Dvouvýběrové testy

14.12.2017

Úvodní nastavení.

– Z internetové stránky www.karlin.mff.cuni.cz/~hudecova/education/ si stáhněte data Iq2.txt a zdrojové kódy cviceni11.R a figKS.R.
– Otevřete si program R Studio, změňte si pracovní adresář a vyčistěte pracoviště

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

– Do proměnné alpha is uložte testovací hladinu 0.05.

Posuzování normality a symetrie pomocí grafů: Je potřeba vzít v úvahu, že pracujeme s konečnými daty a že i v případě dat generovaných přímo z normálního rozdělení nedostaneme „ideální“ grafy. Vyzkoušejte několikrát za sebou spustit následující příkazy a sledujte tvar histogramů a Q-Q grafů.

```
x=rnorm(50)
par(mfrow=c(1,2))
hist(x)
qqnorm(x)
qqline(x)
```

Vyzkoušejte, že i při vyšším rozsahu výběru můžeme dostat na první pohled nesymetrické histogramy a že body Q-Q grafu neleží perfektně na přímce.

Data. Data Iq2.txt obsahují informace o náhodně vybraných žácích osmé třídy základní školy. Pro každého máme k dispozici jeho IQ, pohlaví (kódování 1 dívky a 0 chlapci) a dále pak průměrnou známku na vysvědčení v pololetí sedmé třídy a v pololetí osmé třídy.

IQ	hodnota IQ,
ZN7	průměrná známka na vysvědčení v pololetí sedmé třídy,
ZN7	průměrná známka na vysvědčení v pololetí osmé třídy,
Pohlavi	pohlaví žáka (1- žena, 0 - muž)
FPohlavi	jinak kódované pohlaví žáka (kategorie dívka, chlapec).

Dvouvýběrový problém

1. Načtěte si data Iq2.txt a ujistěte se, že se Vám data dobře načetla.

```
Iq = read.table("Iq2.txt",header=T)
summary(Iq)

attach(Iq)
```

Pohlédněte si základní charakteristiky polohy jednotlivých veličin pomocí funkce summary. Všimněte si rozličného výstupu pro veličiny Pohlavi a FPohlavi.
Příkaz attach nám zpřístupní jednotlivé proměnné z dat, takže je budeme moci volat přímo jejich názvem (např. IQ namísto Iq$IQ).

2. Pomocí vhodného obrázku vizualizujte genderové složení našeho datového vzorku.

3. Budeme se zabývat otázkou, zda se nějak liší IQ u chlapců a dívek. Začneme nejprve prohlídkou popisných statistik.

```
IQchlapci=IQ[FPohlavi=="chlapec"]
IQdivky=IQ[FPohlavi=="divka"]

#charakteristiky polohy dle pohlavi
summary(IQchlapci)
summary(IQdivky)

# nebo rychleji pomoci 1 prikazu
tapply(IQ,FPohlavi,summary)

# podminene sd
tapply(IQ,FPohlavi,sd)
```

   Co usuzujete na základě těchto číselných charakteristik?

4. Problém si ještě budeme vizualizovat:

```
boxplot(IQ~FPohlavi)

# totez jako
boxplot(IQchlapci,IQdivky,names=c("chlapci","divky"))

# muzeme vybarvit a doplnit popisky:
boxplot(IQ~FPohlavi,ylab="IQ",col=c("lightblue","pink"))
```

   Co usuzujeme z tohoto obrázku?

