ZihaoHuang-midterm 1

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#stat223 midterm
#Zihao Huang
#1
cork <- read.table("C:\\Users\\simon\\Documents\\stat223\\cork.txt",header=T)
##a
##violate Independence, four borings are taken at the same tree for each sample.
##b
##1 -1 0 0
##1 0 -1 0
##0 1 0 -1
##c

c1<-matrix(c(1,1,0,-1,0,1,0,-1,0,0,0,-1),nrow=3)
y<-(as.matrix(cork[,2:5]))
(matvar<-var(y))
```

```
##          N        E        S        W
## N 290.4061 223.7526 288.4378 226.2712
## E 223.7526 219.9299 229.0595 171.3743
## S 288.4378 229.0595 350.0040 259.5410
## W 226.2712 171.3743 259.5410 226.0040
```

```
n<-nrow(y)
p<-ncol(y)
meany<- apply(y,2,mean)
##c1%*%meany
T2<-nrow(y)*t(c1%*%meany)%*%solve(c1%*%matvar%*%t(c1))%*%(c1%*%meany)
T2
```

```
##          [,1]
## [1,] 20.74202
```

```
#(critval <- p*(n-1)/(n-p)*qf(.95,p,n-p))
#(pvalT2 <- 1-pf((n-p)/((n-1)*p)*T2,p,n-p))
(critval <- (4-1)*(n-1)/(n-4+1)*qf(.95,4-1,n-4+1))
```

```
## [1] 9.691621
```

```
(pvalT2 <- 1-pf((n-4+1)/((4-1)*(n-1))*T2,3,25))
```

```
##             [,1]
## [1,] 0.002280399
```

```
##given T2= 20.74202>9.6916 and p-value 0.0023<0.05, we reject the null hypothesis and accept that Ha.
##d
varc1mu<-c1%*%matvar%*%t(c1)
```

```
c1mu<-c1%*%meany
t1<-c()
for (i in 1:3){
     t1[i]<-c1mu[i]/sqrt(varc1mu[i,i]/n)
   }
t1
```

```
## [1] 2.9086731 0.5690203 0.5209192
```

```
(p.value =2*pt(-abs(t1), n-1))
```

```
## [1] 0.007178626 0.574045784 0.606669393
```

```
##t-value are 2.9086731 0.5690203 0.5209192
##p-value are 0.007178626 0.574045784 0.606669393
##contrast 1 reject H0.

#2
sparrow<-read.table("C:\\Users\\simon\\Documents\\stat223\\sparrow.txt",header=T)
##a
##sample sizes are different
##b
y1<-sparrow[1:21,2:6]
y2<-sparrow[22:49,2:6]
s1<-var(y1)
s2<-var(y2)
n1<-nrow(y1)
n2<-nrow(y2)
meany1<- apply(y1,2,mean)
meany2<- apply(y2,2,mean)
meandiff<-(meany1-meany2)
n1<-21
n2<-28
p<-5
sp<-((n1-1)*s1+(n2-1)*s2)/(n1+n2-2)
(T2<-(n1*n2/(n1+n2))*t(meandiff)%*%solve(sp)%*%(meandiff))
```

```
##          [,1]
## [1,] 2.823698
```

```
## T2 is 2.823698
p*(n1+n2-2)/(n1+n2-p-1)*qf(.95,p, n1+n2-p-1)
```

```
## [1] 13.29246
```

```
##critical value 13.29246
a1 <- 1/(p*(n1+n2-2)/(n1+n2-p-1))
1-pf(a1*T2, p, n1+n2-p-1)
```

```
##           [,1]
## [1,] 0.7621709
```

```
##p-value is 0.7621709
t.2<-c()
for (i in 1:ncol(y1)){
  t.2[i]<-meandiff[i]/sqrt(sp[i,i]*(1/n1+1/n2))
}
```

```
##t statistics are -0.9929537 -0.3871246 -0.1951942  0.3257939 -0.1029179
(p.value =2*pt(-abs(t.2), n1+n2-2))
```

```
## [1] 0.3258173 0.7004114 0.8460823 0.7460264 0.9184660
```

```
#p-values are 0.3258173 0.7004114 0.8460823 0.7460264 0.9184660
(a<-t(solve(sp)%*%meandiff))
```

```
##          totlen     alarext    beaklen   humlen   sternlen
## [1,] -0.1553257 -0.02649058 -0.0928576 1.032474 0.06932512
```

