Kontingenční tabulky a testy shody
4.1.2018

Kontingenční tabulky

1. Tabulka 1 shrnuje osudy pasažérů lodě Titanic, která tragicky ztroskotala v roce 1912. Zajímá nás, zda existuje nějaká souvislost mezi třídou, ve které cestující cestoval, a přežitím, nebo zda jsou tyto dva faktory nezávislé.

	Přežil	
Třída	Ne	Ano
1	122	203
2	167	118
3	528	178
Posádka	673	212

Tabulka 1: Data o Titanicu.

   Jaký model teď budeme uvažovat a pro co? Co všechno je v modelu náhodné a co naopak není?

2. Nejprve si do R zadáme tabulku 1.

```
titanic=matrix(c(122, 167, 528, 673, 203, 118, 178, 212),ncol=2)
dimnames(titanic)=list(Trida=c("1","2","3","Posadka"),Prezil=c("Ne","Ano"))
titanic
```

   Zkontrolujte, že máme tabulku správně zadanou. Kdybychom si chtěli dopočítat marginální četnosti, provedeme to následovně:

```
apply(titanic,1,sum)
apply(titanic,2,sum)
```

3. Opakování: Vhodným obrázkem graficky ilustrujte marginální rozdělení zkoumaných dvou veličin.

4. Vrátíme se zpět ke kontingenční tabulce. Podíváme se na tabulky relativních četností

```
prop.table(titanic)
prop.table(titanic,marg=1)
prop.table(titanic,marg=2)
```

   Co nám jednotlivé relativní četnosti odhadují? Jak by měly tabulky přibližně vypadat v případě nezávislosti?
   Podíváme se na tutéž věc i graficky:

```
barplot(titanic,beside=T,legend=T)
barplot(prop.table(titanic,mar=2),beside=T,legend=T)
barplot(t(titanic),beside=T,legend=T)
barplot(prop.table(t(titanic),mar=2),beside=T,legend=T)
```

Prozkoumejte jednotlivé obrázky a jak se mezi sebou liší. Co si na základě čísel a grafů myslíte o vztahu zkoumaných dvou veličin? Jsou nezávislé?

5. Provedeme χ^2 test nezávislosti.

```
chisq.test(titanic,correct=FALSE)
```

Jaký je náš závěr?

(a) Připomeneme, jak se spočítá testová statistika χ^2 testu:

```
a1=apply(titanic,1,sum)
a2=apply(titanic,2,sum)
n=sum(titanic)

E=a1%o%a2/n
sum((titanic-E)^2/E )
```

Kolik má stupňů volnosti příslušné asymptotické χ^2 rozdělení? Jak se toto číslo spočítá?

(b) Ještě si prohlédneme jednotlivé položky, které máme k dispozici po použití funkce chisq.test:

```
 CH=chisq.test(titanic,correct=FALSE)
names(CH)

CH$residuals
```

Co přesně jsou tato „rezidua“? Které kategorie tabulky nejvíce přispívají k výsledné hodnotě χ^2 statistiky a tím „porušují“ nezávislost?

Jak bychom shrnuli naše poznatky týkající se přežití pasažérů z jednotlivých tříd?

6. A není to s tím Titanicem celé trochu jinak? Podíváme se na úplně kompletní data, která jsou k dispozici v R :

```
data(Titanic)
Titanic

#nase tabulka 1
apply(Titanic,c(1,4),sum)

#dalsi tabulky:
(t1=apply(Titanic,c(2,4),sum))
prop.table(t1,mar=1)

(t2=apply(Titanic,c(1,2),sum))
prop.table(t2,mar=1)
```

7. Uvažujme 2×2 tabulku uloženou v `t1`, která shrnuje vztah pohlaví a přežití pasažérů.

(a) Otestujte nezávislost těchto dvou veličin pomocí χ^2 testu.
(b) Podívejte se na problém jinak a otestujte shodu pravděpodobností přežití pro muže a pro ženy, pomocí funkce prop.test.
(c) Jak se liší uvažované dva modely v (a) a (b)? V jakém vztahu jsou testové statistiky v (a) a (b)?
(d) Uvažujme model jako v (a). Jaké je rozdělení marginálních řádkových četností n_{+1} a n_{+2}? Jaké je rozdělení četností v tabulce, podmíníme-li marginálními četnostmi n_{+1} a n_{+2}?