5. Nyní se podíváme na to, zda je možné (resp. vhodné) předpokládat, že oba výběry pocházejí z normálního rozdělení.

```
par(mfrow=c(1,2))
hist(IQchlapci,prob=T,xlim=c(60,150))
hist(IQdivky,prob=T,xlim=c(60,150))

par(mfrow=c(1,2))
qqnorm(IQchlapci,main="Chlapci")
qqline(IQchlapci)
qqnorm(IQdivky,main="Divky")
qqline(IQdivky)
par(mfrow=c(1,1))
```

Dvouvýběrový Kolmogorovův-Smirnovův test

6. Pomocí Kolmogorovova-Smirnovova testu budeme testovat, zda je rozdělení IQ chlapců i dívek stejné.

(a) Co předpokládáme za model?
(b) Jak zní nulová a alternativní hypotéza?
(c) Provedeme test:
```
ks.test(IQdivky, IQchlapci, exact = FALSE)
```
Jaký je náš závěr?

7. Připomeňte si, jak se spočítá testová statistika tohoto testu. Ověříme, že R počítá podle stejného předpisu:

```
e1=ecdf(IQchlapci)
e2=ecdf(IQdivky)
max(abs(e1(IQ)-e2(IQ)))
```

Kdybychom chtěli zjistit, kde maximum nastává:

```
IQ[which.max(abs(e1(IQ)-e2(IQ)))]
```

Celou situaci si můžeme i vizualizovat pomocí následujícího obrázku:

```
source("figKS.R")
figKS(IQ, FPohlavi)
```

Dvouvýběrový t-test

8. Pomocí dvouvýběrového t-testu budeme testovat shodu středních hodnot IQ u chlapců a dívek. Nejprve použijeme verzi pro shodné rozptyly.

(a) Jaký předpokládáme model? Jsou tyto předpoklady pro naše data reálná? Který z předpokladů lze do určité míry ignorovat a test zůstane v platnosti alespoň asymptoticky? Nesplnění kterého předpokladu může naopak vést k nedodržení předepsané hladiny testu?
(b) Napište, jaké hypotézy testujeme a porovnejte je s hypotézami K-S testu.
(c) Připomeňte si tvar testové statistiky tohoto testu.
(d) Provedeme test:
```
t.test(IQ ~ FPohlavi, var.equal = TRUE)
# nebo jiny zpusob zadani:
t.test(IQdivky, IQchlapci, var.equal = TRUE)
```
Jaký je náš závěr ohledně středních hodnot IQ? Co odhaduje uvedený interval spolehlivosti?

9. Spočítáme hodnotu testové statistiky manuálně:

```
xmean=mean(IQdivky)
ymean=mean(IQchlapci)
nx=length(IQdivky)
ny=length(IQchlapci)
S2=1/(nx+ny-2)*((nx-1)*var(IQdivky)+(ny-1)*var(IQchlapci))

(T.stat=(ymean-xmean)/(sqrt(S2*(1/nx+1/ny))))
```

```
# kriticka hodnota
qt(1-alpha/2,df=nx+ny-2)
#p-hodnota
2 * pt(-abs(T.stat), df = nx+ny-2)
```

10. Nyní si vyzkoušíme jinou verzi t-testu, která předpoklad shody rozptylů neuvažuje. Jde o tzv. Welchův test.
    (a) V čem se liší naše předpoklady nyní oproti předpokladům v 8.? Liší se nějak hypotézy?
    (b) Provedeme test:
    ```
    t.test(IQ ~ FPohlavi)
    #nebo
    t.test(IQchlapci,IQdivky)
    ```
    Jaký je náš závěr? Jaké rozdělení má asymptoticky uvedená testová statistika? A na základě jakého rozdělení je spočtená p-hodnota? Co stačí předpokládat pro asymptotickou verzi testu?
    Jak vidíme, tato verze t-testu je v R nastavena defaultně. Jelikož předem většinou nevíme, zda je možné předpokládat shodu rozptylů zkoumaných dvou výběrů, je lepší rovnou použít tuto obecnější verzi testu. V případě, že rozptyly ve skutečnosti shodné jsou, má tento obecnější test jen zanedbatelně nižší sílu ve srovnání s testem odvozeným pro výběry se shodným rozptylem, viz simulace ve 12.

