Testy o proporci
12.12.2019

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, pozor na zjevnou chybu v posledním řádku tabulky). 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 $\mathsf{N}(0,1)$.

   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))
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. 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.

11. 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. 45). Zjistěte, zda je pravděpodobnost narození chlapce na Slovensku vyšší než 1/2.

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

prop.test(chlapciSR,nSR,alternative="greater")
```

12. 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?

DVOUVÝBĚROVÝ PROBLÉM PRO BINÁRNÍ DATA. Zajímá nás, zda je pravděpodobnost narození chlapce stejná v ČR a na Slovensku.

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

14. 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 $\widetilde{T}_d^2$.
    - Provedeme test:
      `prop.test(c(chlapci,chlapciSR),c(n,nSR),correct=FALSE)`

    Jaký učiníme závěr na základě tohoto testu?

15. 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)))
```

16. 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)))
```

17. 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?

18. Proveďte ručně test založený relativním riziku, kdy testujeme $H_0 : r_X = p_1/p_2 = 1$, 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 $\dfrac{\log \widehat{o} - \log o_X}{\widehat{V}_0^{1/2}}$ 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)))
```

19. 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. Proveďte jak test založený na rozdílu pravděpodobností, tak test založený na relativním riziku, tak i test založený na poměru šancí.

20. 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.

21. V roce 2014 se narodilo 56 636 chlapců, z nichž 56 410 živě. Dívek se narodilo 53 616, přičemž 53 450, viz zdroj. Liší se proporce živě narozených jedinců mezi dívkami a chlapci?

Doplňky

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

(iii) 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.