```
#a          totlen     alarext     beaklen   humlen   sternlen
#         -0.1553257 -0.02649058 -0.0928576 1.032474 0.0693251
#humlen contributes most


#3
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.4.3
```

```
usstates <- read_excel("~/stat223/USStates.xlsx")
colnames(usstates)
```

```
##  [1] "State"            "HouseholdIncome"  "IQ"
##  [4] "McCainVote"       "Region"           "ObamaMcCain"
##  [7] "Population"       "EighthGradeMath"  "HighSchool"
## [10] "GSP"              "FiveVegetables"   "Smokers"
## [13] "PhysicalActivity" "Obese"            "College"
## [16] "NonWhite"         "HeavyDrinkers"    "ElectoralVotes"
```

```
us<-usstates[,12:16]
ne<- subset(usstates, Region=="NE", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))
w<- subset(usstates, Region=="W", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))
s<- subset(usstates, Region=="S", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))
mw<- subset(usstates, Region=="MW", select=c(Smokers,PhysicalActivity,Obese,College,NonWhite))

n1 <- nrow(ne)
mean1 <- colMeans(ne)
S1 <- var(ne)
n2 <- nrow(w)
mean2 <- colMeans(w)
S2 <- var(w)
n3 <- nrow(s)
mean3 <- colMeans(s)
S3 <- var(s)
n4 <- nrow(mw)
mean4 <- colMeans(mw)
S4 <- var(mw)

p<-5
N <- nrow(usstates)
k<-4
n <- c(n1,n2,n3,n4)
Spl <- ((n1-1)*S1+(n2-1)*S2+(n3-1)*S3+(n4-1)*S4)/(N-k)

m1 <- manova(cbind(Smokers,PhysicalActivity,Obese,College,NonWhite)~Region,data=usstates)
```

```r
summary(m1,test="Wilks")
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## Region     3 0.12565   8.6865     15 116.34 4.3e-13 ***
## Residuals 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
H <- summary(m1,test="Wilks")$SS$Region
E <- summary(m1,test="Wilks")$SS$Residuals
(wilks <- det(E)/det(E+H))
```

```
## [1] 0.1256536
```

```r
##wilks is 0.12565
##wilks lambda table shows p=5, k-1=3, N-k=46, the value is about 0.522, lambda<0.522
(f.stat <- (N-k-p+1)/p*(1-sqrt(wilks))/sqrt(wilks))
```

```
## [1] 15.29692
```

```r
df.num <- 2*p
df.den <- 2*(N-k-p+1)
qf(0.95,df.num,df.den)
```

```
## [1] 1.945361
```

```r
pf(f.stat, df1=10, df2=84, lower.tail=FALSE)
```

```
## [1] 3.609482e-15
```

```r
##F statistics is 15.29692>1.94536, p-value is 3.609482e-15<0.05,

summary(aov(Smokers~Region,data=usstates))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Region       3  144.1   48.03   6.412 0.00101 **
## Residuals   46  344.6    7.49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##F is 6.412, p-value is 0.00101,
summary(aov(PhysicalActivity~Region,data=usstates))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Region       3  295.0   98.33   11.02 1.42e-05 ***
## Residuals   46  410.4    8.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##F is 11.02, p-value is 1.42e-05
summary(aov(Obese~Region,data=usstates))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Region       3  262.5   87.51   21.01 1.02e-08 ***
## Residuals   46  191.6    4.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##F is 21.01, p-value is 1.02e-08
summary(aov(College~Region,data=usstates))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Region       3  656.9  218.95   13.73 1.59e-06 ***
## Residuals   46  733.6   15.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##F is 13.73, p-value is 1.59e-06
summary(aov(NonWhite~Region,data=usstates))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Region       3   2224   741.5   4.281 0.00954 **
## Residuals   46   7967   173.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##F is 4.281, p-value is 0.00954

##e
chi<--(N-k-1-0.5*p+0.5*k)*log(wilks)

#4.

R <- cor(us)
(u <- det(R))
```

```
## [1] 0.08519845
```

```
(u1 <- -(N-1-1/6*(2*p+5))*log(u))
```

```
## [1] 114.5189
```

```
df <- .5*(p^2-p)
(critval <- qchisq(.95,df))
```

```
## [1] 18.30704
```

```
u1 > critval
```

```
## [1] TRUE
```

```
pchisq(u1,df=10,lower.tail=FALSE)
```

```
## [1] 6.524936e-20
```

```
## u1 is 114.5189>18.30704, p-value is 6.524936e-20
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.