Neparametrické jednovýběrové testy a párové testy
7.12.2017

Úvodní nastavení.

- Z internetové stránky `www.karlin.mff.cuni.cz/~hudecova/education/` si můžete si stáhnout zdrojový kód k dnešnímu cvičení cviceni10.R.
- Otevřete si program R Studio, změňte si pracovní adresář a vyčistěte pracoviště
```
setwd("H:/nmsa331")
rm(list=ls())
```
- Načtěte si data Hosi.txt a ujistěte se, že se Vám data dobře načetla.
- Do proměnné alpha is uložte testovací hladinu 0.05.
- Opět se budeme zabývat porodní hmotností a opět budeme nejprve pracovat pouze s (náhodným) podvýběrem o rozsahu $n = 100$ pozorování.
```
set.seed(07122017)
hmot100 = sample(Hosi$por.hmot, 100)
n=length(hmot100)

hmot=Hosi$por.hmot
```

Jednovýběrový znaménkový test

1. Minule jsme se zabývali otázkou, zda je průměrná porodní hmotnost chlapců rovna 3,3 kg. Zopakujte tuto analýzu (pomocí t-testu) na „dnešní“ data.
   - Připomeňte si pravděpodobnostní model, který jsme uvažovali a nulovou i alternativní hypotézu.
   - Zformulujte závěr.
2. Nyní se budeme snažit zodpovědět tutéž otázku pomocí znaménkového testu.
   - Připomeňte si předpoklady tohoto testu (tj. uvažovaný model pro naše data).
   - Formulujte nulovou a alternativní hypotézu.
   - Testová statistika má tvar
   $$Y_n = \sum_{i=1}^{n} \mathrm{I}[X_i > m_0],$$
   kde m_0 je námi testovaná hodnota 3300 g. Jaké má přesné a jaké asymptotické rozdělení?

   Provedeme asymptotickou variantu znaménkového testu nejprve „manuálně“. Nejprve musíme vyřadit pozorování, která jsou přesně rovna 3300, a budeme pracovat jen se zbylými daty.
```
sum(hmot100 == 3300)

(nLess <- sum(hmot100 < 3300))
(nMore <- sum(hmot100 > 3300))

(N <- nLess + nMore)
```

Naše analýza tedy bude založena na N počtu pozorování, v proměnné nMore je hodnota testové statistiky Y_n. Pro provedení asymptotického testu si ji příslušně znormujeme. Následně ji porovnáme s kvantilem normálního rozdělení, nebo spočítáme přímo p-hodnotu:

```
(Z = (nMore - N/2) / sqrt(N*1/4))
qnorm(1-alpha/2)

# p- hodnota (dve moznosti, jak spocitat):
2*(1-pnorm(abs(Z)))
2*pnorm(-abs(Z))
```

K jakému závěru jsme došli?

3. Celý výpočet můžeme provést rychleji pomocí funkce prop.test, která slouží k testování pravděpodobnosti (proporce) v alternativním rozdělení.

```
prop.test(x=nMore, n=N,p=1/2,correct=FALSE)
```

Všimněte si, že jsme dostali opravdu stejnou p-hodnotu. Testová statistika je ale jiná. Najdete souvislost s naším výsledkem uloženým v Z? Jaké asymptotické rozdělení má tedy testová statistika funkce prop.test?
Nastavení correct=FALSE zakazuje tzv. korekci pro spojitost, která je jinak defaultně zapnutá. Více v samostatném cvičení č. 3.

4. Vyzkoušejte, co by se stalo, kdybychom celý výpočet prováděli s nLess namísto nMore.

5. Jelikož znaménkový test převádí naše naměřená data na binární, podíváme se na to, jak vizualizovat kategoriální veličinu:

```
Y <- hmot100[hmot100 != 3300]; # vybereme hmotnosti ruzne od 3300
Y <- factor(Y > 3300) # udelame z nich 0-1 veliciny a chapeme je jako "faktor"
levels(Y) <- c("< 3300", "> 3300") # pojmenujeme kategorie
summary(Y)   # kontrola s nLess a nMore

plot(Y,col=2:3);
# totez jako:
barplot(table(Y),col=2:3)
# nebo
barplot(c(nLess, nMore), col=2:3,legend=c("< 3300", "> 3300"))

#proporcionalne:
pie(table(Y),col=2:3)
# nebo
barplot(prop.table(as.matrix(c(nLess, nMore))), col=2:3, legend=c("< 3300", "> 3300"))
```

Jednovýběrový Wilcoxonův test

6. Stále se budeme zabývat otázkou, zda je průměrná porodní hmotnost chlapců rovna 3,3 kg. Nyní k testu použijeme Wilcoxonův test

(a) Jaký model předpokládáme pro naše data nyní?

