hw1 - Jin Kweon (3032235207)

Jin Kweon

9/20/2017

Problem 1

Test I choose:

Repeated measure design.

Reason to choose the test:

There are 32 subjects in the data. The problem states that half of the subjects (16 of the entire set) got assigned Arabic first and the other got assigned words first. So, in here, to test the treatment effects, I need to find subjects, quantitative variables, factors, and levels.

For this question, subjects are the same (n = 32).

And, there are 1 quantitative variable to measure: median reaction times.

And, the groups are defined by 2 factors: format and parity.

And, we have 2 levels for each factor: "Arabic digits or words for format" and "same and different for parity." Each subject receives each treatment once over successive periods of time, assuming no left-over effects. (So, four different treatments: "Arabic with same parity," "Arabic with different parity," "Words with same parity," and "Words with different parity" are compared with respect to a single variable)

Thus, **the same population with four independent treatments with respect to a single variable** implies that the test will be a "repeated measure design."

Test:

As we learned in the class, the hypothesis can be formulated as $H_o : C\mu \ = \ 0$.

Let the table has mean each:

$$Mean\ Table: \begin{bmatrix} Word\ Diff = & \mu 1 \\ Word\ Same = & \mu 2 \\ Arabic\ Diff = & \mu 3 \\ Arabic\ Same = & \mu 4 \end{bmatrix}$$

I have four hypotheses, below:

1. Null effect: $\mu1 = \mu2 = \mu3 = \mu4$

$$\begin{pmatrix} \mu2 - \mu1 \\ \mu3 - \mu2 \\ \mu4 - \mu3 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu1 \\ \mu2 \\ \mu3 \\ \mu4 \end{pmatrix} = C\mu$$

2. Main effect of Arabic or words: $(\mu1 + \mu2) - (\mu3 + \mu4)$

$$\begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu1 \\ \mu2 \\ \mu3 \\ \mu4 \end{pmatrix} = C\mu$$

3. Main effects of parity: $(\mu2 + \mu4) - (\mu1 + \mu3)$

$$\begin{pmatrix} -1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu1 \\ \mu2 \\ \mu3 \\ \mu4 \end{pmatrix} = C\mu$$

4. Interaction of parity and format: $(\mu2 + \mu3) - (\mu1 + \mu4)$

$$\begin{pmatrix} -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu1 \\ \mu2 \\ \mu3 \\ \mu4 \end{pmatrix} = C\mu$$

My test statistics is $T^2 = n(C\bar{X})^T(CSC^T)^{-1}C\bar{X}$, and distribution under H_0 is $T^2 \sim \frac{(n-1)\tilde{q}}{n-\tilde{q}} F_{\tilde{q},\, n-\tilde{q}}$, where $C\bar{X} \sim N_{\tilde{q}}(C\mu,\, n^{-1}CSC^T)$. I want to test all four hypotheses.

$\tilde{q}$ will be 3, 1, 1, 1, and $n - \tilde{q}$ will be 29 (= 32 - 3), 31 (= 32 - 1), 31, and 31.

Conclusion:

This is really interesting test. They test how people prcoess numbers in the brain depends on whether they are expressed in words or Arabic digits.

The means for each column are not all the same.

And, the effects of Arabic digits and words are not the same.

And, the effects of parity are not the same.

