# 11 Hodnocení kontingenčních tabulek

**Příklad 11.1. Testování hypotézy o nezávislosti, měření síly závislosti**
V roce 1950 zkoumali Yule a Kendall barvu očí a vlasů u 6800 mužů. Výsledky zkoumání jsou uvedeny v následující tabulce a v souboru vlasy_oci.csv.

| Barva očí | Barva vlasů | | | |
|---|---|---|---|---|
| | světlá | kaštanová | černá | rezavá |
| modrá | 1768 | 807 | 189 | 47 |
| šedá/zelená | 946 | 1387 | 746 | 53 |
| hnědá | 115 | 438 | 288 | 16 |

Na asymptotické hladině významnosti $\alpha = 0.05$ testujte hypotézu o nezávislosti barvy očí a barvy vlasů. Vypočtěte Cramérův koeficient.

### Řešení příkladu 11.1

- $H_0$ : Barva očí a barva vlasů ................................ stochasticky nezávislé.
- $H_1$ : Barva očí a barva vlasů ................................ stochasticky nezávislé.
- Hladina významnosti $\alpha =$ ............... .

### Podmínka dobré aproximace

```
data <- read.delim(..., header = T, row.names = 1) # nacteni datoveho souboru
# funkce chisq.test()
#   1. argument ... tabulka dat
#   correct = F ... zakaze provedeni korekce Pearsonova (viz priklad 11.3)
# vystup funkce:
#   chisq.test() ... kompletni vystupni zprava Pearsonova chi^2 testu
#   chisq.test()$expected ... tabulka ocekavanych cetnosti
round(chisq.test(data, correct = F)$expected, 1) # tabulka ocekavanych cetnosti
                                                  # (zaokrouhlena na 1 des. misto)
```

```
             svetla kastanova cerna rezava
modra        1167.3    1086.0 500.9   47.9
seda/zelena  1304.7    1213.9 559.9   53.5
hneda         357.0     332.1 153.2   14.6
```

Podmínky dobré aproximace ................... splněny. Všechny teoretické četnosti jsou ......................... než 5.

### Pearsonův $\chi^2$ test

```
chisq.test(data, correct = F) # Pearsonuv chi-kvadratovy test
alpha <- ... # hladina vyznamnosti alpha
r <- ... # pocet variant znaku X (barva oci)
s <- ... # pocet variant znaku Y (barva vlasu)
q <- qchisq(1 - alpha, (r - 1) * (s - 1)) # dolni hranice kritickeho oboru
```

```

        Pearson's Chi-squared test

data:  data
X-squared = 1088.1, df = 6, p-value < 2.2e-16
```

```
         q
1 12.59159
```

a) **Test kritickým oborem**
Hodnota testovací statistiky $K$ je ......................................... . Kritický obor má tvar ......................................... .
Protože ............................., $H_0$................................ na hladině významnosti $\alpha =$............................ .

b) **Test $p$-hodnotou**
$P$-hodnota vyšla ........................ Protože $p$-hodnota ...................., $H_0$ ................................ na hladině významnosti $\alpha$ =............................ .

Pro zjištění míry závislosti v kontingenční tabulce použijeme ................................. koeficient.

```
lsr::cramersV(data) # Crameruv koeficient
```

```
  Crameruv.koeficient
1           0.2830494
```

Hodnota Cramérova koeficientu je .......................... .

**Interpretace výsledků:** Znaky barva očí a barva vlasů jsou / nejsou stochasticky nezávislé. Mezi barvou očí a barvou vlasů existuje .............................. stupeň závislosti.

**Příklad 11.2. Fisherův faktoriálový test**
100 náhodně vybraných mužů a žen bylo dotázáno, zda dávají přednost nealkoholickému nápoji A či B. Údaje jsou uvedeny ve čtyřpolní kontingenční tabulce.

| pref. nápoj | pohlaví | |
|---|---|---|
| | muž | žena |
| A | 20 | 30 |
| B | 30 | 20 |

Na hladině významnosti $\alpha = 0.05$ testujte pomocí Fisherova faktoriálového testu hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

**Řešení příkladu 11.2**

- $H_0$ : Znaky pohlaví a preference ................................ stochasticky nezávislé.
- $H_1$ : Znaky pohlaví a preference ................................ stochasticky nezávislé.
- Hladina významnosti $\alpha$ = ............... .

