Testy o proporci a testy v multinomickém rozdělení
21.12.2017

Úvodní nastavení.

– Z internetové stránky www.karlin.mff.cuni.cz/~hudecova/education/ si můžete stáhnout zdrojový kód cviceni12.R.

– Otevřete si program R Studio a proveďte úvodní nastavení jako vždy:

```
setwd("H:/nmsa331")
rm(list=ls())
alpha=0.05
```

Jednovýběrový problém pro binární data

1. Načtěte si data Hosi.txt. Z nich nás opět bude zajímat porodní hmotnost, a to konkrétně náhodný podvýběr o rozsahu 200:

```
Hosi = read.table("Hosi.txt",header=T)
set.seed(21122017)
hmot=sample(Hosi$por.hmot,200)
```

   Podvýběr děláme proto, abychom si mohli nechat vypsat všechny analyzované hodnoty (a případné nové proměnné) a bylo to přehledné.

2. Novorozenec s váhou nižší než 2500 g bývá označován za novorozence s nízkou porodní hmotností. Na stránce wikiskripta.eu se uvádí, že v roce 2009 mělo nízkou porodní hmotnost 7.8% dětí. Naším úkolem je zjistit, zda jsou naše data v souladu s touto informací.
   Zavedeme si novou proměnnou NPH, která může nabývat dvou hodnot: nph pro chlapce s nízkou porodní hmotností a ok pro ostatní.

```
NPH=factor(ifelse(hmot<2500,"nph","ok"))

sum(NPH=="nph")
(tabulka=table(NPH) )
(Ptabulka=prop.table(table(NPH)) )# relativni cetnosti

round(Ptabulka*100,2)
```

   Takto jsme si nechali vypsat tabulku četností a tabulku relativních četností. Rovnou jsme si je i uložili, protože se nám budou pro další práci hodit. Ještě si vše znázorníme graficky:

```
pie(tabulka)
pie(tabulka, labels=c("Nizka hmotnost","Normalni hmotnost"))

barplot(tabulka, ylab = "Cetnost")
barplot(tabulka,ylab="Cetnost",names=c("Nizka hmotnost","Normalni hmotnost"))

barplot(Ptabulka,ylab="Relativni cetnost",names=c("Nizka hmotnost",
           "Normalni hmotnost"))
```

3. Co odhadují relativní četnosti a v jakém modelu? Uložíme si tento odhad do proměnné phat:

```
(phat =Ptabulka[1] )
# totez jako:
Ptabulka["nph"]
```

4. Přistoupíme k testu výše uvedené domněnky o výskytu dětí s nízkou porodní hmotností.
   – Jaký předpokládáme model a jak budeme formulovat hypotézy?
   – Nejprve budeme uvažovat asymptotický Wilsonův test, který je založený na testové statistice
   $$W = \sqrt{n}\frac{\widehat{p}_n - p_0}{\sqrt{p_0(1-p_0)}},$$
   která má asymptoticky normální rozdělení.
   – Test provedeme:
   ```
   p0 <- 0.078
   (Xn <- tabulka["nph"])
   n <- length(NPH)

   prop.test(x = Xn, n = n, p = p0, correct = FALSE)

   #nebo primo
   prop.test(tabulka,p=p0,correct=FALSE)
   ```
   Jaký je náš závěr? Jsou naše data v souladu s uváděným procentuálním výskytem novorozenců s nízkou hmotností?
   – Jakou testovou statistiku nám zde uvádí prop.test? A jaké má asymptotické rozdělení? Ověříme vše manuálním výpočtem:
   ```
    W <- (phat - p0)/sqrt(p0*(1-p0)/n)
   W^2
   # p-hodnota
   2*pnorm(-abs(W))
   1 - pchisq(W^2, df=1)
   ```

   A jak se spočítá uvedený interval spolehlivosti?

5. Nyní provedeme test téže hypotézy pomocí přesného testu:

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

   Jaký je nyní náš závěr?
   Připomeňte si, že přesný test je založen na binomickém rozdělení a že nulovou hypotézu zamítáme pro velmi malé a velmi velké hodnoty. Tedy, zamítáme, pokud $X_n \leq C_1$ nebo $X_n \geq C_2$, kde

   ```
   (C1 <- qbinom(alpha/2, size=n, prob=p0))
   (C2 <- qbinom(1-alpha/2, size=n, prob=p0) - 1)
   ```

Odpovídá to výsledku funkce binom.test?

6. Jelikož alternativní rozdělení splňuje předpoklady centrální limitní věty, můžeme použít i asymptotický t-test. K jeho provedení musíme nejprve převést naši proměnnou NPH na 0-1 veličinu:

```
NPH2=abs(as.numeric(NPH)-2)

t.test(NPH2,mean=p0)
```

7. Podle zprávy Hospodářských novin ze dne 16.12.2017 (zpráva *Na českých vysokých školách přibývá studentů z bývalého Sovětského svazu a pomalu vytlačují Slováky*) na českých vysokých školách studuje 14 % cizinců. Na základě dat z našeho cvičení otestujte domněnku, že na matfyzu je to statisticky významně více. Na toto cvičení chodí 20 studentů, z nichž 6 je ze Slovenska (tj. jsou cizinci).
Jaký model předpokládáme? Jak vypadají hypotézy? Je vhodnější použít přesnou variantu testu nebo asymptotickou verzi?

