11 Testování nezávislosti v kontingenčních tabulkách

11.1 Obecné kontingenční tabulky

Předpokládáme, že máme n objektů, na nichž zjišťujeme hodnoty $(x_1, y_1)^T, \dots, (x_n, y_n)^T$ dvou nominálních veličin X a Y, veličina X má r variant, veličina Y má s variant. Tyto dvojice považujeme za realizace náhodného výběru rozsahu n z dvourozměrného rozložení, kterým se řídí dvourozměrný diskrétní náhodný vektor $(X, Y)^T$. Zjištěné absolutní simultánní četnosti n_{jk} dvojice variant $(x_{[j]}, y_{[k]})^T$ uspořádáme do kontingenční tabulky (viz tabulka 11.1):

Tabulka 11.1: Kontingenční tabulka absolutních četností

	Y			
X	$y_{[1]}$	$\dots$	$y_{[s]}$	$n_{j.}$
$x_{[1]}$	n_{11}	$\dots$	n_{1s}	$n_{1.}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$x_{[r]}$	n_{r1}	$\dots$	n_{rs}	$n_{r.}$
$n_{.k}$	$n_{.1}$	$\dots$	$n_{.s}$	n

Tabulku doplníme o marginální četnosti: $n_{j.} = n_{j1} + \dots + n_{js}$ je marginální absolutní četnost varianty $x_{[j]}$, $j = 1, \dots, r$, $n_{.k} = n_{1k} + \dots + n_{rk}$ je marginální absolutní četnost varianty $y_{[k]}$, $k = 1, \dots, s$.

Na hladině významnosti α testujeme H_0: X, Y jsou stochasticky nezávislé náhodné veličiny proti H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny.

Testová statistika $K = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n}\right)^2}{\frac{n_{j.}n_{.k}}{n}}$, kde $\frac{n_{j.}n_{.k}}{n}$ je tzv. teoretická četnost dvojice variant $(x_{[j]}, y_{[k]})^T$.

Platí-li H_0, pak K se asymptoticky řídí rozložením $\chi^2((r-1)(s-1))$. Kritický obor: $W = \langle \chi^2_{1-\alpha}((r-1)(s-1)); \infty)$. Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti α, když $K \in W$. Tento test se nazývá Pearsonův χ^2 test nezávislosti.

Podmínka dobré aproximace: Rozložení statistiky K lze aproximovat rozložením $\chi^2((r-1)(s-1))$, pokud teoretické četnosti aspoň v 80 % případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20 % neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant. Sílu závislosti mezi veličinami X, Y měří Cramérův koeficient: $V = \sqrt{\frac{K}{n(m-1)}}$, kde $m = \min\{r, s\}$. Cramérův koeficient V nabývá hodnot mezi 0 a 1. Čím blíže je k 1, tím je závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější. Stupnice míry závislosti je uvedena v tabulce 11.2.

Tabulka 11.2: Stupnice míry závislosti pro Cramérův koeficient

Cramérův koeficient V	**Interpretace**
$\langle 0,0; 0,1)$	zanedbatelný stupeň závislosti
$\langle 0,1; 0,3)$	slabý stupeň závislosti
$\langle 0,3; 0,7)$	střední stupeň závislosti
$\langle 0,7; 1,0\rangle$	silný stupeň závislosti

Příklad 11.1. Řešený příklad
Načtěte datový soubor 22-multinom-palmar-lines.txt obsahující údaje o zakončení tří dlaňových linií (vysoké, střední, nízké) a o barvě vlasů (světlá, střední, tmavá) u 100 mužů a 100 žen. Na asymptotické hladině významnosti $\alpha = 0,01$ testujte hypotézu o nezávislosti typu zakončení tří dlaňových linií a barvě vlasů u mužů. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti kvantifikujte Cramérovým koeficientem.

Řešení příkladu 11.1
Nejprve je třeba ověřit podmínku dobré aproximace. Aby byla podmínka splněna, je třeba, aby v alespoň 80 % případů nabývaly teoretické četnosti $\frac{n_{j.}n_{.k}}{n}$ hodnoty větší nebo rovné 5 a ve zbylých 20 % případů nebyly menší než 2. V softwaru R získáme teoretické četnosti jako výstup expected funkce chisq.test().

```
data <- read.delim('22-multinom-palmar-lines.txt', sep = '\t', dec = '.')
data.M <- data.frame(data[, 2:4], row.names = data[, 1])
chisq.test(data.M)$expected
```

```
       Hi    Mi   Lo
LiH  7,04  5,28 3,68
MH  18,48 13,86 9,66
DaH 18,48 13,86 9,66
```

Z výstupu vidíme, že pouze jedna teoretická četnost z devíti (tj. 11.11 %) je menší než 5, přičemž tato četnost nabývá hodnoty $3,68$, což ve více než 2. Podmínka dobré aproximace je tedy splněna.

