HW3

STA104

Filip Wilhelm Sjostrand

2022-11-10

Question I:

a)

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_a : at least one μ_i is different

$\alpha = 0.05$

```r
# Packages ----------
library(dplyr)
library(readr)

# Data ----------
g1 <- c(2.9736, 0.9448, 1.6394, 0.0389, 1.2958)
g2 <- c(0.7681, 0.8027, 0.2156, 0.074, 1.5076)
g3 <- c(4.8249, 2.2516, 1.5609, 2.0452, 1.0959)
outcome <- c(g1,g2,g3)
treatment <- c(rep(1,length(g1)),rep(2,length(g2)),rep(3,length(g3)))
df <- data.frame(outcome, treatment)
outcome <- df$outcome
treatment <- df$treatment

# Observed F stat ----------
model <- summary(lm(outcome~treatment))
Fobs <- model[[10]][1]

# Permutation ----------
tot <- 10000
p <- c()
f <- c()

for(i in 1:tot){
  permut <- sample(outcome)
  model1 <- lm(permut~treatment)
  f[i] <- summary(model1)[[10]][1]
  p[i] <- (f[i]>=Fobs)+0
}

pvalue1 <- sum(p)/tot
```

Given that $p-value = 0.2358 > \alpha$ we fail to reject the null. Thus, there is insufficient evidence at significance level 5% that there is at least one population mean different from the others.

b)

Reject if $F_{obs} > F_{crit}$.

```
# Critical value approach ----------
critical <- qf(0.05, 3-1, 15-3, lower.tail = FALSE)
```

Given that $F_{obs} = 1.628$ and $F_{crit} = 3.8853$ we fail to reject the null. Hence, with the critical value approach we still conclude at significance level 5% there is insufficient evidence that at least one population mean differs from the others.

Question II:

```
# Get data ----------
df <- read_delim("~/Documents/UC Davis/Courses/STA 104/hw3 prob2.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
outcome <- df$`Femur Load`
treatment <- df$Weight

# Observed F stat ----------
model <- summary(lm(outcome~treatment))
Fobs <- model[[10]][1]

# Permutation ----------
tot <- 10000
p <- c()
f <- c()

for(i in 1:tot){
  permut <- sample(outcome)
  model1 <- lm(permut~treatment)
  f[i] <- summary(model1)[[10]][1]
  p[i] <- (f[i]>=Fobs)+0
}

pvalueP <- sum(p)/tot

# Critical value approach ----------
k <- df %>% select(Weight) %>% n_distinct()
N <- nrow(df)

pvalueA <- pf(Fobs, k-1, N-k, lower.tail = FALSE)
```

The p-value from the permutation test is 0.342 and for the ANOVA test it is 0.455. Since the both values are approximately equal, it is reasonable to assume that the data is approximately normal. This is justified by the fact that the ANOVA model assume normal distribution of the errors.

Question III:

```
# Data ----------
g1 <- c(2.9736, 0.9448, 1.6394, 0.0389, 1.2958)
g2 <- c(0.7681, 0.8027, 0.2156, 0.074, 1.5076)
g3 <- c(4.8249, 2.2516, 1.5609, 2.0452, 1.0959)
outcome <- c(g1,g2,g3)
treatment <- c(rep(1,length(g1)),rep(2,length(g2)),rep(3,length(g3)))

# Kruskal-Wallis statistic ----------
model <- kruskal.test(outcome~treatment)
ksobs <- model[[1]][1]

# Permutation test KS ----------
tot <- 10000
p=c()
ks=c()

for(i in 1:tot){
  permut <- sample(outcome)
  model <- kruskal.test(permut~treatment)
  ks[i] <- model[[1]][1]
  p[i] <- (ks[i]>=ksobs)+0
}

pvalue=sum(p)/tot
```

Using the permutation F test we got a p-value of 0.24 and from the Kruskal-Wallis permutation test we got a p-value of 0.048. In the first case, we fail to reject and in the second we reject the null at $alpha = 0.05$. Hence, when using the ranks, the permutation test provide significant evidence that there is at least one population mean different from the others.

Question VI:

Kruskal-Wallis test

$H_0 : \bar{R}_1 = \bar{R}_2 = ... = \bar{R}_7$

H_a : at least one $\bar{R}_i$ is different

$\alpha = 0.05$

```
# Get data ----------
df <- headinjury <- read_delim("~/Documents/UC Davis/Courses/STA 104/headinjury.csv",
    delim = ";", escape_double = FALSE, col_types = cols(Type = col_factor(levels = c("1",
        "2", "3", "4", "5", "6", "7"))),
    trim_ws = TRUE)
outcome <- df$`Head injury`
treatment <- df$Type
k <- n_distinct(df$Type)
```

```
# Kruskal-Wallis statistic ----------
model <- kruskal.test(outcome~treatment)
ksobs <- model[[1]][1]
```

From the Kruskal-Wallis test we derived our test statistic $KS = 18.2901$. From table A7 with $df = 6$, we find the critical value to be $= 12.6$. Since $KS > 12.6$, we reject H_0 and conclude that at 5% significance level there is sufficient evidence that at least one of the population means based on the ranks are different from the others.

