\theta})} \geq \chi^2_{k-1-q}(1-\alpha)$$

- jsou zjištěné četnosti fenotypů $n_1 = 18$, $n_2 = 17$, $n_3 = 6$
  v souladu s modelem, tj. s H-W rovnováhou?
- v našem příkladu vyjde

$$\hat{\theta} = \frac{2 \cdot N_1 + N_2}{2n} \quad \left(= \frac{2 \cdot 18 + 17}{82} = 0{,}646 \right)$$

1 st. volnosti, nelze použít chisq.test, napsat nebo v balíku

- příklad: $\chi^2 = 0{,}355 < \chi^2_{3-1-1}(0{,}95) = 3{,}84$
  $p = 55{,}1\ \%$ hypotézu na 5% hladině nemůžeme zamítnout

C. TESTOVÁNÍ NEZÁVISLOSTI DVOU ZNAKŮ

nezávislost **nominálních** znaků

příklad hrách: barva květů a tvar pylových zrnek

- nominální znak s hodnotami $A_1, \ldots, A_r$ (tvar)
- nominální znak s hodnotami $B_1, \ldots, B_c$ (barva)
- N_{ij} kolikrát současně A_i a B_j (**sdružené četnosti**)
- **marginální** četnosti

řádky $N_{i\bullet} = \sum_{j=1}^{c} N_{ij}$ $N_{\bullet j} = \sum_{i=1}^{r} N_{ij}$ sloupce

- **nezávislost** znaků: pro všechny dvojice i, j platí

$$\mathrm{P}(A_i \cap B_j) = \mathrm{P}(A_i)\,\mathrm{P}(B_j)$$

- charakteristika nezávislosti: z **marginálních** pstí jevů A_i, B_j dokážeme rekonstruovat **sdružené psti** jevů $A_i \cap B_j$

C. TESTOVÁNÍ NEZÁVISLOSTI DVOU ZNAKŮ

test nezávislosti dvou kvalitativních znaků

hodnocení kontingenční tabulky

- H_0 : znaky jsou **nezávislé**

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - o_{ij})^2}{o_{ij}}$$

- teoretické četnosti (protějšek N_{ij}) –
  četnosti, které **v průměru očekáváme, platí-li hypotéza**

$$o_{ij} = n \cdot \widehat{P(A_i \cap B_j)} = n \cdot \widehat{P(A_i)} \cdot \widehat{P(B_j)} = n \cdot \frac{N_{i\bullet}}{n} \cdot \frac{N_{\bullet j}}{n} = \boxed{\frac{N_{i\bullet} N_{\bullet j}}{n}}$$

- nezávislost se zamítá pokud $X^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$
- stupně volnosti $n - 1 - q = r \cdot c - 1 - (r-1) - (c-1) = r \cdot c - r - c + 1 = \boxed{(r-1)(c-1)}$
- mělo by být $o_{ij} \geq 5 \ \forall \ (i,j)$ (tj. pro všechny dvojice)

C. TESTOVÁNÍ NEZÁVISLOSTI DVOU ZNAKŮ

příklad: kouření u mužů

data: lchs

empirické sdružené a marg. četnosti

vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	14	55	55	73	197
bývalý k.	11	28	44	42	125
kuřák	14	24	24	17	79
silný k.	78	189	175	106	548
celkem	117	296	298	238	949

očekávané sdružené a marg. četnosti

vzdělání	zákl.	odb.	mat.	VŠ	celk.
nekuřák	24,3	61,4	61,9	49,4	197
bývalý k.	15,4	39,0	39,3	31,3	125
kuřák	9,7	24,6	24,8	19,8	79
silný k.	67,6	170,9	172,1	137,4	548
celkem	117	296	298	238	949

[chisq.test(matrix(c(14,11,14,78, 55,28,24,189, 55,44,24,175, 73,42,17,106),nr=4,nc=4))]

závislost jsme na 5% hladině prokázali

$$\chi^2 = \frac{(14-24{,}3)^2}{24{,}3} + \cdots + \frac{(106-137{,}4)^2}{137{,}4} = 38{,}68$$

$$f = (4-1)(4-1) = 9$$

$$p < 0{,}0001$$

lze na to nahlížet i jako na „populaci“ nekuřáků atd.

