PH245 HW1 Solution

Xiangyu Hu and Toki Sherbakov

October 10, 2014

Problem 1.
Test: Repeated Measures:

Reasoning: There is **one single variable: median reaction time**, and it was measured four times repeatedly on the **same subject**, assuming that there is no left-over effect of treatments.

Hypotheses:
$H_0 : \mathbf{C}\boldsymbol{\mu} = 0$, or equivalently $\mu_1 = \mu_2 = \mu_3 = \mu_4$, or in words: there is no difference in mean median reaction time across 4 treatments
H_1: There is difference in mean median reaction time across 4 treatments

```
# set work directory
setwd("/Users/huxiangyu/Documents/ZhengShi/Berkeley/GSI/PH245/HW/hmwk1")
# Import dataset
cognition = read.table("Data-06-17.dat")

# Obtain value for n
n = nrow(cognition)

# Obtain X_bar matrix
dat.mean = matrix(colMeans(cognition), nrow = ncol(cognition))

# Obtain covariance matrix
dat.cov = cov(cognition)

# Obtain C matrix
C = matrix(c(-1, 1, 0, 0, 0, -1, 1, 0, 0, 0, 1, -1), nrow = 3, ncol = 4, byrow = TRUE)

# Get hotellings T^2 statistic
T2 = n * t(C %*% dat.mean) %*% solve(C %*% dat.cov %*% t(C)) %*% (C %*% dat.mean)
T2
```

```
##       [,1]
## [1,] 153.7
```

```
# Obtain value for q-tilda
q.tilda = nrow(C)

# Set confidence level
alpha = 0.05

# Obtain cutoff for test statistic
f.cutoff = ((n - 1)*q.tilda)/(n - q.tilda) * qf(1 - alpha, q.tilda, n - q.tilda)
f.cutoff
```

```
## [1] 9.409
```

```
# Test to see if T2 is larger than the F distribution cutoff
# If so, then we reject the null that the four measures of reaction times are the same
T2 > f.cutoff
```

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

```
# Obtain p-value if needed (need to convert it back to standard F distribution)
pf(T2 / (((n - 1)*q.tilda)/(n - q.tilda)), q.tilda, n - q.tilda, lower.tail = F)
```

```
##           [,1]
## [1,] 2.328e-11
```

Conclusions: The test statistic we obtained is 153.7, and critical value obtained by setting $\alpha = 0.05$ is 9.409. The test statistic is much bigger than the critical value, so we can reject the null hypothesis and conclude that mean median reaction time is different across 4 treatments.
Alternatively, we can use test statistic to calculate p-value which is 2.328e-11. It is much smaller than $\alpha = 0.05$, so we can reject the null hypothesis.

Problem 2. *Graded, total 10 points*
Test:(*1pt*) Multivariate two sample comparison

Reasoning:(*3pt*) The two populations we are comparing are independent. Although having different sample sizes is not a necessary requirement(which means that sometimes sample sizes could be the same), these two samples have different sample sizes and by no mean they can be paired up. Three different types of costs were measured, so this is a multivariate problem.

Hypotheses:(*2pt*)
H_0 : $\boldsymbol{\mu_{gasoline}} = \boldsymbol{\mu_{diesel}}$, where $\boldsymbol{\mu_{gasoline}}$ and $\boldsymbol{\mu_{diesel}}$ are mean vectors that looks like the following:

$$\boldsymbol{\mu_{gasoline}} = \begin{pmatrix} \mu_{fuel:gasoline} \\ \mu_{repair:gasoline} \\ \mu_{capital:gasoline} \end{pmatrix} \boldsymbol{\mu_{diesel}} = \begin{pmatrix} \mu_{fuel:diesel} \\ \mu_{repair:diesel} \\ \mu_{capital:diesel} \end{pmatrix}$$