DVOUVÝBĚROVÝ PROBLÉM PRO BINÁRNÍ DATA

8. Bude nás zajímat, zda je pravděpodobnost narození dítěte s nízkou porodní hmotností stejná pro ženy pod 35 let a ženy, které mají alespoň 35 let. Pro zkoumání tohoto problému použijeme všechna dostupná data. Nejprve provedeme přípravu dat, tj. zavedeme novou veličinu, která nám bude kategorizovat matky na mlade a stare podle toho, zda je jejich věk menší než 35 let nebo nikoliv.

```
Fmatka=factor(ifelse(Hosi$vek.matky>=35,"stara","mlada"))
NPH=factor(ifelse(Hosi$por.hmot<2500,"nph","ok"))
```

Podíváme se na počty případů v jednotlivých kategoriích a další vhodné charakteristiky:

```
tapply(NPH, Fmatka,summary)
table(NPH, Fmatka)
table(Fmatka,NPH)
prop.table(table(Fmatka,NPH),mar=1) ## margin = 2 --> podminuj sloupecky
                                    ## margin = 1 --> podminuj radky
```

Vše si můžeme i graficky znázornit:

```
plot(NPH~Fmatka)

barplot(table(NPH,Fmatka),beside=T,legend=T)
barplot(prop.table(table(NPH,Fmatka)),beside=T,legend=T)

par(mfrow=c(1,2))
pie(table(NPH,Fmatka)[,1],main="Mlade matky",col=2:3)
pie(table(NPH,Fmatka)[,2],main="Stare matky",col=2:3)
par(mfrow=c(1,1))
```

Co si myslíte o zkoumaném problému na základě těchto údajů a grafů? Závisí pravděpodobnost nízké porodní váhy na věku matky?

9. Provedeme dvouvýběrový test o proporci založený na rozdílu pravděpodobností.
   – Jaký předpokládáme model? Jak zní testované hypotézy?
   – Provedení testu pomocí R funkce:
   ```
   (tabulka=table(NPH,Fmatka))
   (pocty.NPH=tabulka["nph",])
   (pocty.n=table(Fmatka))
   # nebo: pocty.n=margin.table(tabulka,2)

   prop.test(x = pocty.NPH, n = pocty.n, correct = FALSE)

   # nebo jine zadani:
   prop.test(t(tabulka),correct=F)
   ```
   Jaký učiníme závěr na základě tohoto testu?

10. Manuální výpočet testové statistiky:
   ```
   prumer.M=pocty.NPH[1]/pocty.n[1]
   prumer.S=pocty.NPH[2]/pocty.n[2]
   prumer.all=sum(pocty.NPH)/sum(pocty.n)

   var1=prumer.all*(1-prumer.all)*(1/pocty.n[1]+1/pocty.n[2])
   (T1=(prumer.S-prumer.M)/sqrt(var1))

   T1^2
   #porovname s
   prop.test(t(tabulka),correct=F)$stat

   2*pnorm(-abs(T1))
   prop.test(t(tabulka),correct=F)$p.val
   ```
   Alternativně bychom mohli odhadnout rozptyl v každém výběru zvlášť:
   ```
   var2=prumer.M*(1-prumer.M)/pocty.n[1]+prumer.S*(1-prumer.S)/pocty.n[2]
   (T2=(prumer.S-prumer.M)/sqrt(var2))
   2*pnorm(-abs(T2))
   ```

11. Pro danou situaci bychom mohli použít i dvouvýběrový asymptotický t-test
   ```
   NPH2=abs(as.numeric(NPH)-2)
   t.test(NPH2~Fmatka)
   ```
   Porovnejte p-hodnotu a interval spolehlivosti s předchozím výsledkem funkce prop.test.

12. Poznámka pro uživatele LaTeX: Tabulku z R snadno převedeme do LaTeX-u pomocí funkce xtable z knihovny xtable následovně:

```
library(xtable)
xtable(tabulka)
```

Testy dobré shody pro multinomické rozdělení se známými parametry

13. V rámci přednášky pro studenty chemie PřF MFF 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

Na základě těchto dat ověřte, zda se lidé rodí rovnoměrně během roku, nebo dochází k nějakému systematickému porušení této rovnoměrnosti.

(a) Jaký model budeme předpokládat a jak budeme formulovat nulovou hypotézu?

(b) Nejprve si musíme spočítat teoretické pravděpodobnosti narození v jednotlivých měsících, které jsou za nulové hypotézy rovny relativnímu počtu dní v daném měsíci vzhledem k celkovému počtu dní v roce. Pro jednoduchost zanedbáme přestupné roky a budeme brát 365 dní v roce a pro únor 28 dní. Dále si do vektoru `x` uložíme příslušné počty z tabulky výše:

```
tf=c(31,28,31,30,31,30,31,31,30,31,30,31)/365

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

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

Jaký je náš závěr ohledně rozložení data narození v průběhu roku?

14. 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 domněnku, že jedináčci se vyskytují v poměru k osobám s jedním sourozencem a osobám s více než dvěma sourozenci v poměru 1:3:1.

Samostatná práce

1. Mezi 325 studenty chemie bylo jen 21 cizinců. Otestujte, zda jsou tato data v souladu s dříve uvedeným tvrzením, že na českých vysokých školách studuje 14 % cizinců.

2. 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?

3. Odhadněte relativní riziko (tj. poměr rizik) pro narození chlapce s nízkou porodní hmotností pro matky pod 35 let a nad 35 let včetně. Pro sestrojení intervalového odhadu využijte vzorec ze skript na str. 109.

4. Na datech z příkladu 7. ověřte, že R ve funkci binom.test nepočítá p-hodnotu podle vzorce na str. 109. Podle této definice 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])
```

5. Porovnání hladiny testu a síly statistik T_d a $\tilde{T}_d$ pro dvouvýběrový problém:

```
opak=1000
n1=50
n2=80
p.T1=numeric(n)
p.T2=numeric(n)
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.