Poslední úprava dokumentu: 17. dubna 2024.

Analýza kategoriálních dat (χ^2-testy dobré shody)

Prostudujte si základní pojmy a testy v multinomickému rozdělení a kontingenčních tabulkách. Využít můžete materiály z přednášky, nebo text *Analýza kategoriálních dat* (dále jako *AKD*), který najdete na mých stránkách.

1 Shoda s multinomickým rozdělením

Rozhodněte, zda četnosti 95, 169, 89 odpovídají ideálnímu genotypovému štěpnému poměru 1:2:1. Hladinu testu uvažuje 5 %.

Vektor (Y_1, Y_2, Y_3) představující četnosti jednotlivých genotypů ve skupině n jedinců má multinomické rozdělení s parametry n, π_1, π_2, π_3. Formálně zapsáno: $(Y_1, Y_2, Y_3) \sim M(n, (\pi_1, \pi_2, \pi_3))$. My jsme celkem zkoumali $n = 95 + 169 + 89 = 353$ jedinců a naměřili jsme četnosti $n_1 = 95, n_2 = 169, n_3 = 89$.

Nyní bychom rádi otestovali hypotézu, zda tyto četnosti odpovídají teoretickému poměru $1:2:1$, to jest, jestli se jednotlivé genotypy realizují s pravděpodobnostmi $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. Chceme testovat hypotézu:

$$H_0 : (\pi_1, \pi_2, \pi_3) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$
$$\text{proti: } H_1 : (\pi_1, \pi_2, \pi_3) \neq \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right).$$

1) K testu samozřejmě použijeme χ^2 test dobré shody (viz text *AKD*, sekce 1.1):

```
(cetnosti <- c(95, 169, 89))          # pozorovane cetnosti
(prpsti <- c(1, 2, 1)/4)              # hypoteticke pravdepodobnosti
chisq.test(cetnosti, p=prpsti)        # test dobre shody
```

- Z výstupu se dozvídáme, že:
  - hodnota testové statistiky je $\mathsf{X}^2 = 0.84136$ (viz *AKD*, vzorec (6))
  - df $= 2$ (počet stupňů volnosti je $k - 1 = 3 - 1 = 2$, přičemž $k =$ počet kategorií)
  - p-hodnota $= 0.6566$
- Jaký je závěr?
  P-hodnota je větší než zvolená hladina 0.05, nezamítáme tedy nulovou hypotézu, že četnosti jednotlivých genotypů odpovídají teoretickému poměru 1:2:1.
- Jak byste o nulové hypotéze rozhodli v případě, kdy byste měli k dispozici pouze hodnotu testové statistiky a statistické tabulky?
  Kritickou hodnotu tohoto testu najdeme v tabulkách (viz např. materiály k 8. cvičení), kde je označena jako $\chi^2_n(\alpha) = \chi^2_2(0.05) = 5.99$. Tato kritická hodnota je samozřejmě rovna 95% kvantilu χ^2-rozdělení se 2 stupni volnosti, tj. $q\chi^2_2(0.95)$, což si lze snadno ověřit příkazem

```
qchisq(0.95,2)        # 95% kvantil chi-kvadrat rozdeleni
```

Jelikož hodnota testové statistiky 0.84136 není větší než kritická hodnota 5.99, tak nezamítáme nulovou hypotézu.

2) Nesmíme zapomenout, že použitý χ^2-test dobré shody funguje dobře pouze pro dostatečně velké očekávané četnosti. Musíme tedy zkontrolovat, že všechny očekávané četnosti jsou větší nebo rovny 5.

```
(ex <- chisq.test(cetnosti, p=prpsti)$expected)     # očekávané četnosti
```

✧ Jsou všechny očekávané četnosti dost velké? Ano, všechny očekávané četnosti jsou ≥ 5.

3) Očekávané četnosti lze získat také jednoduchým příkazem:

```
(ex <- prpsti * sum(cetnosti))
```

Nicméně předchozí konstrukce se nám bude hodit i u testů dobré shody v kontingenčních tabulkách.

4) V rámci procvičení programování v R si můžeme zkusit testovou statistiku a p-hodnotu spočítat „ručně“:

```
(Xs <- (cetnosti - ex)^2 / ex)     # komponenty testové statistiky
(X <- sum(Xs))                     # testová statistika
1 - pchisq(X, df=2)                # p-hodnota
```

5) Rozhodněte, zda lze považovat za reprezentativní vzorek dospělých žen, v němž je

180	žen svobodných,
239	žen vdaných,
75	žen rozvedených,
4	ženy ovdovělé,

když v odpovídající věkové populaci jsou skutečné podíly žen rovny po řadě 34,27 %, 52,02 %, 12,50 % a 1,21 %.