Na asymptotické hladině významnosti $\alpha = 0,01$ testujeme H_0: *X a Y jsou stochasticky nezávislé náhodné veličiny.* proti H_1: *X a Y nejsou stochasticky nezávislé náhodné veličiny.* K testování použijeme Pearsonův χ^2 test nezávislosti implementovaný ve funkci chisq.test(). Výstupem funkce je realizace testové statistiky K a p-hodnota. Dolní hranici kritického oboru $\chi^2_{1-\alpha}((r-1)(s-1))$, kde $r = 3$ a $s = 3$, dopočítáme příkazem qchisq(). Nakonec kvantifikujeme míru závislosti pomocí Cramérova koeficientu. Jeho hodnotu vypočítáme pomocí funkce cramersV() z knihovny lsr.

```
alpha <- 0.01
r <- 3; s <- 3
qchisq(1 - alpha, (r - 1) * (s - 1)) # 13,2767
chisq.test(data.M)
lsr::cramersV(data.M) # 0,1014841
```

```

        Pearson's␣Chi-squared␣test

data:␣␣data.M
X-squared␣=␣2,0598,␣df␣=␣4,␣p-value␣=␣0,7248
```

Realizace testové statistiky $k = 2,0598$, kritický obor $W = \langle 13,2767; \infty)$. Protože $k \notin W$, H_0 nezamítáme na asymptotické hladině významnosti $\alpha = 0,01$. Protože p-hodnota $= 0,7248$ je větší než $\alpha = 0,01$, H_0 nezamítáme na asymptotické hladině významnosti $\alpha = 0,01$. Mezi typem zakončení tří dlaňových linií a barvou vlasů u mužů neexistuje statisticky významná stochastická závislost. Mezi typem zakončení tří dlaňových linií a barvou vlasů u mužů existuje slabý stupeň závislosti ($V = 0,1015$). ★

Příklad 11.2. Neřešený příklad
Načtěte datový soubor 20-more-samples-probabilities-pubis.txt obsahující údaje o frekvenci výskytu tří stupňů změn ((i) žádné změny, (ii) stopové až malé změny, (iii) střední až výrazné změny) kostního reliéfu na vnitřní straně stydké kosti (*os pubis*) v blízkosti spony stydké (*symphysis pubica*) u žen z kosterních souborů tří populací: evropského původu, afrického původu a Inuitů. Na asymptotické hladině významnosti $\alpha = 0,10$ testujte hypotézu o nezávislosti míry změn kostního reliéfu na vnitřní straně stydké kosti a populace. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti kvantifikujte Cramérovým koeficientem.
Výsledky: podmínka dobré aproximace je splněna (devět z devíti (100 %) teoretických četností je větších než 5); $k = 9,4351$, $r = 3$, $s = 3$, $W = \langle 7,7794; \infty)$; p-hodnota $= 0,0511$, $\alpha = 0,10$; H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,10$; Cramérův koeficient: $V = 0,1517$. ★

11.2 Čtyřpolní kontingenční tabulky

Je-li $r = s = 2$, jedná se o čtyřpolní kontingenční tabulku, v níž používáme označení: $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$.

Hypotézu o nezávislosti veličin X, Y můžeme ve čtyřpolní kontingenční tabulce testovat několika způsoby.

a) Pomocí Pearsonova χ^2 testu nezávislosti (v tomto případě má kritický obor tvar $W = \langle\chi^2_{1-\alpha}(1); \infty))$.

b) Pomocí Fisherova přesného testu (p-hodnotu tohoto testu, kterou nám poskytne statistický software, porovnáme s hladinou významnosti α. Je-li p-hodnota $\leq \alpha$, pak hypotézu o nezávislosti zamítáme na hladině významnosti α).

c) Pomocí výběrového podílu šancí $OR = \frac{ad}{bc}$. Za platnosti H_0 je OR blízký 1 testová statistika $T_0 = \dfrac{\ln OR}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$ se asymptoticky řídí rozložením $N(0, 1)$. Kritický obor: $W = (-\infty; -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}; \infty)$.
H_0 zamítáme na asymptotické hladině významnosti α, když $T_0 \in W$. Test o nezávislosti lze provést i pomocí $100(1-\alpha)\%$ asymptotického intervalu spolehlivosti pro logaritmus teoretického podílu šancí $o\rho$, který je dán vzorcem: $(d; h) = \left(\ln OR - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}u_{1-\alpha/2}; \ln OR + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}u_{1-\alpha/2}\right)$.

