Analýza rozptylu (jednoduché třídění)
11.1.2018

Úvodní nastavení.
- Z internetové stránky `www.karlin.mff.cuni.cz/~hudecova/education/` si stáhněte data Med.txt.
- Otevřete si program R Studio a načtěte si výše uvedená data

```
setwd("H:/nmsa331")
rm(list=ls())
Med=read.table("Med.txt",header=T)
```

Popis dat. Na pěti různých místech A, B, C, D a E bylo z řeky vyloveno vždy 7 ryb a byla zjišťována koncentrace mědi v jejich játrech. Naměřená data jsou obsažena v datech Med.txt. Otázkou je, zda je znečištění řeky stejné na všech zkoumaných místech nebo zda se nějak významně liší.

1. Prohlédněte si data Med.txt. V analýze budeme pracovat s logaritmem koncentrace, tj. s proměnnou lnCu.
   - Podívejte se, kolik máme pozorování pro jednotlivá místa A, B, C, D a E.
   - Porovnáme průměry a směrodatné odchylky na jednotlivých místech. Vše si znázorníme i graficky.

   ```
   attach(Med)
   tapply(lnCu,Misto,mean)
   tapply(lnCu,Misto,sd)

   boxplot(lnCu~Misto,col="orange")
   ```

2. Na náš problém budeme chtít použít analýzu rozptylu. Připomeňte si, jaké všechny předpoklady tato metoda má a posuďte graficky, zda se zdají být tyto předpoklady v našich datech splněny nebo nikoliv.

3. Dále si připomeňte, na jakých principech je analýza rozptylu založena: co je to celkový součet čtverců, součet čtverců skupin a reziduální součet čtverců.
   Jednotlivé body a průměry si můžeme znázornit i graficky:

   ```
   (prumery=tapply(lnCu,Misto,mean))
   plot(lnCu~as.numeric(Misto), ,col="gray40", cex=1,pch=19,
      xlab="Misto",xaxt="n",xlim=c(0.5,5.5))
   mtext(levels(Misto),1,line=1.2,at=1:5)
   points(prumery~c(1:5),col="blue",pch=17,cex=1.2)
   (celk.prumer=mean(lnCu))
   abline(h=celk.prumer,col="red",lwd=2)
   ```

4. Otestujte, zda je znečištění řeky na zkoumaných pěti místech stejné. Formulujte H_0 a H_1. Test provedeme následovně:

```
model<-aov(lnCu~Misto)
anova(model)
#totez jako
summary(model)
```

5. Nyní si spočítáme jednotlivé položky z tabulky manuálně.

```
(ni=table(Misto))
(N=sum(ni))
(SSc=sum((lnCu-celk.prumer)^2) )
(SSa=sum(ni*(prumery-celk.prumer)^2))
(SSe=sum((lnCu-fitted(model))^2))
# nebo zde taky takto:
(SSe=sum((lnCu-rep(prumery,7))^2))

p=length(levels(Misto))

SSa/(p-1)
SSe/(N-p)

# testova statistika
(Fa=SSa/(p-1)/(SSe/(N-p)))

# p-hodnota
1-pf(Fa,df1=p-1,df2=N-p)
```

6. F-test analýzy rozptylu zamítl nulovou hypotézu shody středních hodnot a prokázal, že existují (alespoň dvě) místa v řece, kde se znečištění statisticky významně liší. Přirozeně nás tedy zajímá, která místa řeky se od sebe významně liší co se týče znečištění. Jinými slovy nás zajímá porovnání znečištění v místě i a místě j pro všechna $i \neq j$.
   (a) Jak bychom postupovali, kdybychom chtěli provést pouze jedno porovnání míst A a B?
   (b) Kolik porovnání musíme provést, když chceme porovnat všechny dvojice?
   (c) Uvažujte hypotetickou situaci, kdy uvažujeme sto různých nulových hypotéz a jim odpovídajících sto různých testů. Nechť jsou tyto testy nezávislé a všechny provádíme na hladině 5%. Předpokládejme, že všechny nulové hypotézy platí.
      – Kolik budeme očekávat falešných zamítnutí?
      – Jaká je pravděpodobnost, že alespoň jeden test falešně zamítne platnou nulovou hypotézu?
      – Jak se předchozí čísla změní, budeme-li uvažovat 20 hypotéz a jim příslušných nezávislých testů?

