Testy o proporci

18.12.2018

Jednovýběrový problém pro binární data. V roce 2008 se v České republice živě narodilo 119 570 dětí, z toho 58 244 dívek a 61 326 chlapců (zdroj ČSÚ). Zajímá nás, zda je pravděpodobnost narození chlapce 1/2.

1. Jaký předpokládáme model a jaké budeme testovat hypotézy?

2. Nejprve budeme uvažovat Wilsonův test založený na testové statistice

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

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

```
alpha=0.05
n=119570
divky=58244
chlapci=61326

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

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

   Jaký je nyní 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. Jak je konstruován uvedený interval spolehlivosti ve funkci prop.test? A jaký jiný intervalový odhad se spolehlivostí 95 % byste uměli zkonstruovat?

6. Nyní provedeme test téže hypotézy pomocí přesného testu

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

   Na jakém rozdělení je založený testo test? Jaký je nyní náš závěr?

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?

8. 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)
```

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

10. Na rozmyšlení na doma: Odhadněte intervalově, kolikrát je pravděpodobnost narození chlapce vyšší než pravděpodobnost narození dívky.

Dvouvýběrový problém pro binární data. Na Slovensku se v roce 2010 narodilo 60 410 dětí, z nichž bylo 30 544 chlapců a 29 866 děvčat (zdroj pluska.cz). Zajímá nás, zda je pravděpodobnost narození chlapce stejná v ČR a na Slovensku.

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

12. 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

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

     kde $\widetilde{p}$ je odhad společné pravděpodobnosti úspěchu za nulové hypotézy. Funkce nám ve výstupu dává $\widetilde{T}_d^2$.
   -
```
nSR=60410
chlapciSR=30544
divkySR=29866

prop.test(c(chlapci,chlapciSR),c(n,nSR),correct=FALSE)
```

     Jaký učiníme závěr na základě tohoto testu?

13. Ještě spočítáme testovou statistiku ručně

```
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)))
```

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

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

známou z přednášky:

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

15. Pro danou situaci bychom mohli použít i dvouvýběrový asymptotický t-test (Welchův 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?

16. V roce 1970 se v ČSSR živě narodilo 117 137 chlapců a 111 394 děvčat. Zjistěte, zda se pravděpodobnost narození chlapce v ČR v roce 2008 liší oproti situaci v ČSSR v roce 1970.

Testy pro multinomické rozdělení

17. V rámci přednášky pro studenty chemie PřF UK v letech 2006-2013 bylo zjišťováno mimo jiné, v jakém měsíci slaví narozeniny. Naměřena byla následující data:

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

Data zapsaná v R formátu:

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

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 $\mathbf{p} = (p_1, \ldots, p_{12})^\top$.

- Odhadněte parametry tohoto multinomického rozdělení.
- Formulujte nulovou a alternativní hypotézu.
- 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 $\mathbf{V} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$.
- Pomocí R test 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 %*%. Matici $\mathbf{V}$ vytvoříme následovně

```
V=diag(p)-p%*%t(p)
```

18. Otestujte podobně, zda jsou pravděpodobnosti narození dítěte v I. a III. čtvrtletí stejné. Formulujte opět nulovou a alternativní hypotézu a proveďte ručně vhodný test.

19. Studenti dále uváděli počet svých sourozenců. Z 326 studentů 50 uvedlo, že nemá žádného sourozence, 183 má jednoho a zbytek má dva a více sourozenců. Otestujte zda je pravděpodobnost jednoho sourozence 3 krát větší než pravděpodobnost žádného sourozence.

Samostatná práce

(i) V aplikacích (nebo např. v domácím úkolu) se nám může hodit umět převést vektor spojitých hodnot na kategoriální veličinu. Např. uvažujme vektor x, který si nagenerujeme z normálního rozdělení $\mathsf{N}(0,1)$, pro který chceme vytvořit vektor s, který bude identifikátor toho, zda je $x_i > 0$. Ten pak můžeme ještě dále převést na tzv. typ factor:

```
(x=rnorm(20,0,1))
(s=ifelse(x>0,1,0)) # 1 priradime tam, kde x>0 a jinde bude 0
summary(s)

sf=factor(s)
summary(sf)
```

Všimněte si rozdílného výstupu funkce summary. Počet „úspěchů“ ve vektoru s pak spočteme jako sum(s) nebo sum(sf=="1").

(ii) Ve skriptech máte kromě Wilsonova intervalu spolehlivosti a přesného intervalu spolehlivosti pro proporci uvedený také interval spolehlivosti založený na logitu a klasické asymptotické metodě (dále Waldův interval). Stáhněte si z internetu 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?

(iii) Funkce binom.test nepočítá p-hodnotu podle vzorce ze skript na str. 112. 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])
```

(iv) Sestrojte intervalový odhad podílu pravděpodobností narození chlapce v ČR a SR.

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

```
opak=1000
n1=20
n2=40
p.T1=numeric(opak)
p.T2=numeric(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.