**Fisherův faktoriálový test**

```
data <- data.frame(muz = c(20, 30), zena = c(30, 20),
                   row.names = c('A', 'B')) # vytvoreni datove tabulky

# fisher.test()
#  1. argument ... datova tabulka
#  alternative ... typ alternativni hypotezy
fisher.test(data, alternative = 'two.sided') # Fisheruv faktorialovy test
```

```

        Fisher's Exact Test for Count Data

data:  data
p-value = 0.07134
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1846933 1.0640121
sample estimates:
odds ratio
 0.4481632
```

Protože $p$-hodnota = ...................................... je ........................... než $\alpha = 0.05$, nulovou hypotézu $H_0$ o nezávislosti preferovaného typu nápoje na pohlaví ..................................... na hladině významnosti $\alpha$ =............................ .

**Příklad 11.3. Podíl šancí**
Pro údaje z příkladu 11.2 vypočtěte podíl šancí a sestrojte 95 % asymptotický interval spolehlivosti pro logaritmus podílu šancí. Pomocí tohoto intervalu spolehlivosti testujte na asymptotické hladině významnosti $\alpha = 0.05$ hypotézu, že preferovaný typ nápoje nezáleží na pohlaví respondenta.

### Řešení příkladu 11.3

- $H_0$ : ..................................... → .....................................
- $H_1$ : ..................................... → .....................................
- Hladina významnosti $\alpha =$ ............... .

### Podmínka dobré aproximace

Podmínky dobré aproximace .................... splněny. Všechny teoretické četnosti jsou .......................... než 5.

```
a <- 20; b <- 30; c <- 30; d <- 20
data <- data.frame(muz = c(a, c), zena = c(b, d),
                   row.names = c('A', 'B')) # vytvoreni datove tabulky
chisq.test(..., correct = ...)$expected # tabulka ocekavanych cetnosti
```

```
  muz zena
A  25   25
B  25   25
```

### Výpočet (logaritmu) podílu šancí

```
OR <- (a * d) / (b * c) # podil sanci
lnOR <- log(OR) # logaritmus podilu sanci
```

```
         OR       lnOR
1 0.4444444 -0.8109302
```

Podíl šancí $OR =$.................................. . Logaritmus podílu šancí $\ln(OR) =$.................................. .

a) **Test kritickým oborem**

```
t0 <- log(OR) / sqrt(1 / a + 1 / b + 1 / c + 1 / d) # hodnota testovaci statistiky
q1 <- qnorm(alpha / 2) # horni hranice kritickeho oboru
q2 <- qnorm(1 - alpha / 2) # dolni hranice kritickeho oboru
```

```
         t0        q1       q2
1 -1.986365 -1.959964 1.959964
```

Hodnota testovací statistiky $t_0$ je ......................................... . Kritický obor má tvar ......................................... .
Protože ............................., $H_0$................................. na hladině významnosti $\alpha =$............................. .

b) **Test intervalem spolehlivosti**

```
dh <- log(OR) + sqrt(1 / a + 1 / b + 1 / c + 1 / d) * qnorm(alpha / 2) # dolni
    hranice IS
hh <- log(OR) + sqrt(1 / a + 1 / b + 1 / c + 1 / d) * qnorm(1 - alpha / 2) # horni
    hranice IS
```

```
         dh          hh
1 -1.611082 -0.01077827
```

Interval spolehlivosti má tvar ......................................... . Protože ............................., $H_0$.................................
na hladině významnosti $\alpha =$............................. .

c) **Test $p$-hodnotoou**

```
p.val <- 2 * min(pnorm(t0), 1 - pnorm(t0)) # p-hodnota
```

```
           p
1 0.04699278
```

$P$-hodnota vyšla ........................ Protože $p$-hodnota ...................., $H_0$ ................................ na hladině významnosti $\alpha =$............................ .

*Interpretace výsledků:* Znaky pohlaví a preference jsou / nejsou stochasticky nezávislé.

*Poznámka:* Uvedený výsledek je v rozporu s výsledkem, ke kterému dospěl Fisherův faktoriálový (přesný) test. Je to způsobeno tím, že test pomocí asymptotického intervalu spolehlivosti je pouze přibližný. Ke stejnému závěru, jaký jsme dostali u testování pomocí podílu šancí, dospějeme, pokud použijeme Pearsonův chí-kvadrát test o nezávislosti.

```
chisq.test(data, correct = F)
```

```

        Pearson's Chi-squared test

data:  data
X-squared = 4, df = 1, p-value = 0.0455
```

Ve funkci `chisq.test()` však můžeme zadat parametr `correct` = `T`, který provede korekci Pearsonova testu pro kontingenční tabulky typu $2 \times 2$. Výsledek takto provedeného testu je již v souladu s Fisherovým faktoriálovým testem.

```
chisq.test(data, correct = T)
```

```

        Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 3.24, df = 1, p-value = 0.07186
```