   Z výše uvedených důvodů provedeme mnohonásobné porovnání pomocí Bonferroniho metody. Porovnáme tedy všechny dvojice (i, j) na hladině významnosti α/m, kde m je počet porovnání z (b).

```
lev.mista=levels(Misto)
```

```
alpha=0.05
m=5*4/2

#vsechny testy na hladine:
alpha/m
for(i in 1:4) for(j in (i+1):5){
  print(paste(lev.mista[i],"-",lev.mista[j]))
  print(t.test(lnCu[Misto==lev.mista[i]],lnCu[Misto==lev.mista[j]])$p.val,var.equal=T)
}
```

Které místa se významně liší?
Můžeme si vytvořit i přehlednější výstup:

```
tabulka=matrix(0,ncol=4,nrow=m)
rownames(tabulka)=1:10
k=1
for(i in 1:4) for(j in (i+1):5){
  rownames(tabulka)[k]=paste(lev.mista[i],"-",lev.mista[j])
  test=t.test(lnCu[Misto==lev.mista[i]],lnCu[Misto==lev.mista[j]],
     conf.level = 1-alpha/m,var.equal=T)
  int=test$conf.int
  tabulka[k,1]=test$estimate[1]-test$estimate[2]
  tabulka[k,2]=int[1]
  tabulka[k,3]=int[2]
  tabulka[k,4]=ifelse(test$p.val<alpha/m,1,0)
  k=k+1
  }
tabulka
```

7. Při vyšším počtu porovnání (tj. vyšším počtu kategorií) bývá Bonferroniho metoda příliš přísná. Existují tedy i specializované metody pro mnohonásobné porovnání, např. tzv. Tukeyho metoda.

   ```
   TukeyHSD(model)
   plot(TukeyHSD(model))
   ```

   Na kterých místech v řece se znečištění statisticky významně liší?

8. V případě, že nelze předpokládat shodu rozptylů, můžeme namísto F-testu analýzy rozptylu použít Welchovu variantu testu, která je jakýmsi zobecněním Welchova dvouvýběrového t-testu pro více výběrů, viz skripta str. 129. Tento test využívá asymptotické chování příslušné testové statistiky pro $n_i \to \infty$, $n_i/N \to \lambda > 0$. V našem případě máme $n_i = 7$, takže užití asymptotického testu není vhodné, ale ukážeme si, jak bychom jej v R zavolali:

   ```
   oneway.test(lnCu~Misto)
   ```

   V případě zamítnutí nulové hypotézy tímto testem, pak opět provedeme mnohonásobné porovnání, a to pomocí Bonferroniho metody a Welchova testu (Tukeyho metoda předpokládá shodu rozptylů, takže ji v tomto případě nelze použít).

   Manuální výpočet (viz vzorec ze skript):

```
(Si2=tapply(lnCu,Misto,var))
(wi=ni/Si2)
(Lambda=sum(1/(ni-1)*(1-wi/sum(wi))^2)/(p^2-1))
(Ywbar=weighted.mean(prumery,w=wi))
(FW=sum(wi*(prumery-Ywbar)^2)/(p-1)*1/(1+2*Lambda*(p-2)))
oneway.test(lnCu~Misto)$stat

#df
oneway.test(lnCu~Misto)$param
1/(3*Lambda)

#p-hodnota
1-pf(FW,p-1,1/(3*Lambda))
oneway.test(lnCu~Misto)$p.val
```

9. Podíváme se, jaký je vztah dvouvýběrového t-testu a analýzy rozptylu pro případ $p = 2$. Z našich dat si tedy vybereme pouze místa A a B a ta porovnáme jak t-testem, tak pomocí F-testu.

```
detach(Med)
AB=Med[Med$Misto=="A"|Med$Misto=="B",]
AB$Misto=factor(AB$Misto)

(t=t.test(lnCu~Misto,data=AB,var.equal=T))
(a=anova(modelAB<-aov(lnCu~Misto,data=AB)))

#porovnani testovych statistik a p-hodnot
t$stat^2
a$F
t$p.val
a$Pr
```

Pomocí jakých rozdělení jsou spočtené výše uvedené p-hodnoty?