D. TESTOVÁNÍ SYMETRIE („PÁROVOST“)

McNemarův test (test symetrie)

nezaměňovat s testem nezávislosti!

- **párový** test pro nominální veličinu s hodnotami $B_1, \ldots, B_k$
- zjišťujeme hodnoty nominálního znaku na **stejných** objektech za **dvojích** okolností (před ošetřením, po ošetření)
- N_{ij} počet objektů, u nichž první měření B_i a druhé měření B_j
- **nulová hypotéza**: pravděpodobnosti možných hodnot znaku jsou **stejné** za obojích okolností (před ošetřením i po něm)

$$X^2 = \sum_{i<j} \sum \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}}$$

- hypotézu zamítneme při $X^2 \geq \chi^2_{k(k-1)/2}(1 - \alpha)$
- výrazy ve jmenovateli musí být kladné!
- nezávisí na počtu objektů, kdy vyšly oba výsledky stejně (N_{ii})

D. TESTOVÁNÍ SYMETRIE („PÁROVOST“)

příklad **stromy**

1994	1995			celkem
	1	2	3	
1	4	3	3	10
2	7	21	11	39
3	1	15	35	51
celkem	12	39	49	100

- stav týchž stromů ve dvou sezónách
- celkem 100 stromů

$$\chi^2 = \frac{(3-7)^2}{3+7} + \frac{(3-1)^2}{3+1} + \frac{(11-15)^2}{11+15} = 3{,}215$$

- $\chi^2_3(0{,}95) = 7{,}8147, \quad p = 36{,}0\ \%$
- rozdíl mezi sezónami jsme neprokázali
- [mcnemar.test(matrix(c(4,7,1,3,21,15,3,11,35),3,3))]

2x2 TABULKA

čtyřpolní tabulka (tabulka 2×2)

znovu test nezávislosti či homogenity

a c	b d	$a+b$ $c+d$
$a+c$	$b+d$	n

- speciální případ kontingenční tabulky pro $r = c = 2$
- test nezávislosti i test homogenity
  statistiku lze upravit na pohodlnější vyjádření

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

zamítá se pro $X^2 \geq \chi^2_1(1-\alpha) \qquad (= z(1-\alpha/2)^2)$

2x2 TABULKA

- je-li některá očekávaná četnost malá, pak lze u čtyřpolní tabulky použít upravený postup: **Yatesova korekce**

$$X^2_Y = \frac{n(|ad - bc| - n/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

- **Fisherův exaktní** test počítá přímo dosaženou hladinu p
- pro tabulku s velkými četnostmi je výpočet Fisherova testu výpočetně náročný (paměťové nároky, trvání výpočtu)
- existuje zobecnění Fisherova testu i pro větší tabulky, než je čtyřpolní

Princip Fisherova testu: pravděpodobnost vzniku konkrétní tabulky při pevných marginálních četnostech

$$\frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{n! * a! * b! * c! * d!}$$

a c	b d	$a+b$ $c+d$
$a+c$	$b+d$	n

FISHERŮV EXAKTNÍ TEST

pravděpodobnost realizace tabulky II.:

$$\frac{9! * 9! * 13! * 5!}{18! * 1! * 8! * 4! * 5!} = \frac{136080}{1028160} = 0.132$$