Comparison Test

$H_0 : \hat{D}_i^R = 0$

$H_a : \hat{D}_i^R > 0$

$\alpha = 0.05$

LSD:

```
# Ranked means ----------
rm <- df %>%
  mutate(rank=rank(df$`Head injury`)) %>%
  group_by(Type) %>%
  summarise(rankMean=mean(rank))

# Differences ----------
comb <- data.frame(combn(rm$rankMean, 2))
d <- c()

for(i in 1:ncol(comb)){
  d[i] <- comb[,i][1]-comb[,i][2]
}

pairs <- c(
  "1-2", "1-3", "1-4", "1-5", "1-6", "1-7",
  "2-3", "2-4", "2-5", "2-6", "2-7",
  "3-4", "3-5", "3-6", "3-7",
  "4-5", "4-6", "4-7",
  "5-6", "5-7",
  "6-7"
  )

# LSD ----------
z <- qnorm(0.05/2, lower.tail = FALSE)
N <- max(rank(df$`Head injury`))
mseReplace <- (N*(N+1))/(12)
n <- df %>% filter(Type==1) %>% count()
s <- sqrt(mseReplace*(2/n))

result <- c()
for(i in 1:length(d)){
  if(abs(d[i])>=s*z){
   result[i] <- TRUE
  }
  else{
    result[i] <- FALSE
  }
}

LSD <- data.frame(Pair=pairs, Difference = d, Significance=result )
```

Table 1: LSD: Ranked Poupulation Mean Differences

Pair	Difference	Significance
1-2	14.7	FALSE
1-3	-2.9	FALSE
1-4	12.2	FALSE
1-5	-11.5	FALSE
1-6	-6.9	FALSE
1-7	-14.7	FALSE
2-3	-17.6	FALSE
2-4	-2.5	FALSE
2-5	-26.2	TRUE
2-6	-21.6	TRUE
2-7	-29.4	TRUE
3-4	15.1	FALSE
3-5	-8.6	FALSE
3-6	-4.0	FALSE
3-7	-11.8	FALSE
4-5	-23.7	TRUE
4-6	-19.1	TRUE
4-7	-26.9	TRUE
5-6	4.6	FALSE
5-7	-3.2	FALSE
6-7	-7.8	FALSE

Thus we conclude that at 5% family wise significance level, $\mu_2 < \mu_5$, $\mu_2 < \mu_6$, $\mu_2 < \mu_7$, $\mu_4 < \mu_5$, $\mu_4 < \mu_6$, and $\mu_4 < \mu_7$. We fail to reject the other ranked mean differences.

HSD:

```
# HSD ----------
q <- 4.17
mseReplace <- (N*(N+1))/(24)
s <- sqrt(mseReplace*(2/n))

result <- c()
for(i in 1:length(d)){
  if(abs(d[i])>=s*q){
   result[i] <- TRUE
  }
  else{
    result[i] <- FALSE
  }
}

HSD <- data.frame(Pair=pairs, Difference = d, Significance=result )
```

Table 2: HSD: Ranked Poupulation Mean Differences

Pair	Difference	Significance
1-2	14.7	FALSE
1-3	-2.9	FALSE
1-4	12.2	FALSE
1-5	-11.5	FALSE
1-6	-6.9	FALSE
1-7	-14.7	FALSE
2-3	-17.6	FALSE
2-4	-2.5	FALSE
2-5	-26.2	FALSE
2-6	-21.6	FALSE
2-7	-29.4	TRUE
3-4	15.1	FALSE
3-5	-8.6	FALSE
3-6	-4.0	FALSE
3-7	-11.8	FALSE
4-5	-23.7	FALSE
4-6	-19.1	FALSE
4-7	-26.9	TRUE
5-6	4.6	FALSE
5-7	-3.2	FALSE
6-7	-7.8	FALSE

Thus we conclude that at 5% family wise significance level, $\mu_2 < \mu_7$ and $\mu_4 < \mu_7$. We fail to reject the other ranked mean differences.

Question VIII

```
# Permutation HSD ---------
tot <- 1000
d <- c()
thsd <- c()

for(i in 1:tot){
  permut <- sample(outcome)
  df <- data.frame(treatment, outcome)
  model <- lm(permut~treatment)
  mse <- (sum((summary(model)$residuals)^2))/(N-k)

  m <- df %>%
    group_by(treatment) %>%
    summarise(mean=mean(outcome))

  mean <- m$mean
  maxmean <- max(mean)
  minmean <- min(mean)

  thsd[i] <- (maxmean-minmean)/( (1/10)*(sqrt(mse)) )

}

cv5 <- quantile(thsd,.95)
cv10 <- quantile(thsd, .90)
critical <- c(cv5, cv10)
table <- data.frame(critical)
```

Table 3: Critical Values for HSD permutation

	critical
95%	16.16580
90%	15.96782