(b) Jak budeme teď formulovat testované hypotézy?
   - Porovnejte předpoklady Wilcoxonova testu s předpoklady přesného t-testu a s předpoklady asymptotického t-testu.
   - Porovnejte, kdy jsou testované hypotézy ekvivalentní a kdy naopak nejsou.
   - Uveďte příklad situace, pro kterou můžeme použít Wilcoxonův test a t-test by nebyl vhodný, a situace opačné.

(c) Připomeňte si testovou statistiku tohoto testu: Je-li $Z_i = X_i - \delta_0$, kde δ_0 je pro nás 3300, pak

$$W = \sum_{i \in I} R_i,$$

kde R_i jsou pořadí $|Z_i|$ a I je množina indexů i takových, že $Z_i > 0$. Test můžeme provést jak přesný (pro rozumně malé n; viz samostatné cvičení č. 4.) nebo asymptotický založený na tom, že

$$\frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

má asymptoticky $\mathsf{N}(0,1)$.
Pro provedení testu musíme opět nejprve odstranit pozorování rovné přímo δ_0. V případě výskytu shod je potřeba jmenovatel zmenšit o tzv. korekci rozptylu (viz přednáška, str. 82). V R se toto děje defaultně. Navíc se uvažuje ještě korekce pro spojitost, kdy se jmenovatel snižuje o 1/2 podobně jako u znaménkového testu (viz samostatné cvičení č.3.). Toto lze vypnout pomocí correct=FALSE.

(d) Provedeme test
```
wilcox.test(hmot100, mu = 3300,correct=FALSE)
```
Jaký je náš závěr?

7. Ověříme, že uvedená testová statistika je skutečně W uvedené výše

```
x <- hmot100 - 3300
x <- x[x != 0]    ### nebudeme pouzivat porodni hmotnosti 3300
(R <- rank(abs(x)))
(Sign <- -1*(x < 0) + 1*(x > 0))
(Splus  <- sum(R[x > 0]))
```

8. Vypočítáme p-hodnotu pro test bez korekce i s korekcí pro spojitost. Vzorec pro korekci byl na přednášce, viz str. 82.

```
ESplus <- N*(N+1)/4
# totez jako 1/2*sum(1:N)
varSplus_NoTies <- N * (N + 1) * (2 * N + 1) / 24

#vypocet korekce:
(tx <- table(R)) # Pocet jednotlivych hodnot
varCorrectWithTies <- sum(tx^3 - tx) / 48   # Korekce v pripade shod
varSplus <- varSplus_NoTies - varCorrectWithTies
```

```
(U <- (Splus - ESplus) / sqrt(varSplus))
(U.corr <- (Splus - ESplus - 0.5) / sqrt(varSplus))


### P-hodnoty
(P          <- 2*pnorm(-abs(U)))
(P.corr <- 2*pnorm(-abs(U.corr)))
```

Což můžeme srovnat s p-hodnotami:

```
wilcox.test(hmot100, mu = 3300, correct = FALSE)$p.val
wilcox.test(hmot100, mu = 3300)$p.val
```

Výsledné p-hodnoty zkuste sami porovnat s p-hodnotou, kterou bychom dostali bez korekce jmenovatele.

9. Porovnáme sílu t-testu, znaménkového testu a Wilcoxonova testu pro normální rozdělení.

```
opak=1000
n=100

p.t=numeric(opak)
p.z=numeric(opak)
p.w=numeric(opak)
for(i in 1:opak){
 x=rnorm(n,mean=0.2,sd=1)
 p.t[i]=t.test(x,mu=0)$p.val
 p.w[i]=wilcox.test(x,mu=0)$p.val
 p.z[i]=prop.test(sum(x>0),n,p=1/2)$p.val
}

mean(p.t<=0.05)
mean(p.w<=0.05)
mean(p.z<=0.05)
```

Zkuste změnit velikost rozsahu n nebo alternativu, za které simulujeme, a sledujte, jak se mění síla. Lze udělat nějaký obecný závěr?

