Testy o proporci a testy v multinomickém rozdělení

12. cvičení

I Jednovýběrový problém pro binární data.

V roce 2015 se v České republice živě narodilo 110 764 dětí, z toho 53 947 dívek a 56 817 chlapců (zdroj ČSÚ, Tabulka 1, odhalíte zjevnou chybu v posledním řádku tabulky?). Zajímá nás, zda je pravděpodobnost narození chlapce 1/2. Označme ji dále jako p.

1. Jaký předpokládáme model a jaké budeme testovat hypotézy? Jak vychází bodový odhad pravděpodobnosti p narození chlapce?

2. Nejprve budeme uvažovat Wilsonův test založený na testové statistice

$$W_n = \frac{\sqrt{n}\,(\widehat{p}_n - p_0)}{\sqrt{p_0(1-p_0)}},$$

   která má asymptoticky normální rozdělení $\mathsf{N}(0,1)$.

   Test můžeme provést buď „ručně“ nebo za pomocí funkce `prop.test`.

```
alpha=0.05
n=110764
divky=53947
chlapci=56817

#rucne:
p0=1/2
W=sqrt(n)*(chlapci/n - p0)/sqrt(p0*(1-p0))
(q=qnorm(1-alpha/2))
2*(1-pnorm(abs(W)))

# pomoci funkce v R
 prop.test(chlapci,n,p=1/2,correct=FALSE)
```

   Jaký je náš závěr?

3. Připomeňte si vztah mezi W_n a testovou statistikou uváděnou funkcí prop.test. Vypočtěte p-hodnotu pomocí rozdělení této statistiky.

4. Mohli bychom uvažovat i jinou testovou statistiku Z_n, která by měla také asymptoticky $\mathsf{N}(0,1)$ rozdělení? Proveďte ručně test pomocí této statistiky.

5. Vzpomeňte si, jak byste zkonstruovali (klasický) asymptotický interval spolehlivosti pro p? Jedná se o interval, který vrací funkce prop.test?

6. Nyní provedeme test téže hypotézy pomocí přesného testu. Připomeňte si princip tohoto testu.

```
binom.test(chlapci,n,p=1/2)
```

Na jakém rozdělení je založený tento test? Jaký je nyní náš závěr?
Pozor, reportovaná p-hodnota je počítaná jinak, než je uvedeno ve skriptech. Viz bod (i) *v Doplňujících informacích.*
Jak bychom spočetli uváděný interval spolehlivosti? Je přesný nebo asymptotický?

7. Jelikož alternativní rozdělení splňuje předpoklady centrální limitní věty, mohli bychom použít i asymptotický t-test.
   - Odvoďte, jak vypadá v tomto případě testová statistika T_n. Jak se liší od testové statistiky Z_n?
   - Nyní t-test provedeme (potřebujeme ale naše data ve formě vektoru 0 a 1). Ten vyrobíme následovně:
     ```
     data=c(rep(1,chlapci),rep(0,divky))
     t.test(data,mu=0.5)
     ```

8. V rámci přednášky jste probírali přesný interval spolehlivosti, klasický asymptotický (dále Waldův interval) a Wilsonův interval. Ve skriptech lze nalézt ještě i interval spolehlivosti založený na logitu. Stáhněte si z Moodlu a načtěte si soubor pokryti.R, který obsahuje předem připravenou funkci, která počítá pro všechny čtyři metody skutečné pokrytí pro různé hodnoty parametru p a pro zadaný rozsah výběru `n`. Výsledkem je tedy graf skutečného pokrytí v závislosti na p a tabulka délky jednotlivých intervalů pro několik různých p. Vyzkoušejte tuto funkci pro několik různých voleb n:

   ```
   source("pokryti.R")

   pokryti(n=20)
   pokryti(n=50)
   pokryti(n=200)
   ```

   Jak je to se skutečným pokrytím přesného intervalu spolehlivosti? Který z intervalů spolehlivosti Vám připadá nejlepší?

II Dvouvýběrový problém pro binární data

Zajímá nás, zda je pravděpodobnost narození chlapce stejná v ČR a na Slovensku. Na Slovensku se v roce 2015 narodilo živě 55 615 dětí, z nichž bylo 28 703 chlapců a 26 912 děvčat (zdroj Národné centrum zdravotníckych informácií, Tabulka 33 na str. 42).

```
nSR= 55615
chlapciSR=28703
divkySR=26912
```

9. Porovnejte procentuální zastoupení chlapců mezi narozenými dětmi pro ČR a SR. Vykreslete i vhodné obrázky.

10. Provedeme dvouvýběrový test o proporci založený na rozdílu pravděpodobností.

– Jaký předpokládáme model? Jak zní testované hypotézy?
– Testová statistika, kterou počítá R ve funkci prop.test je založená na statistice

$$T_d = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widetilde{p}(1-\widetilde{p})(\frac{1}{n}+\frac{1}{m})}},$$

kde $\widetilde{p}$ je odhad společné pravděpodobnosti úspěchu za nulové hypotézy. Funkce nám ve výstupu dává $\widetilde{T}_d^2$.
– Provedeme test:
`prop.test(c(chlapci,chlapciSR),c(n,nSR),correct=FALSE)`

Jaký učiníme závěr na základě tohoto testu?