However, when we make an interaction between parity and format, their means are quite the same.

```r
setwd("/Users/yjkweon24/Desktop/Cal/2017 Fall/PB HLTH 245/HW/HW1")
cog_dat <- read.table("Data-HW1-Cognition.dat")
colnames(cog_dat) <- c("x1", "x2", "x3", "x4")

dim(cog_dat)
```

```
## [1] 32  4
```

```
n <- nrow(cog_dat)
xbar <- apply(cog_dat, 2, mean)
S <- var(cog_dat)

#hypothesis 1
c1 <- matrix(c(-1, 0, 0, 1, -1, 0, 0, 1, -1, 0, 0, 1), 3, 4)

c1xbar <- c1 %*% xbar
c1stc1 <- c1 %*% S %*% t(c1)
t1 <- as.numeric(n * t(c1xbar) %*% solve(c1stc1) %*% c1xbar)
pf((t1*29)/(31*3), 3, 29, lower.tail = F) #same as 1 - pf((t1*29)/(31*3), 3, 29)
```

```
## [1] 2.32844e-11
```

```
#Reject the null!!!




#hypothesis 2
c2 <- matrix(c(1, 1, -1, -1), 1, 4)

c2xbar <- c2 %*% xbar
c2stc2 <- c2 %*% S %*% t(c2)
t2 <- as.numeric(n * t(c2xbar) %*% solve(c2stc2) %*% c2xbar)
pf((t2*31)/(31*1), 1, 31, lower.tail = F)
```

```
## [1] 9.02949e-10
```

```
#Reject the null




#hypothesis 3
c3 <- matrix(c(-1, 1, -1, 1), 1, 4)
c3xbar <- c3 %*% xbar
c3stc3 <- c3 %*% S %*% t(c3)
t3 <- as.numeric(n * t(c3xbar) %*% solve(c3stc3) %*% c3xbar)
pf((t3*31)/(31*1), 1, 31, lower.tail = F)
```

```
## [1] 2.013627e-09
```

```
#Reject the null




#hypothesis 4
c4 <- matrix(c(-1, 1, 1, -1), 1, 4)
c4xbar <- c4 %*% xbar
c4stc4 <- c4 %*% S %*% t(c4)
t4 <- as.numeric(n * t(c4xbar) %*% solve(c4stc4) %*% c4xbar)
pf((t4*31)/(31*1), 1, 31, lower.tail = F)
```

```
## [1] 0.2142716
#Do not reject the null
```

Problem 2

Test I choose:

Two sample T^2 test

Reason to choose the test:

So, this time, the number of subjects for gasoline and diesel are different. (36 gasoline and 23 diesel trucks) Subject is 59 trucks for this question.

I can say 3 quantitative variables/measurements: costs of fuel, repair, and capital.

There is 1 factor: types of trucks.

And, I have 2 levels: gasoline and diesel. And, I want to test for differences in the mean costs between the gasoline and diesel trucks.

So, **one factor for two populations with different subjects** implies that I need to use "Two sample T^2 test." Here, we assume gasoline and diesel trucks are independent (not paried).

Test:

H_0 will be $\mu 1 - \mu 2 = 0$ if $\mu 1$ is the mean of gasoline truck group and $\mu 2$ is the mean of diesel truck group.

I am using two different approaches to solve this problem. (because it is subjective when I compare two covariance matrix, and different people will make different conclusions.)

First

For the first one, I would consider covariance matrix is not that different. Test statistic is $T^2 = (\bar{X}_1 - \bar{X}_2)^T\{S_p(\frac{1}{n_1} + \frac{1}{n_2})\}^{-1}(\bar{X}_1 - \bar{X}_2)$ where $S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$ (where S_1 and S_2 are sample variance of each group) is the pooled sample covariance matrix (pg.422 from stat 135 textbook).

Second

(312 - 317 from the textbook) For the second one, I would say covariance matrix is different. So, I am not going to use the pooled variance for this time. Instead, I will replace the pooled variance with $\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2$, and get $T^2 = (\bar{X}_1 - \bar{X}_2)^T\{\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\}^{-1}(\bar{X}_1 - \bar{X}_2)$

And, the distribution under H_0 is $T^2 \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1}F_{p,\, n_1 + n_2 - p - 1}$. I am also going to stick with F test instead of chi-square test, since it is not really clear whether the population of each group are large (subjective). Also, even when population are large, I can still use F test, so just to be safe, I will stick with F test for this question.

Conclusion:

The mean for diesel and gasoline groups are differnt. (for both methods)

First method (pooled variance)

```r
transportation <- read.table("Data-HW1-Transportation.dat")
colnames(transportation) <- c("fuel", "repair", "capital", "type")

dim(transportation)
```

```
## [1] 59  4
```

```r
gas <- nrow(transportation[transportation$type == "gasoline",])
diesel <- nrow(transportation[transportation$type == "diesel",])

x1bar <- as.vector(apply(transportation[1:gas, 1:3], 2, mean))
s1 <- var(transportation[1:gas, 1:3])

x2bar <- as.vector(apply(transportation[(gas+1):nrow(transportation), 1:3], 2, mean))
s2 <- var(transportation[(gas+1):nrow(transportation), 1:3])

sp <- ((gas-1)*s1 + (diesel-1)*s2)/(gas+diesel-2)

t <- as.numeric(t(x1bar - x2bar) %*% solve(((1/gas)+(1/diesel))*sp) %*% (x1bar-x2bar))

pf((t*(gas+diesel-4))/((gas+diesel-2)*3), 3, 55, lower.tail = F)
```

```
## [1] 1.000461e-07
```

```
#Reject the null at the alpha = 0.05!!!
```