2 Analýza obecné kontingenční tabulky

Uvažujme následující kontingenční tabulku, jež udává počty novomanželských párů s jednotlivými kombinacemi vzdělání ženicha a nevěsty získáné v jistém období na nejmenované radnici.

	Nevěsta		
Ženich	základní	maturita	VŠ
základní	24	12	3
maturita	7	24	3
VŠ	3	9	15

Test nezávislosti (v zadané kontingenční tabulce)

1) Máme tedy dva znaky: X = vzdělání nevěsty, Y = vzdělání ženicha. Jako první nás zajímá, zda jsou tyto veličiny závislé.

$$H_0: \; X \text{ a } Y \text{ jsou nezávislé}$$
$$H_1: \; X \text{ a } Y \text{ jsou závislé}$$

K tomuto účelu lze použít χ^2-test nezávislosti, případně Fisherův test. Hladinu testu budeme uvažovat 5 %.

✧ Nejprve si musíme tabulku výše zadat do R:

```
TAB <- matrix(c(24,7,3, 12,24,9, 3,3,15), nrow=3)     # vytvoreni matice cisel
rownames(TAB) <- c("zakladni", "maturita", "VS")      # pojmenovani radku
colnames(TAB) <- c("zakladni", "maturita", "VS")      # pojmenovani sloupcu
print(TAB)
```

✧ Nyní provedeme χ^2-test nezávislosti (viz text *AKD*, sekce 2.1)

```
chisq.test(TAB)                    # chi^2 test
```

✧ Testová statistika vyšla $\mathsf{X}^2 = 43.219$ a p-hodnota je $9.32 \cdot 10^{-9}$. Jelikož je p-hodnota mnohem menší než hladina 0.05, zamítáme H_0. Na hladině 5 % jsme tedy prokázali, že vzdělání snoubenců jsou závislá.

2) Jsou všechny očekávané četnosti dostatečně velké?

```
chisq.test(TAB)$expected          # ocekavane cetnosti pri nezavislosti
```

Je to jen tak tak, ale všechny očekávané četnosti v tabulce jsou ≥ 5. Rozdělení χ^2_4 by tedy mělo poměrně dobře aproximovat rozdělení testové statistiky, a test by měl tudíž pro naše data fungovat dobře.

3) Kdybychom pro rozhodování měli k dispozici pouze statistické tabulky, museli bychom si najít příslušnou kritickou hodnotu. Naše tabulka má rozměry $I \times J$, kde $I = J = 3$. Víme, že testová statistika má asymptoticky χ^2 rozdělení se stupni volnosti $df = (I-1)(J-1) = 2 \cdot 2 = 4$. Najdeme si tedy kritickou hodnotu $\chi^2_4(\alpha) = 9.49$, která odpovídá kvantilu

```
qchisq(0.95, df=4)
```

Jelikož testová statistika $\mathsf{X}^2 = 43.219$ překračuje tuto kritickou hodnotu, docházíme ke stejnému závěru jako pomocí p-hodnoty, a to zamítnout nezávislost veličin.

4) Všechny očekávané četnosti v tabulce jsou sice větší než 5, ale někdy jenom těsně. Pokud bychom se v tomto případě nechtěli spoléhat na asymptotiku χ^2-testu, můžeme použít Fisherův faktoriálový test (viz *AKD*, sekce 2.2), který je přesný a není založen na žádné asymptotice.

```
fisher.test(TAB)                  # Fisheruv test
```

✧ Tento test nepočítá hodnotu žádné testové statistiky, ale z pravděpodobnosti realizace naší konkrétní tabulky při platnosti nulové hypotézy počítá přímo p-hodnotu. Ta v tomto případě činí $5.472 \cdot 10^{-8}$. Je tedy menší než zvolená hladina 0.05, a docházíme tedy ke stejnému závěru jako pomocí χ^2-testu. Na hladině 5 % zamítáme hypotézu, že vzdělání snoubenců jsou nezávislá.