Podobné srovnání pak dostaneme pro Welchův test a Welchovu statistiku jednoduchého třídění

```
t.test(lnCu~Misto,data=AB)
oneway.test(lnCu~Misto,data=AB)
```

Samostatná práce

1. Data dieta.txt porovnávají efekt tří diet na hubnutí. Pro 76 osob máme k dispozici jejich pohlaví, věk, výšku, typ diety (kódováno 1, 2 a 3) a hmotnost před a po 6 týdnech diety. Zajímá nás, zda mají zkoumané tři diety stejný vliv na úbytek hmotnosti, nebo zda je mezi nimi významná odlišnost.

   (a) Načtení dat

```
dieta=read.table("dieta.txt",header=T)
summary(dieta)
dim(dieta)
head(dieta)

dieta$ubytek=dieta$pre.weight-dieta$weight6weeks
dieta$Diet=factor(dieta$Diet)

attach(dieta)
table(Diet)
```

(b) Zkoumaný problém si vizualizujte pomocí boxplotů. Dále si spočtěte průměry a směrodatné odchylky jednotlivých kategorií (použijte příkaz tapply). Podívejte se také na QQ grafy úbytků hmotnosti pro jednotlivé diety a posuďte splnění předpokladů analýzy rozptylu.

(c) Pomocí analýzy rozptylu otestuje shodu středních hodnot úbytků hmotnosti pro zkoumané tři diety.

(d) Pomocí mnohonásobného porovnání (Bonferroniho nebo Tukeyho metodou) zjistěte, které diety se od sebe významně liší.

2. Podívejte se na hustoty $F_{m,n}$ rozdělení pro různé n a m.

```
curve(df(x,4,30),lwd=2,0,4,ylab="f",xlab="x",ylim=c(0,1))
curve(df(x,4,100),lwd=2,add=T,col="red")
curve(df(x,4,1000),lwd=2,add=T,col="brown")
curve(df(x,8,30),lwd=2,add=T,col="darkgreen")
curve(df(x,8,100),lwd=2,add=T,col="purple")
curve(df(x,8,1000),lwd=2,add=T,col="blue")
legend("topright",c("F(4,30)","F(4,100)","F(4,1000)","F(8,30)","F(8,100)","F(8,1000"),
   col=c("black","red","brown","darkgreen","purple","blue") ,lty=rep(1,6),lwd=rep(2,6))
```

(a) Ověřte (přibližně graficky), zda je pravda, že hustota $F_{n,m}$ rozdělení dosahuje maximum v bodě $\frac{n-2}{n}\frac{m}{m+2}$.

(b) Jak se chová $F_{m,n}$ rozdělení pro m pevné a $n \to \infty$? Porovnejte grafy hustot $F_{m,n}$ pro nějaké pevně zvolené m (malé) a n (velké) s tímto limitním rozdělením.

(c) Podobně si vykreslete graf distribuční funkce (namísto df použijte pf) a sledujte, jak se chová $F_{n,m}$ rozdělení pro $n, m \to \infty$. Zdůvodněte vše i teoreticky.

3. Závěrečné opakování: Rozhodněte, jaký test (postup) byste použili pro zkoumání následujících problému vztahujících se k datům dieta.txt:

   – Měla dieta 1 efekt na hubnutí? Tj. mají osoby po jejím absolvování nižší hmotnost než před tím?
   – Je pravdivé tvrzení, že díky dietě 3 lidé zhubnou v průměru více než 5 kg?
   – Závisí úbytek hmotnosti na pohlaví?
   – Je pravděpodobnost zhubnutí stejná pro muže a pro ženy?
   – Jsou věkové skupiny < 30 let, $30 - 50$ let, > 50 zastoupeny v populaci hubnoucích lidí v poměru $1 : 2 : 1$?
   – Je pravděpodobnost zhubnutí stejná pro výše uvedené tři věkové skupiny?