11. Opět provedeme manuální výpočet testové statistiky

```
(T.W=(ymean-xmean)/(sqrt(var(IQdivky)/nx+var(IQchlapci)/ny)))
```

12. Na základě následujících simulací porovnáme sílu t-testu se shodou rozptylů a Welchova testu pro situaci, kdy shoda rozptylů platí

```
opak=1000
n1=100
n2=150
p.s=numeric(opak)
p.w=numeric(opak)
for(i in 1:opak){
  x=rnorm(n1,0.2,sd=1)
  y=rnorm(n2,0,sd=1)
  p.s[i]=t.test(x,y,var.equal=TRUE)$p.val
  p.w[i]=t.test(x,y)$p.val
  }
mean(p.s<=0.05)
mean(p.w<=0.05)
```

13. V předchozích simulacích proveďte následující změnu: simulujte `x` za nulové hypotézy, tj. nastavte střední hodnotu na 0, ale změňte směrodatnou odchylku na 3. Tím se podíváme na to, jaká je hladina testů za nulové hypotézy v případě neshody rozptylů. Měli bychom vidět, že varianta t-testu, která shodu předpokládá, nedodržuje předepsanou hladinu testu.

DVOUVÝBĚROVÝ WILCOXONŮV TEST (MANNŮV-WHITNEYŮV TEST)

14. Opět se vrátíme k původní otázce, a to, zda se liší IQ u chlapců a dívek. Nyní použijeme dvouvýběrový Wilcoxonův test.
    (a) Jaký model předpokládáme pro data? Co musí být splněno, aby test správně fungoval?
    (b) Jak zní nulová a alternativní hypotéza?
    (c) Připomeňte si tvar testové statistiky a pro jaké její hodnoty budeme zamítat.
    (d) Provedeme asymptotickou verzi testu (bez korekce pro spojitost):
    `wilcox.test(IQ~FPohlavi, correct = FALSE)`
    Jaký je náš závěr?
    Přesnou verzi testu bychom zavolali pomocí nastavení exact=TRUE v předchozí funkci. Zkuste.

15. Zkusíme manuální výpočet jednotlivých položek testu. Spočítáme testovou statistiku podle vzorce z přednášky:

```
r=rank(IQ)
(W1=sum(r[FPohlavi=="divka"]))
(W2=sum(r[FPohlavi=="chlapec"]))
```

Dostali jsme stejný výsledek jako funkce wilcox.test?
Ještě vypočítáme p-hodnotu pomocí aproximace normálním rozdělením:

```
n1=length(IQdivky)
n2=length(IQchlapci)

EW=n1*(n1+n2+1)/2

#vypocet korekce rozptylu:
t=table(r)
kor=sum(t^3-t)/((n1+n2)*(n1+n2-1))
varW.kor=n1*n2*(n1+n2+1-kor)/12

U=(W1-EW)/sqrt(varW.kor)

2*pnorm(-abs(U)) # p-hodnota
```

Dostali jsme nyní stejnou p-hodnotu jako funkce wilcox.test?

16. Vypočteme Mannovu-Whitneyovu statistiku:

```
x1 <- matrix(rep(IQdivky, n2), ncol = n2) # ve sloupcich IQdivky
x2 <- matrix(rep(IQchlapci, n1), ncol = n2, byrow = TRUE) # v radcich IQchlapi
(U1 <- sum(x1 < x2) + 0.5 * sum(x1 == x2))
(U2 <- sum(x1 > x2) + 0.5 * sum(x1 == x2))
```

(viz poznámka na str. 99 pro případ dat se shodami pozorování).