Test symetrie

5) Lze se též ptát, zda je sdružené rozdělení vzdělání ženicha a nevěsty symetrické. Použijeme-li opět naše značení X = vzdělání nevěsty, Y = vzdělání ženicha, pak symetrie rozdělení (X, Y) odpovídá testu hypotézy

$$H_0 : P(X = i \ \& \ Y = j) = P(X = j \ \& \ Y = i) \quad \text{pro všechny dvojice } (i, j)$$
$$H_1 : P(X = i \ \& \ Y = j) \neq P(X = j \ \& \ Y = i) \quad \text{pro některou dvojici } (i, j)$$

kde $i, j = 1$ (ZŠ), 2 (maturita), 3 (VŠ). K tomu se použije Bowkerův test symetrie (viz text *AKD*, sekce 2.3), který je v R nazýván McNemarův:

```
mcnemar.test(TAB)
```

Hodnota testové statistiky (AKD, vzorec (9)) vyšla 4.3158, p-hodnota je 0.2293. Pro úplnost dodejme, že náš počet kategorií vzdělání je $I = 3$, a tudíž počet stupňů volnosti příslušného χ^2 rozdělení je $I(I-1)/2 = 3 \cdot 2/2 = 3$, což je taktéž uvedeno ve výstupu pod označením `df` (= degrees of freedom).

- ✧ Jaký je závěr? Hladinu testu uvažujte 5 %.
  P-hodnota je vyšší než zvolená hladina 0.05, tudíž nezamítáme, že rozdělení vzdělání snoubenců je symetrické.
- ✧ To, že testová statistika `McNemar's chi-squared` z výstupu souhlasí s teoretickým vzorcem (9) z textu AKD si můžeme ověřit ručním výpočtem:

  ```
  (12-7)^2/(12+7) + (3-3)^2/(3+3) + (3-9)^2/(3+9)
  ```
- ✧ Pokud bychom neměli k dispozici p-hodnotu, ale pouze hodnotu testové statistiky, museli bychom spočítat kritickou hodnotu. V tabulkách snadno najdeme, že kritická hodnota $\chi^2_{I(I-1)/2}(\alpha) = \chi^2_3(0.05)$ je 7.81, což si lze ověřit i výpočtem odpovídajícího kvantilu $q\chi^2_3(0.95)$

  ```
  qchisq(0.95, df=3*2/2)
  ```

  Jelikož $Q = 4.3158 < 7.81 = \chi^2_3(0.05)$, tedy hodnota testové statistiky nepřekročila kritickou hodnotu, nezamítáme naši nulovou hypotézu, že rozdělení vektoru (X, Y) je symetrické. (Což je tedy stejný závěr jako pomocí p-hodnoty).

Test nezávislosti (včetně přípravy kontingenční tabulky)

Nyní se vrátíme k datům `Kojeni`. Načtěme si dříve uložená data do RStudia.

```
load("Kojeni.RData")
Kojeni$fOtec <- factor(Kojeni$Otec, labels=c("ne", "ano"))
Kojeni$fDudlik <- factor(Kojeni$Dudlik, labels=c("ne", "ano"))
Kojeni$fPlan <- factor(Kojeni$Plan, labels=c("ne", "ano"))
save(Kojeni, file="Kojeni.RData")
attach(Kojeni)
```

6) Souvisí přítomnost otce u porodu (proměnná Otec, resp. fOtec) se vzděláním matky (Vzdelani)? To jest, jsou tyto veličiny závislé? Pokuste se sami interpretovat výsledky následujících příkazů.

```
(TAB <- table(Vzdelani, fOtec))     # kontingencni tabulka
(PTAB <- prop.table(TAB, margin=1) * 100)        # radkove proporce (v %)
chisq.test(TAB)                     # chi^2 test dobré shody
chisq.test(TAB)$expected            # ocekavane cetnosti pri nezavislosti
fisher.test(TAB)                    # Fisheruv test
```

3 Test nezávislosti ve čtyřpolní tabulce

Liší se podíl dětí, které dostaly dudlík (proměnná Dudlik, resp. fDudlik) mezi dětmi „plánovanými“ a „neplánovanými“ (Plan, resp. fPlan)? Jinými slovy - jsou veličiny Dudlik a Plan závislé? Na hladině $\alpha = 0.05$ budeme testovat

$$H_0 : \text{veličiny } \texttt{Dudlik} \text{ a } \texttt{Plan} \text{ jsou nezávislé}$$
$$\text{proti } H_1 : \text{veličiny } \texttt{Dudlik} \text{ a } \texttt{Plan} \text{ jsou závislé .}$$