Second method (different covariance matrix)

```
s1
```

```
##              fuel    repair   capital
## fuel    23.013361 12.366395  2.906609
## repair  12.366395 17.544111  4.773082
## capital  2.906609  4.773082 13.963334
```

```
s2
```

```
##              fuel     repair   capital
## fuel    4.3623166  0.7598872  2.362099
## repair  0.7598872 25.8512360  7.685732
## capital 2.3620992  7.6857322 46.654400
```

```
t_second <- as.numeric(t(x1bar - x2bar) %*% solve((1/gas)*s1+(1/diesel)*s2) %*% (x1bar-x2bar))

pf((t_second*(gas+diesel-4))/((gas+diesel-2)*3), 3, 55, lower.tail = F)
```

```
## [1] 7.420961e-07
```

```
#Reject the null at the alpha = 0.05!!!
```

Problem 3

Test I choose:

One-way MANOVA

Reason to choose the test:

Subjects will be 90 male Egyptian skull. (different subjects as we have population from period 1, 2, and 3. So, it is "between subject comparison")

And, there are 4 quantitative measurements/variables: X_1 (maximum breadth of skull), X_2 (base height of skull), X_3 (base length of skull), and X_4 (nasal height of skull).

They ask me to test for differences in skull size over differnt time periods, and we have three different time periods/groups. So, it implies that we have 1 factor: time period.

And, there are 3 levels: period1 (4000 BC), period2 (3300 BC), and period3 (1850 BC).

Thus, **more than two different populations with one factor** implies that we need to use One-way MANOVA.

Test:

Refer to the pg.303 from the textbook "Applied multivariate statistcal analysis."

My null hypothesis is: $H_0 : \mu 1 = \mu 2 = \mu 3 = 0$ (where each μ_i stands for mean for each group). And, the test statistics will be distribution of Wilk's lambda, $\Lambda^* = \frac{|W|}{|B + W|} = \frac{|\sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \bar{x_l})(x_{lj} - \bar{x_l})^T|}{|\sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \bar{x})(x_{lj} - \bar{x})^T|}$,where d.f. for numerator is $\sum_{l=1}^{g} (n_l - g) = 90 - 3 = 87$ and denominator is $\sum_{l=1}^{g} (n_l - 1) = 90 - 1 = 89$. $\bar{x_l}$ stands for mean of population group1, $\bar{x_l}$ is overall mean, and x_{lj} is each observation. $= (\frac{\sum n_2 - p - 2}{p})(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}) \sim F_{2p,\ 2(\sum nl - p - 2)}$ with d.f. of 8 and 168 as p (number of quantitative variables) = 4.

So, $\Lambda^* = 0.8301$ (please refer to the R code I built). So, $\frac{\sum n_l - p - 2}{p} \frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} = \frac{84}{4} \frac{1 - \sqrt{0.8301}}{\sqrt{0.8301}} \approx 2.049 \sim F_{8,168}$. So, $P(2.049 > F_{8,168}) \approx 0.04359$, according to calculation on R.

Tip:

The mathematical method I used are well expained in pg.306-307. (For example, W can be computed easiy without calculating doulbe summations)

Anova is based on decomposition of observation.

x_{lj} (observation) = $\bar{x}$ (overall mean) + ($\bar{x_l}$ - $\bar{x}$) (estimated treatment/between effect ~ how the mean of group deviates from overall mean) + (x_{lj} - $\bar{x_l}$) (residual/within ~ how individual deviates from overall mean). So, SSobs = SSmean + SStr + SSres, where SS = Sum of Squares.

Conclusion:

P-value is around 0.04359, which is really closed to 0.05 (our significance level). However, we would still reject the null: "no treatment between groups." Thus, there are differences in skull size over differnt time periods.

