Homework 5

Joshua Oswari - A14751270

05/19/2019

Problem 1 (Permutation test)

Question: Is smoking associated with a decrease in fitness level?

```
load("smokers.rda")
smokers
```

```
##   non light moderate heavy
## 1  69    55       66    91
## 2  52    60       81    72
## 3  71    78       70    81
## 4  58    58       77    67
## 5  59    62       57    95
## 6  65    66       79    84
```

```
# Non-smokers, light smokers, moderate smokers and heavy smokers (six in each group) undertook sustaine
# Their heart rates were measured after resting for three minutes.
```

Null hypothesis: distribution of heart rates for non/light/moderate/heavy smokers is the same.

Convert to the appropriate form.

```
g = numeric(0)
g = c(g, rep("non", 6))
g = c(g, rep("light", 6))
g = c(g, rep("moderate", 6))
g = c(g, rep("heavy", 6))
g
```

```
##  [1] "non"      "non"      "non"      "non"      "non"      "non"
##  [7] "light"    "light"    "light"    "light"    "light"    "light"
## [13] "moderate" "moderate" "moderate" "moderate" "moderate" "moderate"
## [19] "heavy"    "heavy"    "heavy"    "heavy"    "heavy"    "heavy"
```

```
y = unlist(smokers, use.names=FALSE)
y
```

```
##  [1] 69 52 71 58 59 65 55 60 78 58 62 66 66 81 70 77 57 79 91 72 81 67 95
## [24] 84
```

One way ANOVA:

```
mydf = data.frame(smoke = g, heartrate = y)
mydf
```

```
##       smoke heartrate
## 1       non        69
## 2       non        52
## 3       non        71
## 4       non        58
## 5       non        59
## 6       non        65
```

```
## 7     light        55
## 8     light        60
## 9     light        78
## 10    light        58
## 11    light        62
## 12    light        66
## 13 moderate        66
## 14 moderate        81
## 15 moderate        70
## 16 moderate        77
## 17 moderate        57
## 18 moderate        79
## 19    heavy        91
## 20    heavy        72
## 21    heavy        81
## 22    heavy        67
## 23    heavy        95
## 24    heavy        84
```

```
res.aov <- aov(heartrate ~ smoke, data=mydf)
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## smoke        3   1464   488.0    6.12 0.00398 **
## Residuals   20   1595    79.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# permute data, and get another F. compare it to this F

# how to get f value
mylm <- lm(heartrate ~ smoke, data=mydf)
myF <- summary(mylm)$fstatistic[1]
myF
```

```
##    value
## 6.120284
```

What's wrong with this procedure? Why consider permutation test? The p-value calculated above relies on the assumption that the observations are * iid * normal * homoscedastic (groups have equal variance)

```
# permF.test <- function(y, g, B=2000) {
#   g = as.factor(g)
#   dat = split(y, g)
#   F_obs = oneway.test(y ~ g)$stat
#   F_pi = numeric(B)
#   #bootstrap
#   dat_centered = lapply(dat, scale, center = T, scale = F)
#   for (b in 1:B){
#    # dat_boot = lapply(dat_centered, bootstrap)
#     y_boot = stack(dat_boot)$values
#     g_boot = stack(dat_boot)$ind
#     F_pi[b] = oneway.test(y_boot ~ g_boot)$stat
#   }
#   # return ratio of number of times F_pi is larger than F_obs
#   p_val = (sum(F_pi >= F_obs, na.rm = T) + 1)/(B+1)
```

```
#   return(p_val)
#
# }
```

```
# Part c
dat <- read.csv("cars.csv")

#permF.test(dat$City.mpg, dat$Transmission)
```

Problem 2 (Multiple Testing):

Why? Suppose we have data consisting of wage and 10,000 variables that may be associated with wage. Doing pairwise tests with alpha=0.05, we will conclude that ~500 variables are associated with wage, when they have nothing to do with wage!

FWER: family-wise error rate (FWER) is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests.

FDR: The false discovery rate (FDR) of a given multiple testing procedure is the expected proportion of false rejections it makes in a given situation.

FWER control => FDR control

```
fifa <- read.csv("fifa.csv")
#fifa

# convert Wage column to numeric
library(stringr)
W <- as.character(fifa$Wage)
fifa$Wage <- as.numeric(str_extract(W, "[[:digit:]]+"))

# age, nationality, club,
# whatever is not an index
```

Note: Some variables have +2, +3 etc.. Remove it and convert the column to integer type.

For 2 variables..

```
pval = numeric(2)

lm1 <- lm(Wage ~ GKKicking, data=fifa)
pval[1] <- summary(lm1)$coef[,4][2]

lm2 <- lm(Wage ~ Acceleration, data=fifa)
pval[2] <- summary(lm2)$coef[,4][2]

# corrected p-values
pval.bon = p.adjust(pval, "bon") # Theorem: The Bonferroni procedure controls the FWER at alpha.
pval.holm = p.adjust(pval, "holm")
pval.hoch = p.adjust(pval, "hoch")
pval.bh = p.adjust(pval, "BH")
pval.by = p.adjust(pval, "BY")

# Check output
pval
```

```
## [1] 1.348197e-04 3.958930e-64
```

```
pval.bon
```

```
## [1] 2.696393e-04 7.917859e-64
```

```
pval.holm
```

```
## [1] 1.348197e-04 7.917859e-64
```

```
pval.bh
```

```
## [1] 1.348197e-04 7.917859e-64
```

```
pval.by
```

```
## [1] 2.022295e-04 1.187679e-63
```

```
# rejections at the 10% level without adjustement for multiple testing
reject = (pval <= 0.10)
R = sum(reject) # total number of rejections

# rejections at the 10% FWER level (the last one does not guarantee control since assumption of indepen
reject.bon = (pval.bon <= 0.10)
R.bon = sum(reject.bon)
reject.holm = (pval.holm <= 0.10)
R.holm = sum(reject.holm)
reject.hoch = (pval.hoch <= 0.10)
R.hoch = sum(reject.hoch)

# rejections at the 10% FDR level (the first one does not guarantee control since assumption of indepen
reject.bh = (pval.bh <= 0.10)
R.bh = sum(reject.bh)
reject.by = (pval.by <= 0.10)
R.by = sum(reject.by)
```