Veličiny Dudlik a Plan nabývají obě hodnot 0 (= ne) a 1 (= ano). Příslušná kontigenční tabulka má následující tvar:

		fDudlik						fDudlik		
		ne	ano					ne	ano	
fPlan	ne	n_{11}	n_{12}	$n_{1\cdot}$	=	fPlan	ne	9	32	41
	ano	n_{21}	n_{22}	$n_{2\cdot}$			ano	14	44	58
		$n_{\cdot 1}$	$n_{\cdot 2}$	n				23	76	99

a v R si ji snadno vytvoříme příkazem

```
(TAB <- table(fPlan, fDudlik))     # konting. tabulka
```

Pro zajímavost si můžeme spočítat i řádková procenta

```
(PTAB <- prop.table(TAB, margin=1) * 100)       # radkove proporce (v %)
```

Jsou-li obě veličiny nezávislé, měly by být (v ideálním případě) oba řádky shodné.

1) První možností, jak přistoupit k testu nulové hypotézy o nezávislosti, je pomocí χ^2-testu (viz text *AKD*, sekce 4.2).

```
chisq.test(TAB, correct=FALSE)     # chi^2 test (bez Yatesovy korekce)
```

- ✧ Testová statistika (*AKD*, vzorec (15)) má hodnotu $\mathsf{X}^2 = 0.0644$ a příslušná p-hodnota vyšla 0.7997. Jelikož $0.7997 > 0.05$, tak nezamítáme nulovou hypotézu, že veličiny jsou nezávislé. Data tedy neprokázala, že by plánovanost dítěte měla nějakou souvislost s používáním dudlíku.
- ✧ Očekávané četnosti jsou

  ```
  chisq.test(TAB)$expected
  ```

  a jelikož jsou všechny větší než 5, měla by být aproximace rozdělení statistiky X^2 rozdělením χ^2_1 dobrá. Tudíž použití χ^2-testu nezávislost bylo v pořádku.
- ✧ Nicméně můžeme si vyzkoušet i použití χ^2-testu s Yatesovou korekcí, která je vhodná zejména v případě malých četností:

  ```
  chisq.test(TAB)                  # chi^2 test s Yatesovou korekci
  ```

  Hodnota testové statistiky i p-hodnota se samozřejmě trochu změnily, ale náš závěr by byl stejný.

2) Další možností je použít Fisherův faktoriálový test (viz text *AKD*, sekce 4.3).

```
fisher.test(TAB)
```

- ✧ P-hodnota vyšla 1, a tudíž ani Fisherův test nezamítá nulovou hypotézu o nezávislosti. (Ona to není úplně čistá 1, ale je to číslo tak blízké 1, že R při zaokrouhlování výstupu to již uvádí jako 1.)
- ✧ Fisherův faktoriálový test je spjat s pojmem **podílu šancí** (viz *AKD*, vzorec (16)), který má pro naši tabulku tvar

$$\beta = \frac{\text{šance nemít dudlík mezi neplánovanými dětmi}}{\text{šance nemít dudlík mezi plánovanými dětmi}} = \frac{\text{šance být neplánovaný mezi dětmi bez dudlíku}}{\text{šance být neplánovaný mezi dětmi s dudlíkem}}$$

  (záleží jestli se díváme na řádky, nebo na sloupce tabulky, viz Poznámka 7 v textu *AKD*).

- Hypotézu nezávislosti lze v kontextu podílu šancí přeformulovat jako

$$H_0 : \beta = 1$$
$$H_1 : \beta \neq 1.$$

čemuž odpovídá i komentář ve výstupu funkce `fisher.test`: „`alternative hypothesis: true odds ratio is not equal to 1`“.

- Ve výstupu je dále uveden i odhad β, tzv. empirický podíl šancí: $b = 0.885025$. Není to ale přesně ten, na který jsme zvyklí (vzorec (17) v *AKD*), ten by vyšel

$$b = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{9 * 44}{14 * 32} = 0.8839286 \tag{1}$$

R odhad počítá pomocí metody maximální věrohodnosti (populární metoda pro odhad parametrů), a dostává tedy malinko odlišný výsledek.