Vidíme teď nějakou shodu s výstupem funkce wilcox.test?
Výpočet testové statistiky můžeme provést i pomocí for cyklu, který je ale obecně v R mnohem pomalejší než maticové výpočty:

```
U1=0;U2=0

for(i in 1:n1)
  for(j in 1:n2){
    U1=U1+I(IQdivky[i]<IQchlapci[j])+1/2*I(IQdivky[i]==IQchlapci[j])
    U2=U2+I(IQdivky[i]>IQchlapci[j])+1/2*I(IQdivky[i]==IQchlapci[j])
  }
  U1; U2
```

Mezi Wilcoxonovou statistikou a M-W statistickou je jednoznačný vztah známý z přednášky:

```
(n1 * n2 + 0.5 * n1 * (n1 + 1) - W1)            # U1
(n1 * n2 + 0.5 * n2 * (n2 + 1) - W2)            # U2
```

17. V následujících simulacích si porovnáme sílu dnes uvažovaných testů pro několik vybraných situací. Využijeme k tomu funkci simuluj, kterou jsme si načetli spolu s obrázkem K-S testu. Do této funkce zadáme rozsahy uvažovaných výběrů n a m a který z následujících scénářů chceme:

- normal-posun: $X_i \sim \mathsf{N}(0,1)$, $Y_i \sim \mathsf{N}(1,1)$,
- normal-hetero $X_i \sim \mathsf{N}(0,1)$, $Y_i \sim \mathsf{N}(0,3)$,
- normal-exp: $X_i \sim \mathsf{N}(0,1)$ a $Y_i = W_i - 1$, kde $W_i \sim \mathsf{Exp}(1)$,
- normal-exp2: $X_i \sim \mathsf{N}(0,1)$ a $Y_i = W_i - \log 2$, kde $W_i \sim \mathsf{Exp}(1)$.

```
simuluj(n=10, m=30, scenar="normal-posun")

simuluj(n=10, m=30, scenar="normal-hetero")
simuluj(n=100, m=300, scenar="normal-hetero")

simuluj(n=10, m=30, scenar="normal-exp")
simuluj(n=100, m=300, scenar="normal-exp")

simuluj(n=10, m=30, scenar="normal-exp2")
simuluj(n=100, m=300, scenar="normal-exp2")
```

Spusťte si jednotlivé příkazy postupně a zamyslete se, co testujeme a zda generujeme data za nulové hypotézy nebo za alternativy příslušného testu. Co testujeme u Wilcoxonova testu?

Samostatná práce

1. Zjistěte, zda mají dívky v osmé třídě lepší známky než chlapci. Vyberte test vhodný pro naše data a interpretujte výsledek.

2. Zjistěte, zda je rozdíl ve známkách z předchozího bodu větší než 0.5.

3. Předpokládejme, že máme dva nezávislé výběry $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_n$ (oba výběry mají stejný počet pozorování), kde $X_i \sim \mathsf{N}(\mu_1, \sigma_1^2)$ a $Y_i \sim \mathsf{N}(\mu_2, \sigma_2^2)$. Chceme testovat shodu středních hodnot, tj. $H_0 : \mu_1 = \mu_2$. Co když namísto dvouvýběrového t-testu použijeme párový test? Tj. spočteme rozdíly $Z_i = X_i - Y_i$ a pak test založíme na testové statistice

$$T_n = \sqrt{n}\frac{\overline{Z}_n}{S_n^{(Z)}}, \qquad \text{kde} \qquad S_n^{2(Z)} = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z}_n)^2$$

a nulovou hypotézu zamítneme, pokud $|T_n| > t_{n-1}(1 - \alpha/2)$.

4. Rozdělení testové statistiky Welchova testu je aproximováno pomocí t-rozdělení s ν stupni volnosti, kde ν se určí na základě výběrových rozptylů a počtů pozorování v jednotlivých výběrech. Zjistěte, zda R využívá stejný vzorec jako je uvedený ve skriptech na str. 95.

5. Ověřte manuálně výpočet intervalu spolehlivosti ve dvouvýběrovém t-testu se shodou rozptylů.

6. Zajímá nás, zda v osmé třídě dochází k zlepšení prospěchu oproti sedmé třídě. Jaký test je vhodný pro tuto situaci? Proveďte jej a interpretujte závěr.