   Jestliže interval spolehlivosti neobsahuje 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

   Upozornění: Pokud bychom chtěli získat interval spolehlivosti nikoliv pro $\ln o\rho$, ale pro $o\rho$, stačí uvedené meze odlogaritmovat. Při testování hypotézy o nezávislosti veličin X, Y pomocí tohoto intervalu spolehlivosti pak zjišťujeme, zda interval pokrývá číslo 1. Pokud ne, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti α.

Příklad 11.3. Řešený příklad
Načtěte datový soubor 25-one-sample-probability-dermatoglyphs.txt obsahující údaje o frekvenci výskytu dermatoglyfických vzorů *vír*, *smyčka* a *oblouček* na deseti prstech 470 jedinců (235 mužů a 235 žen) bagathské populace z Araku Valley. Na hladině významnosti $\alpha = 0,05$ testujte hypotézu o nezávislosti mezi výskytem dermatoglyfického vzoru *smyčka* a pohlavím u bagathské populace z Araku Valley. Testování proveďte (a) pomocí Pearsonova χ^2 testu nezávislosti ((1) kritickým oborem, (2) p-hodnotou); (b) pomocí Fisherova přesného testu ((1) p-hodnotou); (c) pomocí podílu šancí ((1) kritickým oborem, (2) intervalem spolehlivosti, (3) p-hodnotou). Pro situaci (a) vypočítejte a interpretujte Cramérův koeficient; pro situaci (c) vypočítejte a interpretujte hodnotu výběrového podílu šancí.

Řešení příkladu 11.3
Na hladině významnosti $\alpha = 0,05$ testujeme H_0: *X a Y jsou stochasticky nezávislé náhodné veličiny.* proti H_1: *X a Y nejsou stochasticky nezávislé náhodné veličiny.*

(a) Nejprve je třeba ověřit podmínku dobré aproximace.

```
data <- read.delim('25-one-sample-probability-dermatoglyphs.txt', sep = '\t', dec = '.')
data.L <- data.frame(muzi = c(1246, 235 * 10 - 1246),
                     zeny = c(1349, 235 * 10 - 1349),
                     row.names = c('smycka', 'jine'))
chisq.test(data.L)$expected
```

```
         muzi   zeny
smycka 1297,5 1297,5
jine   1052,5 1052,5
```

Z výstupu vidíme, že všechny teoretické četnosti (tj. 100 %) jsou větší než 5, podmínka dobré aproximace je tedy splněna.

K testování použijeme Pearsonův χ^2 test nezávislosti implementovaný ve funkci chisq.test(). Výstupem funkce je realizace testové statistiky K a p-hodnota. Dolní hranici kritického oboru $\chi^2_{1-\alpha}(1)$ dopočítáme příkazem qchisq(). Nakonec kvantifikujeme míru závislosti pomocí Cramérova koeficientu (funkce cramersV() z knihovny lsr).

```
alpha <- 0.05
qchisq(1 - alpha, 1) # 3,841459
chisq.test(data.L)
lsr::cramersV(data.L) # 0,04364208
```

```

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

data:  data.L
X-squared = 8,9518, df = 1, p-value = 0,002772
```

Realizace testové statistiky $k = 8,9518$, kritický obor $W = \langle 3,8415; \infty)$. Protože $K \in W$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,05$. Protože p-hodnota $= 0,002772$ je menší než $\alpha = 0,05$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,05$. Mezi výskytem dermatoglyfického vzoru *smyčka* a pohlavím u bagathské populace z Araku Valley existuje statisticky významná stochastická závislost. Mezi výskytem dermatoglyfického vzoru *smyčka* a pohlavím u bagathské populace z Araku Valley existuje zanedbatelný stupeň závislosti ($V = 0,04364$).