I. p = 0.0147

0	9	9
5	4	9
5	13	18

II. p = 0.1320

1	8	9
4	5	9
5	13	18

III. p = 0.3530

2	7	9
3	6	9
5	13	18

IV. p = 0.3530

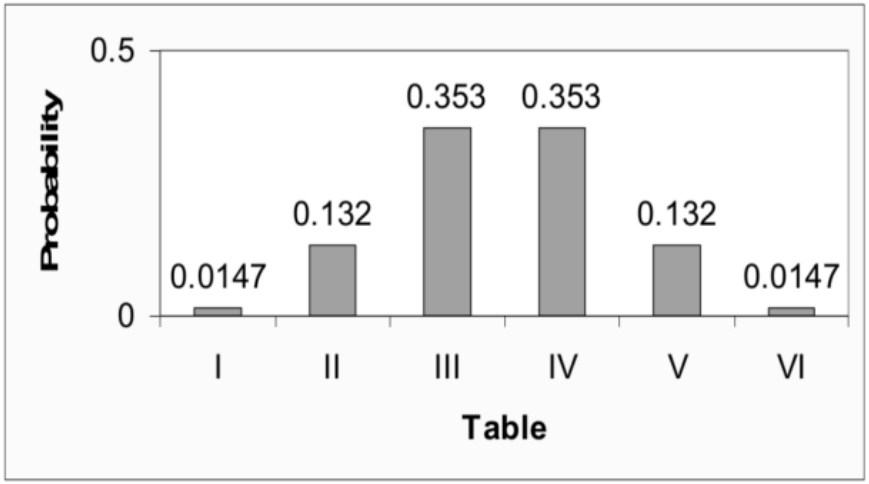
3	6	9
2	7	9
5	13	18

V. p = 0.1320

4	5	9
1	8	9
5	13	18

IV. p = 0.0147

5	4	9
0	9	9
5	13	18



pozorovaná tabulka má pst = 0.132

"extrémy" na obě strany jsou I, II, V a VI

$p_I + p_{II} + p_V + p_{VI} = 0.293$

nezamítáme

RR = RISK RATIO = PODÍL RIZIK

	20 a více odběrů	10 a méně odběrů
onkologické onemocnění	a	c
nemá onkol. onemocnění	b	d
celkem	a+b	c+d

$R_{\geq 20}$ = riziko o.o., pokud 20 a více odběrů = a / (a+b)

$R_{\leq 10}$ = riziko o.o, pokud 10 a méně odběrů = c / (c+d)

RR = risk ratio = podíl rizik = $R_{\geq 20} / R_{\leq 10}$

RR < 1 ... riziko je menší pro osoby s 20 a více odběry

RR > 1 ... riziko je větší pro osoby s 20 a více odběry

- data slouží k odhadu skrytého „skutečného“ RR
- odhad je vždy zatížený možností náhody = je tu možnost chyby

RR = RISK RATIO = PODÍL RIZIK

	20 a více odběrů	10 a méně odběrů
onkologické onemocnění	31	14
nemá onkol. onemocnění	236	251
celkem	267	265

$R_{\geq 20}$ = riziko o.o., pokud 20 a více odběrů = 31 / 267 = 11.6 %

$R_{\leq 10}$ = riziko o.o, pokud 10 a méně odběrů = 14 / 265 = 5.3 %

RR = risk ratio = $R_{\geq 20}$ / $R_{\leq 10}$ = 2.2

RR > 1 ... riziko je větší pro osoby s 20 a více odběry

RR < 1 ... riziko je menší pro osoby s 20 a více odběry

95 % konfidenční interval 1.2 – 4.0, p-hodnota = 0.008

- data slouží k odhadu skrytého „skutečného“ RR
- odhad je vždy zatížený možností náhody = je tu možnost chyby

OR = ODDS RATIO = PODÍL ŠANCÍ

	20 a více odběrů	10 a méně odběrů
onkologické onemocnění	a	c
nemá onkol. onemocnění	b	d
celkem	a+b	c+d

$O_{\geq 20}$ = šanci mít o.o. při 20 a více odběrech = a : b (tedy a / b)

$O_{\leq 10}$ = šanci mít o.o. při 10 a méně odběrech = c : d (tedy c / d)

OR = odds ratio = podíl šancí = $O_{\geq 20} / O_{\leq 10}$ = ad / bc

OR < 1 ... šance je menší pro osoby s 20 a více odběry

OR > 1 ... šance je větší pro osoby s 20 a více odběry

- pro vzácné jevy je OR dobré přiblížení RR
- vhodný vždy když nemáme výběr z populace a při modelování

OR = ODDS RATIO = PODÍL ŠANCÍ

	20 a více odběrů	10 a méně odběrů
onkologické onemocnění	31	14
nemá onkol. onemocnění	236	251
celkem	267	265

