Kontingenční tabulky, testy dobré shody a analýza rozptylu (ANOVA)

6.1.2020 (II)

Testy dobré shody se známými parametry. Mezi 891 studenty pražských vysokých škol byl na podzim 2017 proveden průzkum týkající se jejich konzumace alkoholu. Studenti byli tázáni, jak často pijí alkohol. Výsledky průzkumu jsou uvedeny v následující tabulce 1.

	$>4\times$týdně	$3-4\times$ týdně	$1-2\times$ týdně	$1-3\times$ do měsíce	$<1\times$ za měsíc	vůbec
počet	35	143	293	305	87	28

Tabulka 1: Frekvence konzumace alkoholu mezi studenty pražských vysokých škol.

SZÚ ve své zprávě z roku 2014 uvádí konzumaci alkoholu v ČR v tabulce 2. Zajímá nás, zda je konzumace alkoholu mezi studenty stejná jako v celé ČR.

	$>4\times$týdně	$3-4\times$ týdně	$1-2\times$ týdně	$1-3\times$ do měsíce	$<1\times$ za měsíc	vůbec
%	5	7	23	33	29	3

Tabulka 2: Konzumace alkoholu v ČR v % podle průzkumu SZÚ z roku 2014, zdroj SZÚ.

1. Zadejte si data do R.

```
alkohol=c(35, 143, 293, 305,  87,  28)

p0=c(5,7,23,33,29,3)/100
```

   Znázorněte si data také graficky.

2. Nyní již budeme chtít testovat výše uvedenou domněnku.
   (a) Jaký model budeme uvažovat pro naše data? Jak budeme formulovat hypotézu?
   (b) Odhadněte parametry našeho modelu a porovnejte je graficky:
   `(p.hat=alkohol/sum(alkohol))`

   ```
   barplot(rbind(p.hat,p0),beside=T,legend=c("studenti","CR"))
   ```
   (c) Provedeme test dobré shody
   `chisq.test(alkohol,p=p0,correct=F)`
   Jaký je náš závěr?
   (d) Připomeňte si,
   - zda se jedná o přesný nebo asymptotický test a z jakého rozdělení je spočtena p-hodnota,
   - co by mělo být splněno, aby bylo použití asymptotického testu rozumné,
   - jak byste spočítali hodnotu testové statistiky ručně.

   (e) Kdybychom chtěli vědet, v kterých kategoriích se studenti od ČR nejvíce liší, můžeme se podívat na tzv. rezidua

```
chisq.test(alkohol,p=p0,correct=F)$residuals
```

3. Připomeňte si, jak byste testovali, zda je procentuální zastoupení studentů, kteří pijí často (více než čtyřikrát týdně) stejné jako těch, kteří nepijí alkohol vůbec.

Kontingenční tabulky. Zajímá nás, zda se konzumace alkoholu mezi studenty liší u žen a mužů, neboli zda jsou pohlaví a konzumace alkoholu (výše definovaná) závislé kategoriální znaky. K dispozici máme data uvedená v tabulce 3

	$>4\times$týdně	$3-4\times$ týdně	$1-2\times$ týdně	$1-3\times$ do měsíce	$<1\times$ za měsíc	vůbec
ženy	15	65	155	191	55	20
muži	20	78	138	114	32	8

Tabulka 3: Konzumace alkoholu mezi studenty pražských vysokých škol podle pohlaví.

4. Zadáme si Tabulku 3 do R:

```
tab=matrix(c(15,20,65,78,155,138,191,114,55,32,20,8),nrow=2)
dimnames(tab)=list(Pohlavi=c("zena","muz"),
      Alkohol=c(">4 T","2-3 T","1-2 T","1-3 M","<1 M","nikdy"))
```

Zkontrolujte, že máme tabulku dobře zadanou. Kdybychom si chtěli dopočítat marginální četnosti, provedeme to následovně:

```
rowSums(tab)
colSums(tab)

# nebo ekvivalentne pomoci funkce apply:

apply(tab,1,sum)
apply(tab,2,sum)
```

5. Jaký model teď budeme uvažovat pro naše data? Co všechno je v modelu náhodné a co naopak není?

6. Podíváme se na tabulky relativních četností

```
prop.table(tab)
prop.table(tab,marg=1)
prop.table(tab,marg=2)
```

Co nám jednotlivé relativní četnosti odhadují? Jak by měly tabulky přibližně vypadat v případě nezávislosti?
Podíváme se na tutéž věc i graficky:

```
barplot(tab,beside=T,legend=T)
barplot(prop.table(tab,mar=2),beside=T,legend=T)
barplot(t(tab),beside=T,legend=T,col=rainbow(6))
barplot(prop.table(t(tab),mar=2),beside=T,legend=T,col=rainbow(6))
```

Prozkoumejte jednotlivé obrázky a jak se mezi sebou liší. Co si na základě čísel a grafů myslíte o vztahu zkoumaných dvou veličin? Jsou nezávislé?

7. Provedeme χ^2 test nezávislosti.

```
chisq.test(tab,correct=FALSE)
```

   (a) Jaký je náš závěr?
   (b) Připomeneme, jak se spočítá testová statistika výše uvedeného testu a jaké je její asymptotické rozdělení. Jak spočítáme stupně volnosti?
   (c) Které kategorie tabulky nejvíce přispívají k výsledné hodnotě χ^2 statistiky a tím „porušují“ nezávislost?

Analýza rozptylu (ANOVA). 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.

8. Stáhněte, načtěte a prohlédněte si data Med.txt. V analýze budeme pracovat s logaritmem koncentrace, tj. s proměnnou lnCu.
   – 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")
```

9. Na náš problém budeme chtít použít analýzu rozptylu. Připomeňte si, jaké všechny předpoklady tato metoda má. Formulujte H_0 a H_1.

10. 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)
```

11. Otestujte, zda je znečištění řeky na zkoumaných pěti místech stejné. Test provedeme následovně:

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

Jaký je závěr?

12. Manuální výpočet jednotlivých položek z tabulky analýzy rozptylu:

```
(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)
```

13. Proč jsme nemohli provést test tak, že bychom porovnali (na hladině 5 % pomocí t-testu nebo Welchova testu) všechny dvojice míst a zamítli bychom H_0, pokud alespoň jeden z testů odhalí rozdíl?
Jak lze modifikovat výše uvedený postup, abychom mohli provést mnohonásobné porovnání jednotlivých míst na celkové hladině 5 %?

```
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]],var.equal=T)$p.val)
}
```

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

```
tabulka=matrix(0,ncol=6,nrow=m)
  rownames(tabulka)=1:10
  colnames(tabulka)=c("diff", "lwr","upr", "p"," p adj","reject")
  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$estimate[2]
    tabulka[k,2]=int[1]
    tabulka[k,3]=int[2]
    tabulka[k,4]=test$p.val
    tabulka[k,5]=min(test$p.val*m,1)
    tabulka[k,6]=ifelse(tabulka[k,4]<alpha/m,1,0)
    k=k+1
  }
 tabulka
```

14. 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?