(b) Fisherův přesný test provedeme pomocí funkce fisher.test() implementované v softwaru R. Výstupem funkce je interval spolehlivosti a p-hodnota.

```
fisher.test(data.L, alternative = 'two.sided', conf.level = 0.95)
```

```

        Fisher's␣Exact␣Test␣for␣Count␣Data

data:␣␣data.L
p-value␣=␣0,002768
alternative␣hypothesis:␣true␣odds␣ratio␣is␣not␣equal␣to␣1
95␣percent␣confidence␣interval:
␣0,7451425␣0,9412385
sample␣estimates:
odds␣ratio
␣0,8375093
```

Protože p-hodnota $= 0,002768$ je menší než $\alpha = 0,05$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Mezi výskytem dermatoglyfického vzoru *smyčka* a pohlavím u bagathské populace z Araku Valley existuje statisticky významná stochastická závislost.

(c) Test podílem šancí provedeme pomocí funkce odds.ratio.test() implementované v R-skriptu AS-sbirka-funkce.R, který je součástí této publikace. R-skript načteme příkazem source(). Výstupem funkce odds.ratio.test() je hodnota podílu šancí OR, logaritmus podílu šancí lnOR, realizace testové statistiky t_0, interval spolehlivosti pro logaritmus podílu šancí a p-hodnota. Hranice kritického oboru dopočítáme příkazem qnorm().

```
source('AS-sbirka-funkce.R')
alpha <- 0.05
qnorm(alpha / 2) # -1,959964
qnorm(1 - alpha / 2) # 1,959964
odds.ratio.test(data.L, conf.level = 0.95)
```

```
      OR    lnOR      t0      dh      hh      p
1 0,8375 -0,1774 -3,0203 -0,2925 -0,0623 0,0025
```

Realizace testové statistiky $t_0 = -3,0203$, kritický obor $W = (-\infty; -1,96\rangle \cup \langle 1,96; \infty)$. Protože $t_0 \in W$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,05$. Interval spolehlivosti $IS = (-0,2925; -0,0623)$. Protože

$c = 0 \notin IS$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,05$. Protože p-hodnota $= 0,0025$ je menší než $\alpha = 0,05$, H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,05$. Mezi výskytem dermatoglyfického vzoru *smyčka* a pohlavím u bagathské populace z Araku Valley existuje statisticky významná stochastická závislost. Podíl šancí výskytu dermatoglyfického vzoru *smyčka* mužů ku ženám vyšel 0,8375. V takovém případě bývá nicméně lepší interpretovat převrácenou hodnotu podílu šancí, čili $1/0,8375 = 1,1940$. Šance na výskyt dermatoglyfického vzoru *smyčka* u žen bagathské populace z Araku Valley je 1,1940-krát větší než u mužů bagathské populace z Araku Valley. ★

Příklad 11.4. Neřešený příklad
Načtěte datový soubor 26-two-samples-probabilities-palmar.txt obsahující údaje o frekvenci výskytu vysokého (11.9.7), středního (9.7.5), nízkého (7.5.5) a jiného zakončení dlaňových linií na pravé a levé straně 50 mužů a 50 žen z populace Mech a 105 mužů a 87 žen z populace Rajbanshi. Na hladině významnosti $\alpha = 0,01$ testujte hypotézu o nezávislosti výskytu nízkého zakončení dlaňových linií na pravé straně u mužů z populace Mech a u mužů z populace Rajbanshi. Testování proveďte (a) pomocí Pearsonova χ^2 testu nezávislosti ((1) kritickým oborem, (2) p-hodnotou); (b) pomocí Fisherova přesného testu ((1) p-hodnotou); (c) pomocí podílu šancí ((1) kritickým oborem, (2) intervalem spolehlivosti, (3) p-hodnotou). Pro situaci (a) vypočítejte a interpretujte Cramérův koeficient; pro situaci (c) vypočítejte a interpretujte hodnotu výběrového podílu šancí.
Výsledky: (a) Pearsonův χ^2 test nezávislosti: podmínka dobré aproximace je splněna (čtyři ze čtyř (100 %) teoretických četností jsou větší než 5); $k = 9,4673$, $W = \langle 6,6349; \infty)$; p-hodnota $= 0,002092$, $\alpha = 0,01$; H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,01$; Cramérův koeficient $V = 0,2471$; (b) Fisherův přesný test: p-hodnota $= 0,001676$, $\alpha = 0,01$; H_0 zamítáme na hladině významnosti $\alpha = 0,01$; (c) test podílem šancí: $t_0 = 3,1954$, $W = (-\infty; -2,5758\rangle \cup \langle 2,5758; \infty)$; $IS = (0,2255; 2,1008)$, $c = 0$; p-hodnota $= 0,0014$, $\alpha = 0,01$; H_0 zamítáme na asymptotické hladině významnosti $\alpha = 0,01$; výběrový podíl šancí $OR = 3,2000$. ★