Párový problém

10. Některé zdroje uvádějí, že chlapci v průběhu prvního roku života přiberou v průměru více než 6,5 kg. Máme za úkol ověřit, zda jsou naše data v souladu s tímto tvrzením.
    (a) Jsou porodní hmotnost a hmotnost v jednom roce nezávislé veličiny? Uveďte další možné příklady párového problému.
    (b) Spočítáme si váhový přírůstek během prvního roku:

```
hmot<- Hosi$por.hmot
hmot1 <- Hosi$hmotnost
rozdil=hmot1-hmot
```

(c) Nejprve použijeme t-test.
  - Jaký model předpokládáme (uvažte pomocí vhodných grafů) a jak zní nulová a alternativní hypotéza?
  - Proveďte test a interpretujte výsledek.

(d) Nyní na stejný problém použijeme znaménkový test.
  - Co nyní předpokládáme? A co testujeme? Je obecně nějaká souvislost mezi testovanou kvantitou a charakteristikami původních párových veličin (porodní hmotnosti a hmotnost v 1 roce)?
  - Provedeme test

```
n.more=sum(rozdil>6500)
n=sum(rozdil!=6500)
prop.test(n.more, n, alternative = "greater")
```

    Jaký je náš závěr?

(e) Nakonec si na tomto problému vyzkoušíme ještě i Wilcoxonův test.
  - Co nyní předpokládáme? Co předpokládáme navíc proti jednovýběrovému Wilcoxonovu testu? Jaké testujeme hypotézy?
  - Proveďte test.

11. Zajímá nás věkový rozdíl mezi rodiči chlapců. Konkrétně bychom rádi věděli, zda jsou otcové v průměru o více než 2 roky starší než matky.

   (a) Nejprve si prohlédněte vhodné popisné statistiky a grafy, které ilustrují rozdělení věkového rozdílu.
   (b) Zvolte test, který Vám příjde pro tato data nejvhodnější a proveďte jej. Uvědomte si, jaký uvažujete model, jaké testujete hypotézy a jak budeme interpretovat závěr.

Samostatná práce

1. V rámci tohoto cvičení jsme pro test hypotézy, zda je porodní hmotnost chlapců rovna 3,3 kg, použili několik rozličných testů, které v praxi mohou vést k různým p-hodnotám a různým závěrům. Kterému výsledku tedy máme „věřit“?

2. Pomocí znaménkového testu proveďte test hypotézy, zda je porodní hmotnost nižší než 3,5 kg. Zformulujte řádně nulovou a alternativní hypotézu a proveďte test.

3. Vyzkoušejte si znaménkový test se zapnutou korekcí pro spojitost. Ověřte, že v tomto případě se počítá testová statistika
$$Z^C = \frac{Y_n - \frac{n}{2} - \frac{1}{2}\mathrm{sgn}(Y_n - \frac{n}{2})}{\sqrt{\frac{n}{4}}}$$
(resp. její druhá mocnina). Použijte postup analogický tomu v bodě 2.
Korekce pro spojitost zde vychází z toho, že Y_n má diskrétní rozdělení a $\mathsf{P}(Y_n \leq y) = \mathsf{P}(Y_n < y+1)$. Tudíž při přechodu k asymptotické aproximaci spojitým normálním rozdělením s distribuční funkcí F lze $\mathsf{P}(Y_n \leq y)$ aproximovat jak $F(y+\varepsilon)$ pro všechna $\varepsilon \in [0,1]$. Jako určitý kompromis tedy vezmeme $F(y+1/2)$.

4. Pro Wilcoxonův test můžeme použít i přesný test, založený na přesném rozdělení W. Je však nutné, aby v datech nebyly shody a žádné pozorování nebylo rovno přesně testované hodnotě. Pro vyšší n může být ale výpočet časově náročný. Vyzkoušejte si to na simulovaných datech z rovnoměrného rozdělení:

```
x=runif(5000,0,1)
wilcox.test(x[1:100],mu=1/2,exact=TRUE)
```

   Postupně zvyšujte počet použitých dat ze 100 na 200,500, 1000, 2000.

5. Porovnejte sílu jednovýběrových testů pro rovnoměrné rozdělení (stejně jako v 9.). Například simulujte data z $\mathsf{R}[0, 1.2]$ a testujte, zda je střední hodnota (medián) rovna $1/2$.

6. Znaménkový test má kromě asymptotické verze (kterou jsme dělali výše) i přesnou variantu založenou na binomickém rozdělení. V R můžeme použít funkce `binom.test`, parametry této funkce jsou pak stejné jako u `prop.test`. Proveďte pro porovnání test hypotézy, zda je porodní hmotnost nižší než 3,5 kg, i pomocí přesného testu. Vyzkoušejte použití našich 100 vybraných hodnot i všech dostupných dat.