, or in words: there is no difference in the mean costs between the gasoline and diesel trucks.
H_1: There is difference in the mean costs between the gasoline and diesel trucks.

```
# Import dataset
milk = read.table("Data-06-19.dat")

# Obtain the total number of observations
n = nrow(milk)

# Separate the datasets into gasoline and diesel trucks
gas = milk[milk$V4 == "gasoline", 1:3]
diesel = milk[milk$V4 == "diesel", 1:3]

# Find column means (x.bar) for gas and diesel trucks
gas.bar = colMeans(gas)
diesel.bar = colMeans(diesel)

# Get covariance matrices
s.gas = cov(gas)
s.diesel = cov(diesel)

# Get values for n1 and n2, the number of observations for the two samples
n.gas = nrow(gas)
n.diesel = nrow(diesel)

# Pooled covariance estimate
s.p = ((n.gas - 1) * s.gas + (n.diesel - 1) * s.diesel) / (n.gas + n.diesel - 2)
```

```r
# Obtain T2 statistic
T2 = (gas.bar - diesel.bar) %*% solve((1/n.gas + 1/n.diesel) * s.p) %*%
      as.matrix(gas.bar - diesel.bar)
T2
```

```
##       [,1]
## [1,] 50.91
```

```r
# Set p to be 3 (3 different variables that we are looking at) and set significance level
p = 3
alpha = 0.05

# Get cutoff value
f.cutoff = ((n.gas + n.diesel - 2) * p) / (n.gas + n.diesel - p - 1) *
  qf(1 - alpha, p, n.gas + n.diesel - p - 1)
f.cutoff
```

```
## [1] 8.62
```

```r
# Test to see if T2 is larger than the F distribution cutoff
# If so, then we reject the null that there is no cost difference between gas and diesel trucks
T2 > f.cutoff
```

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

```r
# Obtain p-value if needed (need to convert it back to standard F distribution)
pf(T2 / (((n.gas + n.diesel - 2) * p) / (n.gas + n.diesel - p - 1)),
   p, n.gas + n.diesel - p - 1, lower.tail = F)
```

```
##       [,1]
## [1,] 1e-07
```

Conclusions: The test statistic we obtained is 50.91 (*1pt*), and critical value obtained by setting $\alpha = 0.05$ is 8.62 (*1pt* for either test statistics or p-value). The test statistic is much bigger than the critical value, so we can reject the null hypothesis and conclude that mean costs are different between gasoline and diesel trucks. (*2pt*)
Alternatively, we can use test statistic to calculate p-value which is 1e-07. It is much smaller than $\alpha = 0.05$, so we can reject the null hypothesis.

Problem 3. *Graded, total 10 points*
Test:(*1pt*) One-way MANOVA

Reasoning:(*4pt*) It is one-way because we have only one factor, which is time period. Based on time period, we have three populations, which is more than two population. In this scenario, we should check if we would use ANOVA or MANOVA based on the number of dependent variables. In this problem, there are 4 dependent variables:maximum breadth of skull, base height of skull, base length of skull, nasal height of skull. Therefore, it is a multivariate case, and we should use MANOVA.

Hypotheses:(*2pt*)
$H_0 : \boldsymbol{\mu_{4000B.C.}} = \boldsymbol{\mu_{3300B.C.}} = \boldsymbol{\mu_{1850B.C.}}$, where $\boldsymbol{\mu_{4000B.C.}}$, $\boldsymbol{\mu_{3300B.C.}}$, $\boldsymbol{\mu_{1850B.C.}}$ are mean vectors that looks like the following:

$$\boldsymbol{\mu_{4000B.C.}} = \begin{pmatrix} \mu_{\text{maximum breadth of skull:4000 B.C.}} \\ \mu_{\text{base height of skull:4000 B.C.}} \\ \mu_{\text{base length of skull:4000 B.C.}} \\ \mu_{\text{nasal height of skull:4000 B.C.}} \end{pmatrix}$$

$$\boldsymbol{\mu_{3300B.C.}} = \begin{pmatrix} \mu_{\text{maximum breadth of skull:3300 B.C.}} \\ \mu_{\text{base height of skull:3300 B.C.}} \\ \mu_{\text{base length of skull:3300 B.C.}} \\ \mu_{\text{nasal height of skull:3300 B.C.}} \end{pmatrix}$$

$$\boldsymbol{\mu_{1850B.C.}} = \begin{pmatrix} \mu_{\text{maximum breadth of skull:1850 B.C.}} \\ \mu_{\text{base height of skull:1850 B.C.}} \\ \mu_{\text{base length of skull:1850 B.C.}} \\ \mu_{\text{nasal height of skull:1850 B.C.}} \end{pmatrix}$$

