Testy v multinomickém rozdělení
19.12.2019

V rámci přednášky pro studenty chemie PřF UK v letech 2006-2013 bylo zjišťováno mimo jiné, v jakém měsíci slaví studenti narozeniny. Naměřena byla data uvedená v tabulce 1.

Měsíc	1	2	3	4	5	6	7	8	9	10	11	12
Počet studentů	29	20	23	28	35	25	31	33	31	26	23	24

Tabulka 1: Počty narozených studentů v jednotlivých měsících.

Data zapsaná v R :

```
x=c(29, 20, 23, 28, 35, 25, 31, 33, 31, 26, 23, 24 )
```

Testy pravděpodobností v multinomickém rozdělení

1. Zajímá nás, zda je pravděpodobnost narození v lednu stejná jako pravděpodobnost narození v prosinci.

   Budeme tedy předpokládat, že $\mathbf{X}$ je náhodný vektor s multinomickým rozdělením $\mathsf{Mult}_{12}(n, \mathbf{p})$, kde $n = 328$ a $\mathbf{p} = (p_1, \ldots, p_{12})^\top$.

   (a) Formulujte nulovou a alternativní hypotézu.
   (b) Odhadněte parametry uvažovaného multinomického rozdělení. Vhodně graficky znázorněte pomocí funkce `barplot`.
   (c) Navrhněte vhodnou testovou statistiku. Využijte při tom, že z přednášky víte, že pro vektor $\mathbf{c}$ platí
   $$\sqrt{n}(\mathbf{c}^\top \widehat{\mathbf{p}} - \mathbf{c}^\top \mathbf{p}) \xrightarrow{D} \mathsf{N}(0, V_c), \quad V_c = \mathbf{c}^\top \mathbf{V} \mathbf{c},$$
   kde $\mathbf{V} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$.
   (d) Pomocí R test ručně proveďte. Pro výpočet $\widehat{V}_c$ si buď příslušný výraz zjednodušte a vyjádřete pomocí $\widehat{p}_1$ a $\widehat{p}_{12}$ nebo můžeme použít násobení matic pomocí `%*%`. Např. pro známé `p` matici $\mathbf{V}$ vytvoříme následovně
   `V=diag(p)-p%*%t(p)`

2. Otestujte, zda je pravděpodobnosti narození dítěte v I. čtvrtletí větší než pravděpodobnost narození ve III. čtvrtletí. Formulujte opět nulovou a alternativní hypotézu a proveďte ručně vhodný test.

Testy dobré shody se známými parametry

3. Zajímá nás, zda se děti rodí rovnoměrně během roku.

   (a) Formulujte nulovou hypotézu a alternativu, které nás zajímají.
   (b) Uložte si hodnotu $\mathbf{p}_0$ z H_0 do vektoru `p0`. Dále provedeme test

```
barplot(cbind(p.hat,p0),beside=TRUE)

chisq.test(x,p=p0,correct=FALSE)
```

Jaký je náš závěr?

(c) Připomeňte si,
- zda se jedná o přesný nebo asymptotický test a z jakého rozdělení je spočtena p-hodnota,
- co by mělo být splněno, aby bylo použití asymptotického testu rozumné,
- jak byste spočítali hodnotu testové statistiky ručně.

(d) Kdybychom chtěli vědet, v kterých kategoriích se porovnost od testované nulové nejvíce liší, můžeme se podívat na tzv. rezidua

```
chisq.test(x,p=p0,correct=FALSE)$residuals
```

Testy dobré shody s neznámými parametry

4. Daný gen se vyskytuje ve dvou možných alelách (a, A). Na základě genetického testu 100 náhodně vybraných osob bylo zjištěno, že 12 je nositelem kombinace aa, 46 osob má genotyp aA a 42 osob má kombinaci AA. Zjistěte, zda platí Hardyho-Weinbergovo ekvilibrium. Označme jako q_A pravděpodobnost výskytu alely A.

(a) Zapište tvar pravděpodobností $\mathbf{p}$ za nulové hypotézy.

(b) Parametr q_A musíme odhadnout z rovnice

$$\sum_{k=0}^{2} \frac{X_k}{p_k(q_A)} \frac{\partial p_k(q_A)}{\partial q_A} = 0,$$

kde X_k jsou četnosti jednotlivých kategorií. Následně spočteme $\mathbf{p}(\widehat{q_A})$ a příslušnou χ^2 statistiku.

```
data=c(42,46,12)

# po vyreseni analyticky:
(q.A=(2*data[1]+data[2])/(2*data[1]+2*data[2]+2*data[3]))


# zadani p0 a test:
(p0.HW=c(q.A^2,2*q.A*(1-q.A),(1-q.A)^2))

chisq.test(data,p=p0.HW,correct=FALSE)
```

Co je v posledním výstupu „špatně“? Jak to „opravíme“, aby byl výsledek správný?

Ve zdrojovém kódu můžete najít, jak lze q_A spočítat i numericky jako řešení výše uvedené rovnice nebo přímo jako argument maxima věrohodnostní funkce.

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.

Testy shody s rozdělením

5. 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.

   (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) Porovnejte pozorované a očekávané četnosti graficky a následně proveďte test dobré shody. Jaký je náš závěr?

6. 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
  n=sum(ngoals)
  y=sum(kk*ngoals[1:7])-n*x+ngoals[8]/(1-ppois(6,lambda=x))*x*(1- ppois(5,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)`

Opět je ale potřeba ještě udělat „úpravu“, abychom dostali správný výsledek.

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

Samostatná práce Krevní skupiny jsou kódovány na základě jednoho genu, který se vyskytuje ve třech alelách (A, B, 0). Alely A a B jsou vůči 0 zcela dominantní. Krevní skupině A tedy odpovídají kombinace AA a A0, krevní skupině B kombinace BB a B0, skupině AB genotyp AB a skupině 0 genotyp 00. U několika náhodně vybraných jedinců byl proveden krevní test a byly zjištěny četnosti uvedené v tabulce (3).

1. Uvádí se, že v Evropě se krevní skupiny vyskytují následovně 0 (33%), A (45%), B (16%), AB (6%). Zjistěte, zda jsou výše uvedená data v souladu s tímto tvrzením.

2. Rozhodněte, zda lze tvrdit, že se v ČR krevní skupina A vyskytuje častěji než krevní skupina 0.

3. ⋆ Zjistěte, zda zde platí Hardyho-Weinbergovo ekvilibrium.
   (a) Zapište pravděpodobnosti jednotlivých kategorií za nulové hypotézy.
   (b) Nalezněte odhad neznámých parametrů. (Hledejte je jako argumenty maxima maximálně věrohodné funkce pomocí funkce `optim`.)
   (c) Proveďte test.

Krevní skupina	0	A	B	AB
Počet osob	378	415	141	66

Tabulka 3: Četnost krevních skupin.