- R dále uvádí i 95% interval spolehlivosti pro β, který činí $(0.298, 2.518)$. V tomto intervalu tedy s pravděpodobností 0.95 leží skutečný podíl šancí β. Hodnota 1 v tomto intervalu leží, což korespoduje s naším závěrem nezamítnout nezávislost.

3) Třetí způsob jak otestovat nezávislost je pomocí porovnání dvou binomických rozdělení (viz *AKD*, sekce 4.4). Označme si jako π_0 (populační) proporci dětí bez dudlíku mezi neplánovanými a jako π_1 (populační) proporci dětí bez dudlíku mezi plánovanými. Děti tedy rozdělíme do dvou skupin (na plánované a neplánované) a za „úspěch“ považujeme, když neměli dudlík. Jsou-li veličiny Dudlik a Plan nezávislé, měly by být populační proporce (pravděpodobnosti „úspěchu“) stejné. Budeme tedy testovat $\mathrm{H}_0 : \pi_0 = \pi_1$ proti $\mathrm{H}_1 : \pi_0 \neq \pi_1$.

```
(nemaDudlik <- TAB[, "ne"])              # pocty deti bez dudliku
(pocetDeti <- margin.table(TAB, 1))      # pocty deti neplanovanych a planovanych
prop.test(nemaDudlik, pocetDeti, correct=FALSE)     # bez Yatesovy korekce
```

- Příkaz `prop.test` slouží k otestování rovnosti pravděpodobností dvou (i více) binomických rozdělení. Prvním argumentem jsou počty „úspěchů“ (= dětí bez dudlíku), druhým argumentem jsou počty pokusů (= dětí).
- Hodnota testové statistiky (viz *AKD*, vzorec (18)) vychází $\mathrm{X}^2 = 0.0644$.
- P-hodnota je 0.7997, což je víc než zvolená hladina $\alpha = 0.05$. Nezamítáme tedy nulovou hypotézu o rovnosti proporcí dětí bez dudlíku v obou skupinách (dětmi plánovanými a neplánovanými). Nezamítáme tedy nezávislost veličin `Dudlik` a `Plan`.
- Liší se P-hodnota od χ^2-testu nezávislosti (se stejnou nepřítomností korekce na spojitost)?
Ne, p-hodnota je shodná, nebot' oba testy v principu testují totéž.
- Ve výstupu je dále k dispozici 95% interval spolehlivosti pro rozdíl $\pi_0 - \pi_1$.
Skutečný rozdíl populačních proporcí by měl tedy s pravděpodobností 0.95 ležet v intervalu $(-0.1897, 0.1460)$. Jelikož 0 leží v tomto intervalu, má šancí být skutečnou hodnotou rozdílu $\pi_0 - \pi_1$, což je v souladu s naším předchozím závěrem nezamítnout hypotézu o rovnosti proporcí.
- Posledním údajem jsou odhady π_0 a π_1, které činí:

$$\hat{\pi}_0 = \frac{n_{11}}{n_{1.}} = \frac{9}{41} = 0.2195122$$
$$\hat{\pi}_1 = \frac{n_{21}}{n_{2.}} = \frac{14}{58} = 0.2413793.$$

✧ Opět je k dispozici Yatesova korekce na spojitost:

```
prop.test(nemaDudlik, pocetDeti)                # s Yatesovou korekci
```

P-hodnota je opět shodná jako u χ^2-testu (s Yatesovou korekcí).

4) Samozřejmě by šlo si jako η_0 označit (populační) proporci „neplánovaných“ dětí mezi těmi bez dudlíku a jako η_1 (populační) proporci „neplánovaných“ dětí mezi těmi s dudlíkem, tj. děti bychom rozdělili do dvou skupin podle toho, jestli mají dudlík a za „úspěch“ bychom považovali, že byly neplánované. Pak bychom testovali hypotézu $H_0 : \eta_0 = \eta_1$ proti $H_1 : \eta_0 \neq \eta_1$.

```
(nePlan <- TAB["ne",])                   # pocty neplanovanych deti
(pocetDeti <- margin.table(TAB, 2))      # pocty deti bez a s dudlikem
prop.test(nePlan, pocetDeti, correct=FALSE)    # bez Yatesovych korekci
prop.test(nePlan, pocetDeti)                   # s Yatesovymi korekcemi
```

✧ Tento postup je zcela ekvivalentní s přístupem z bodu 3). P-hodnota vychází stejně a náš závěr je také shodný.

4 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:

File ➡ Save

nebo klávesovou skratkou `Ctrl-s`.