$R_{\geq 20}$ = riziko o.o., pokud 20 a více odběrů = 31 / 236 = 0.13136

$R_{\leq 10}$ = riziko o.o, pokud 10 a méně odběrů = 14 / 251 = 0.05577

RR = risk ratio = $R_{\geq 20}$ / $R_{\leq 10}$ = 2.355

RR > 1 ... riziko je větší pro osoby s 20 a více odběry

RR < 1 ... riziko je menší pro osoby s 20 a více odběry

95 % konfidenční interval 1.2 – 4.0, p-hodnota = 0.008

- data slouží k odhadu skrytého „skutečného“ RR
- odhad je vždy zatížený možností náhody = je tu možnost chyby

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	

OR = ODDS RATIO = PODÍL ŠANCÍ

	má vlastnost	nemá vlastnost
má onemocnění	a	c
nemá onemocnění	b	d
celkem	a+b	c+d

$O_{má}$ = šance mít o. když má vlastnost = a : b (tedy a / b)

$O_{nemá}$ = šance mít o. když nemá vlastnost = c : d (tedy c / d)

OR = odds ratio = podíl šancí = $O_{má} / O_{nemá}$ = ad / bc

OR < 1 ... šance je menší pro osoby, které mají podmínku

OR > 1 ... šance je větší pro osoby, které mají podmínku

- pro vzácné jevy je OR dobré přiblížení RR
- vhodný vždy když nemáme výběr z populace a při modelování

OD OR K LOGISTICKÉ REGRESI

Y / X	má vlastnost	nemá vlastnost
má onemocnění	$P(Y = 1, X = 1) = \pi_1$	$P(Y = 1, X = 0) = \pi_0$
nemá onemocnění	$P(Y = 0, X = 1) = 1 - \pi_1$	$P(Y = 0, X = 0) = 1 - \pi_0$
celkem	$P(X = 1)$	$P(X = 0)$

$O_{Y \mid X = 1}$ = šance mít o. když má vlastnost = $\pi_1 / (1 - \pi_1)$

$O_{Y \mid X = 0}$ = šance mít o. když nemá vlastnost = $\pi_0 / (1 - \pi_0)$

- $\pi_i / (1 - \pi_i)$ nabývá hodnot mezi 0 a $+\infty$
- předpokládáme, že není π_i nulová
- $\log(\pi_i / (1 - \pi_i))$ nabývá hodnot mezi $-\infty$ a $+\infty$
- to už se dá dobře modelovat nějakou křivkou!

$$OR = \frac{\frac{\pi_1}{1 - \pi_1}}{\frac{\pi_0}{1 - \pi_0}}$$

LOGISTICKÁ REGRESE

Y je binární proměnná, Y_1, .., Y_n nezávislé, má binomiální rozdělení $B(n_i,\pi_i)$

$$\begin{aligned}\text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right)\\ &= \beta_0 + \beta_1 x_i\\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}\end{aligned}$$

nebo také

$$\pi_i = Pr(Y_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

potom

$$\begin{aligned}\log(\text{OR}) = \log\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right) &= \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right)\\ &= \beta_0 + \beta_1 x_1 - (\beta_0 + \beta_1 x_0)\\ &= \beta_1 (x_1 - x_0)\end{aligned}$$

LOGISTICKÁ REGRESE

- X může být binární, kategoriální, faktor nebo spojitá
- z modelu logistické regrese vypadávají přímo log(OR) v podobě koeficientů β
- hodnota β vyjadřuje podíl šancí pro osobu s hodnotou vysvětlující proměnné $x+1$ oproti osobě s x (za podmínky, že všechno ostatní je stejné, jako v obvyklé regresi)
- $\beta < 0$ odpovídá $OR < 1$, $\beta > 0$ odpovídá $OR > 1$
- konfidenční, intervaly, testy nulovosti β, ...
- odhady optimalizací podle metody maximální věrohodnosti

LOGISTICKÁ REGRESE - PŘÍKLAD

- plánování těhotenství různí charakteristiky matky
  - primipara
  - věk rodičky
  - vzdělání rodičky
- podrobnější ukázky viz cvičení 7a.R

DĚKUJI ZA POZORNOST

www.biostatisticka.cz