, or in words: there is no difference in the mean skull sizes across 3 periods.(or equivalently, there is no time effect)
H_1: There is difference in the mean skull sizes across 3 periods.(or equivalently, there is time effect)

```r
# We are testing to see if the three time periods differ based on their skull measurements.
# This requires a one-way MANOVA since there are multiple responses (4) that we are comparing
# across 3 groups (more than 2 groups).

# Import dataset
skull = read.table("Data-06-24.dat")

# Change the last column to a factor
skull$V5 = as.factor(skull$V5)

# Change column names of skull to be interpretable
colnames(skull) = c("MaxBreadth", "BaseHeight", "BaseLength", "NasalHeight", "TimePeriod")

# Fit MANOVA
fit = manova(cbind(MaxBreadth, BaseHeight, BaseLength, NasalHeight) ~ TimePeriod, data = skull)

# Obtain the summary of the fit using Wilks test
summary(fit, test = "Wilks")
```

```
##            Df Wilks approx F num Df den Df Pr(>F)
## TimePeriod  2  0.83     2.05      8    168  0.044 *
## Residuals  87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## ADDITIONAL ANALYSIS - Not required for grading the homework ##

# Find anova for each skull measurement to see which one is marginally significant
fit2 = aov(cbind(MaxBreadth, BaseHeight, BaseLength, NasalHeight) ~ TimePeriod, data = skull)

# Obtain the summary of the piecewise ANOVA
summary(fit2)
```

```
##  Response MaxBreadth :
##             Df Sum Sq Mean Sq F value Pr(>F)
## TimePeriod   2    150    75.1    3.66   0.03 *
## Residuals   87   1785    20.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response BaseHeight :
##             Df Sum Sq Mean Sq F value Pr(>F)
## TimePeriod   2     21    10.3    0.47   0.63
## Residuals   87   1924    22.1
##
##  Response BaseLength :
##             Df Sum Sq Mean Sq F value Pr(>F)
## TimePeriod   2    190    95.1    3.84  0.025 *
```

```
## Residuals   87   2153    24.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response NasalHeight :
##             Df Sum Sq Mean Sq F value Pr(>F)
## TimePeriod   2      2    1.01     0.1    0.9
## Residuals   87    840    9.66

# Based on these results, it seems that any differences over time periods are probably due to
# changes in V1 (maximum breath of skull) and V3 (basial vaseolar length). This is purely
# exploratory.
```

Conclusions: The p-value is 0.044 (*1pt*), which is less than $\alpha = 0.05$, so we reject the null hypothesis and conclude that there is a time period effect and that mean skull sizes for these three periods are not all equal to each other.(*2pt*)

Problem 4.
Test: Two-way MANOVA

Reasoning: It is two-way because we have two factors, which are species and time. Based on time, we have three populations (more than two population). Based on species, we have three populations (more than two population). In this scenario, we should check if we would use ANOVA or MANOVA based on the number of dependent variables. In this problem, there are 2 dependent variables:percent spectral reflectance at wavelength 560 nm, percent spectral reflectance at wavelength no nm. Therefore, it is a multivariate case, and we should use MANOVA.

Hypotheses: Here, we have three sets of hypotheses for each effect
For species:
H_0: There is no species effect (or equivalently, there is no difference in the mean percent spectral reflectance at wavelengths across three species)
H_1: There is species effect. (or equivalently, there is a difference in the mean percent spectral reflectance at wavelengths across three species)

For time:
H_0: There is no time effect (or equivalently, there is no difference in the mean percent spectral reflectance at wavelengths across three times)
H_1: There is time effect.(or equivalently, there is a difference in the mean percent spectral reflectance at wavelengths across three times)

For interaction of species and time:
H_0: There is no interaction effect of time and species (or equivalently, there is no difference in the mean percent spectral reflectance at wavelengths across any combinations of species and times)
H_1: There is interaction effect of time and species.(or equivalently, there is a difference in the mean percent spectral reflectance at wavelengths across any combinations of species and times)

```
# Import data
remote = read.table("Data-06-33.dat")

# Don't need the last column of the data (it's the replication variable)
remote = remote[,-5]

# Change the 4th column to a factor
remote$V4 = as.factor(remote$V4)

# Change the column names of remote so that it's interpretable
colnames(remote) = c("Wavelength560", "Wavelength0", "Species", "Time")
```

```r
# Fit MANOVA
fit = manova(cbind(Wavelength560, Wavelength0) ~ Species*Time, data = remote)

# Obtain summary of the fit
summary(fit, test = "Wilks")

##              Df  Wilks approx F num Df den Df  Pr(>F)
## Species       2 0.0688     36.6      4     52 1.6e-14 ***
## Time          2 0.0492     45.6      4     52 < 2e-16 ***
## Species:Time  4 0.0871     15.5      8     52 2.2e-11 ***
## Residuals    27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions: The p-values are 1.6e-14, $< 2e-16$, 2.2e-11 correspondingly for each test. Therefore we reject those null hypotheses, and conclude that there are species effect, time effect, and interaction effect of time and species.