11. Ještě spočítáme testovou statistiku ručně

```
n=110764
xCR=chlapci/n
xSR=chlapciSR/nSR
xall=(chlapci+chlapciSR)/(n+nSR)

(Td=(xCR-xSR)/sqrt(xall*(1-xall)*(1/n+1/nSR)))
2*(1-pnorm(abs(Td)))
```

12. Alternativně bychom mohli odhadnout rozptyl v každém výběru zvlášť a použít tak statistiku

$$\widetilde{T}_d = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}},$$

známou z přednášky:

```
(Td2=(xCR-xSR)/sqrt(xCR*(1-xCR)/n+xSR*(1-xSR)/nSR))
2*(1-pnorm(abs(Td2)))
```

13. Pro danou situaci bychom mohli použít i asymptotický dvouvýběrový t-test:

```
dataSR=c(rep(1,chlapciSR),rep(0,divkySR))

t.test(data,dataSR)
```

Porovnejte výslednou p-hodnotu s výsledkem funkce prop.test. Jak se liší testové statistiky?

14. Proveďte ručně test založený relativním riziku, kdy testujeme $H_0 : r_X = \frac{p_1}{p_2} = 1$, kde použijeme testovou statistiku

$$T_r = \frac{\log \widehat{p}_1 - \log \widehat{p}_2}{\sqrt{\frac{1}{n}\frac{1-\widehat{p}_1}{\widehat{p}_1}+\frac{1}{m}\frac{1-\widehat{p}_2}{\widehat{p}_2}}}.$$

Spočtěte i asymptotický intervalový odhad r_X.

```
r=xCR/xSR
sd=sqrt(1/n*(1-xCR)/xCR+1/nSR*(1-xSR)/xSR)
(Tr=(log(r))/sd)
2*(1-pnorm(abs(Tr)))

# int. odhad
c(r*exp(-q*sd),r*exp(q*sd))
```

Případně bychom mohli test shody pravděpodobností založit na tzv. poměru šancí

$$o_X = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

a testovat $H_0 : o_X = 1$. Z přednášky víme, že $\dfrac{\log \widehat{o} - \log o_X}{\sqrt{\widehat{V}_0}}$ má asymptoticky $\mathsf{N}(0,1)$, kde

$$\widehat{V}_0 = \frac{1}{n\widehat{p_1}} + \frac{1}{n(1-\widehat{p_1})} + \frac{1}{m\widehat{p_2}} + \frac{1}{m(1-\widehat{p_2})}.$$

```
(o=xCR/(1-xCR)*(1-xSR)/xSR)
V=1/chlapci+1/(n-chlapci)+1/chlapciSR+1/(nSR-chlapciSR)
(To=log(o)/sqrt(V))
2*(1-pnorm(abs(To)))
```

III Testy pravděpodobností v multinomickém rozdělení

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 )
```

15. 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)
```

16. 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.

IV Testy dobré shody se známými parametry

17. 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 vědět, v kterých kategoriích se pozorovaná četnost od testované nulové nejvíce liší, můžeme se podívat na tzv. rezidua

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

V Samostatná práce

1. Rozhodněte, zda lze tvrdit, že jsou pravděpodobnosti narození chlapce a dívky v ČR v poměru 21:20.

2. Mění se pravděpodobnost narození chlapce v čase?
   - V roce 2010 se na Slovensku narodilo živě 60 410 dětí, z nichž bylo 30 544 chlapců a 29 866 děvčat. Zjistěte, zda pravděpodobnost narození chlapce na Slovensku od roku 2010 vzrostla.
   - V roce 2008 se v České republice živě narodilo 119 570 dětí, z toho 58 244 dívek a 61 326 chlapců (zdroj ČSÚ). Zjistěte, zda se pravděpodobnost narození chlapce liší pro roky 2008 a 2015.

3. Odhadněte intervalově, kolikrát je pravděpodobnost narození chlapce v ČR vyšší než pravděpodobnost narození dívky.
Musíte si sami odvodit vzoreček. Bude se hodit delta věta.

VI Doplňující informace pro zájemce

(i) Funkce binom.test nepočítá p-hodnotu podle vzorce ze skript, ale trochu jinak. Uvažujme test hypotézy $H_0 : p_X = 0.14$ proti oboustranné alternativě a data $X_n = 6$ a $n = 20$. Podle definice ze skript bychom p-hodnotu spočetli následovně:

```
Xn=6;n=20;p0=0.14
2*min(pbinom(Xn, size = n, p = p0), 1-pbinom(Xn-1, size = n, p = p0))

binom.test(Xn,n,p=p0)
```

To ale neodpovídá p-hodnotě ve funkci binom.test. Tato funkce považuje za hodnoty, které stejně nebo ještě více svědčí proti H_0, ty hodnoty, jejichž pravděpodobnost napozorování za nulové hypotézy je stejná nebo menší, než co jsme napozorovali ve skutečnosti:

```
qq <- as.logical(dbinom(0:n, size = n, p = p0) <= dbinom(Xn, size=n, p=p0));

# p-hodnota
sum(dbinom(0:n, size = n, p = p0)[qq])
```

(ii) Porovnání hladiny testu a síly statistik T_d a $\widetilde{T}_d$ pro dvouvýběrový problém:

```
opak=1000
n1=20
n2=40
p.T1=rep(NA, opak)
p.T2=rep(NA, opak)
for(i in 1:1000){
  x=rbinom(1,size=n1,prob=1/4)
  y=rbinom(1,size=n2,prob=1/4)
  p.T1[i]=prop.test(c(x,y),c(n1,n2),correct=F)$p.val
#
  var2=(x/n1)*(1-x/n1)/n1+(y/n2)*(1-y/n2)/n2
  T2=(y/n2- x/n1)/sqrt(var2)
  p.T2[i]=2*pnorm(-abs(T2))
}

mean(p.T1<=0.05)
mean(p.T2<=0.05)
```

Takto sledujeme hladiny testu. Když změníme 1/4 v předpisu pro generování jednoho z výběrů, tak dostaneme odhad síly testu.