8. Odhadněte poměr šancí na přežití žen a mužů z tabulky t1.

```
# pomoci odhadnutych pravdepodobnosti:
(phat=prop.table(t1,mar=1)[,2])
(odds = phat/(1-phat))
(odds.ratio=odds[2]/odds[1])

# nebo rychleji primo z tabulky:
t1[1,1]*t1[2,2]/(t1[1,2]*t1[2,1])
```

Jak budeme interpretovat toto číslo? Jaká hodnota by odpovídala nezávislosti?
Pro poměr šancí můžeme zkonstruovat i interval spolehlivosti. Vychází z asymptotického normálního rozdělení pro logaritmus poměru šancí, viz skripta str. 109, věta 7.4.

```
sd=sum(1/t1)
log(odds.ratio)+c(-1,1)*qnorm(0.975)*sqrt(sd)
exp(log(odds.ratio)+c(-1,1)*qnorm(0.975)*sqrt(sd))
```

9. Rozhodněte, zda měly ženy více než pětkrát větší šanci na přežití než muži.

TEST SHODY S ROZDĚLENÍM

10. V roce 2008 se v České republice narodilo 119 570 dětí, z toho 58 244 dívek a 61 326 chlapců.
(a) Pomocí χ^2 testu otestujte, zda se dívky a chlapci rodí v poměru 1:1.
(b) Stejnou otázku řešte pomocí jednovýběrového testu o proporci.
(c) Porovnejte testové statistiky z (a) a (b) a jejich rozdělení. Jak se liší uvažované modely?

11. Tabulka 2 udává statistiku počtu vstřelených gólů v německé Bundeslize v roce 2000 (kompletní data jsou obsažena v knihovně vcd, lze je zavolat pomocí příkazu data("Bundesliga").). Zajímá nás, zda je možné modelovat počet vstřelených gólů v jednom zápase pomocí Poissonova rozdělení s parametrem 4.

Počet gólů	0	1	2	3	4	5	6	7 a více
Počet zápasů	25	44	62	65	55	30	14	11

Tabulka 2: Počet vstřelených gólů v Bundeslize v roce 2000.

(a) Jak vypadají původní data, která mají za nulové hypotézy Poissonovo rozdělení? Kolik jich celkem máme? A čemu odpovídají hodnoty v tabulce 2 a jaké mají rozdělení?

(b) Zadáme si hodnoty do R a dopočítáme si teoretické pravděpodobnosti za nulové hypotézy.

```
ngoals=c(25, 44, 62, 65, 55, 30, 14, 11)

(tf=dpois(0:7,lambda=4))
tf[8]=1-sum(tf[1:7])
```

(c) Pozorované a očekávané četnosti si můžeme porovnat graficky:

```
n=sum(ngoals)
barplot(rbind(ngoals,n*tf),beside=T,legend.text=c("pozorovane","ocekavane"))
```

(d) Nakonec provedeme test dobré shody:

```
chisq.test(ngoals,p=tf,correct=F)
```

Jaký je náš závěr?

12. Nyní se podíváme na reálnější situaci: Chtěli bychom zjistit, zda je možné počet vstřelených gólů během zápasu modelovat Poissonovým rozdělením (bez specifikace parametru).

(a) Nejprve tedy musíme neznámý parametr λ odhadnout z rovnice

$$\sum_{k=0}^{7} \frac{X_k}{p_k(\lambda)} \frac{\partial p_k(\lambda)}{\partial \lambda} = 0,$$

kde X_k jsou četnosti v tabulce 2 a $p_k(\lambda)$ jsou příslušné teoretické pravděpodobnosti jednotlivých kategorií. Získaný odhad $\widehat{\lambda}$ pak dosadíme do $p_k(\lambda)$ a následně do χ^2 statistiky.