I include the two differnt ways to do this problem! One is to go through all the math and logic, and the other is to use the manova function embedded in R.

```
skull <- read.table("Data-HW1-Skull.dat")
colnames(skull) <- c("maxbreadth", "baseheight", "baselength", "nasalheight", "period")

dim(skull)
```

```
## [1] 90  5
```

```
#First method

group1 <- skull %>% filter(period == 1)
group2 <- skull %>% filter(period == 2)
group3 <- skull %>% filter(period == 3)


x1bar <- apply(group1[-ncol(group1)], 2, mean)
x2bar <- apply(group2[-ncol(group2)], 2, mean)
x3bar <- apply(group3[-ncol(group3)], 2, mean)

x1bar
```

```
##  maxbreadth  baseheight  baselength nasalheight
##   131.36667   133.60000    99.16667    50.53333
```

```
x2bar
```

```
##  maxbreadth  baseheight  baselength nasalheight
##   132.36667   132.70000    99.06667    50.23333
```

```
x3bar
```

```
##  maxbreadth  baseheight  baselength nasalheight
##   134.46667   133.80000    96.03333    50.56667
```

```
cov1 <- cov(group1[-ncol(group1)])
cov2 <- cov(group2[-ncol(group2)])
cov3 <- cov(group3[-ncol(group3)])

cov1
```

```
##             maxbreadth baseheight baselength nasalheight
## maxbreadth   26.309195  4.1517241  0.4540230   7.2459770
## baseheight    4.151724 19.9724138 -0.7931034   0.3931034
## baselength    0.454023 -0.7931034 34.6264368  -1.9195402
## nasalheight   7.245977  0.3931034 -1.9195402   7.6367816
```

```
cov2
```

```
##             maxbreadth baseheight baselength nasalheight
## maxbreadth   23.136782   1.010345  4.7678161   1.8425287
## baseheight    1.010345  21.596552  3.3655172   5.6241379
## baselength    4.767816   3.365517 18.8919540   0.1908046
## nasalheight   1.842529   5.624138  0.1908046   8.7367816
```

```
cov3
```

```
##             maxbreadth  baseheight baselength nasalheight
## maxbreadth  12.1195402  0.78620690 -0.7747126  0.89885057
## baseheight   0.7862069 24.78620690  3.5931034 -0.08965517
## baselength  -0.7747126  3.59310345 20.7229885  1.67011494
## nasalheight  0.8988506 -0.08965517  1.6701149 12.59885057
```

```r
W <- 29 * (cov1 + cov2 + cov3)
overallbar <- 30/90 * (x1bar + x2bar + x3bar)
B <- (30 * tcrossprod(x1bar - overallbar)) + (30 * tcrossprod(x2bar - overallbar)) +
  (30 * tcrossprod(x3bar - overallbar))

W
```

```
##             maxbreadth baseheight baselength nasalheight
## maxbreadth   1785.4000      172.5   128.9667    289.6333
## baseheight    172.5000     1924.3   178.8000    171.9000
## baselength    128.9667      178.8  2153.0000     -1.7000
## nasalheight   289.6333      171.9    -1.7000    840.2000
```

```r
overallbar
```

```
##  maxbreadth  baseheight  baselength nasalheight
##   132.73333   133.36667    98.08889    50.44444
```

```r
B
```

```
##             [,1]       [,2]       [,3]       [,4]
## [1,]  150.200000  20.300000 -161.83333   5.033333
## [2,]   20.300000  20.600000  -38.73333   6.433333
## [3,] -161.833333 -38.733333  190.28889 -10.855556
## [4,]    5.033333   6.433333  -10.85556   2.022222
```

```r
lambda <- round(det(W)/det(B+W), 4)
lambda
```

```
## [1] 0.8301
```

```r
pf(2.049, 8, 168, lower.tail = F)
```

```
## [1] 0.04359004
```

```r
#Reject the null at 0.05!!!







#Other method!!!!!!

skull.manova <- manova(cbind(maxbreadth, baseheight, baselength, nasalheight) ~
                         as.factor(period), data = skull)
summary(skull.manova, test = "Wilks") #same as lambda above! And, also p-value should be similar as wha
```

```
##                   Df  Wilks approx F num Df den Df  Pr(>F)
## as.factor(period)  2 0.8301   2.0491      8    168 0.04358 *
## Residuals         87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#summary.manova(skull.manova, test = "Wilks") will give the same output as well!

#We are using Wilks, but there are other types: Roy, Pillai, and Hotelling-Lawley.
```

```r
#Let's see how ANOVA works here.
summary.aov(skull.manova, test = "Wilks")
```

```
##  Response maxbreadth :
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(period)  2  150.2  75.100  3.6595 0.02979 *
## Residuals         87 1785.4  20.522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response baseheight :
##                   Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(period)  2   20.6  10.300  0.4657 0.6293
## Residuals         87 1924.3  22.118
##
##  Response baselength :
##                   Df  Sum Sq Mean Sq F value  Pr(>F)
## as.factor(period)  2  190.29  95.144  3.8447 0.02512 *
## Residuals         87 2153.00  24.747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response nasalheight :
##                   Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(period)  2   2.02  1.0111  0.1047 0.9007
## Residuals         87 840.20  9.6575
```

Problem 4

Test I choose:

Two-way MANOVA

Reason to choose the test:

Subjects will be 36 seedlings (but different groups of population).