(b) Po rozepsání výše uvedené rovnice dostaneme:

```
rovnice=function(x){
  kk=0:6
  y=x-(sum(kk*ngoals[1:7])/n +
       ngoals[8]/n*x*(1-ppois(6,lambda=x))/(1-ppois(7,lambda=x)))
  return(y)
  }

# odhad lambda:
(lam.hat=uniroot(rovnice,c(1,10))$root)

#odhad pravdepodobnosti
(tf2=dpois(0:7,lambda=lam.hat))
tf2[8]=1-sum(tf2[1:7])
tf2
```

(c) Vše tedy dosadíme do χ^2 testu:

```
chisq.test(ngoals,p=tf2,correct=F)
```

Co je tady špatně? Jak to „opravíme“, aby byl výsledek správný?

Jaký je nyní náš závěr ohledně rozdělení počtu vstřelených gólů? Jaká je hodnota odhadnutého parametru?

Samostatná práce

1. Proveďte test nezávislosti na data z tabulky 1, kde vyřadíme poslední řádek (posádku). Změní se nějak závěr ohledně nezávislosti cestovní třídy a přežití?

2. Uvažujte dobře známá data Hosi.txt. Bude nás zajímat, zda jsou věk matky a věk otce nezávislé. Nebudeme ale zkoumat naměřená data, ale pouze jejich kategorizované verze, kde budeme brát kategorie do 25 let (včetně), 26–30, 31-35, 36 a výše. Tyto kategoriální veličiny získáme následovně:

   ```
   Fvek.matky=cut(Hosi$vek.matky,breaks=c(0,25,30,35,100))
   Fvek.otce=cut(Hosi$vek.otce,breaks=c(0,25,30,35,100))
   ```

   Proveďte test nezávislosti těchto dvou veličin. V případě, že hypotézu nezávislosti zamítnete, zjistěte, jakým způsobem jsou veličiny asociované.

3. Na webu lidovky.cz byla v listopadu zveřejněná anketa s názvem „Koho byste rádi za nového prezidenta“ (lze ji nalézt pod článkem ze dne 7.11.2017, který představuje nové prezidentské kandidáty). Její výsledky (ke dni 2.1.2018) jsou uvedeny v Tabulce 3.

	MZ	JD	MT	MH	ostatní kandidáti
Počet hlasů	2194	5308	1341	626	1171

   Tabulka 3: Výsledky průzkumu na webu lidovky.cz.

   Naproti tomu agentura Phoenix Research v prosinci uvedla (viz blesk.cz, článek ze dne 30.12.2017), že MZ získá 34 % hlasů, JD 17% hlasů, MT 14 % a MH 10 % hlasů, ostatní kandidáti 12 % hlasů a 13 % voličů ještě není rozhodnuto. Zjistěte na základě výše uvedených dat, zda mají čtenáři webu lidovky.cz stejné volební preference jako celková populace ČR (bereme-li v úvahu pouze rozhodnuté voliče).

   Pro zajímavost, voliče z webu lidovky.cz můžete také graficky porovnat s voliči z webu blesk.cz, příslušnou anketu lze nalézt na stránce http://www.blesk.cz/volba-prezidenta-2018-preference. Které odlišnosti Vám přijdou nejzajímavější?

4. Počet německých bomb, které zasáhly jižní Londýn během druhé světové války, uvádí tabulka 4. Data byla získána následujícím způsobem: Celá sledovaná oblast byla rozdělena na 576 regionů stejné velikosti a byly sledovány počty zásahů jednotlivých oblastí.

Počet zásahů	0	1	2	3	4	5+
Počet regionů	229	211	93	35	7	1

   Tabulka 4: Počet německých bomb v Londýně během druhé světové války.

   Zjistěte, zda se počty zásahů řídí Poissonovým rozdělením (případně s jakým parametrem).