The quantitative measurements/variables are X_1 (percent spectral reflectance of wavelength 560nm) and X_2 (percent spectral reflectance at wavelength no nm). So, there are 2 quantitative measurements.

And, there are 2 factors: species and time (main reason to use two-way, not one-way).

And, there are 3 levels for each factor: SS, JL, and LP for species factor and 1, 2, and 3 for time factor.

Thus, **more than two different populations with two factors** implies that we should use "Two-way MANOVA."

Test:

My null hypotheses H_0 are: (μ and β are means for each group. For example, μ_{11} will be the mean of both first level of both factors)

1. H_0^{int}: $\mu_{11} = \mu_{12} = \mu_{13} = \mu_{21} = \mu_{22} = \mu_{23} = \mu_{31} = \mu_{32} = \mu_{33} = 0$ (no interaction effect)
2. H_0^{fac1}: $\mu_1 = \mu_2 = \mu_3 = 0$
3. H_0^{fac2}: $\beta_1 = \beta_2 = \beta_3 = 0$ (I am following the notations from the textbook)

Test statistics for each H_0 are (where Λ^* is Wilk's lambda and SSP = Sum of Squares and cross Products):

1. $\Lambda^*_{int} = \frac{SSP_{res}}{SSP_{int} + SSP_{res}}$
2. $\Lambda^*_{fac1} = \frac{SSP_{res}}{SSP_{fac1} + SSP_{res}}$
3. $\Lambda^*_{fac2} = \frac{SSP_{res}}{SSP_{fac2} + SSP_{res}}$

I will do a test for interaction before the tests for main factor effects, because if interaction effects exist, the factor effects do not have a clear interpretation. Thus, we do not need to proceed additional multivariate tests (pg.316)

Conclusion:

Reject all of three null hypothesis. Thus, there is no species effect, time effect, and species-time interaction effect.

```r
sensing <- read.table("Data-HW1-Sensing.dat")
colnames(sensing) <- c("560nm", "720nm", "species", "time", "replication")

dim(sensing)
```

```
## [1] 36  5
```

```r
#First method

sensing.manova <- manova(cbind(`560nm`, `720nm`) ~ as.factor(species) * as.factor(time),
                         data = sensing)
```

```r
summary.manova(sensing.manova, test = "Wilks")
```

```
##                                     Df    Wilks approx F num Df den Df
## as.factor(species)                   2 0.068774   36.571      4     52
## as.factor(time)                      2 0.049166   45.629      4     52
## as.factor(species):as.factor(time)  4 0.087070   15.528      8     52
## Residuals                           27
##                                        Pr(>F)
## as.factor(species)                  1.554e-14 ***
## as.factor(time)                     < 2.2e-16 ***
## as.factor(species):as.factor(time) 2.217e-11 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Reject the null hypothesis!!!

#Check anova
summary.aov(sensing.manova, test = "Wilks")
```

```
##  Response 560nm :
##                                     Df  Sum Sq Mean Sq F value    Pr(>F)
## as.factor(species)                   2  965.18  482.59 169.973 5.027e-16
## as.factor(time)                      2 1275.25  637.62 224.578 < 2.2e-16
## as.factor(species):as.factor(time)  4  795.81  198.95  70.073 7.341e-14
## Residuals                           27   76.66    2.84
##
## as.factor(species)                  ***
## as.factor(time)                     ***
## as.factor(species):as.factor(time) ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response 720nm :
##                                     Df Sum Sq Mean Sq F value    Pr(>F)
## as.factor(species)                   2 2026.9 1013.43 15.4622 3.348e-05 ***
## as.factor(time)                      2 5573.8 2786.90 42.5207 4.537e-09 ***
## as.factor(species):as.factor(time)  4  193.5   48.39  0.7383    0.5741
## Residuals                           27 1769.6   65.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```