Hypothesis_Testing_Gender

Yuan Wang

2/6/2017

Description

-

  Goal

  We want to test within selected sample that if, when consider gender only, female and male use different time to finish the game. The overall goal is to see if students in different groups with sample size less than 50 will achieve different conclusions.

-

  Hypothesis

  We hypothesis that gender influences TimeUsed.

-

  Method

  - We first filter out a gender data using v1label1.
  - Then, we mutate a column to the gender data where if the user was male we denote as 1, and if female we denote as 0.
  - We use a for loop to select all groupID's whose sample size is less then 50 and greater than 5. Also we use only groups that have more than 1 male and 1 female observations. (Otherwise t.test would fail).
  - We perform the two-sample t-test within each group.
  - We then perform a two-sample t-test for the over all gender data and yield the p-value as a comparison

Libraries

Import and Trim Data

We use only untimed data with non-zero time used recording.

```
Original <- read.csv("Original.csv")
shape <- Original
shape_untime <- shape[shape$requestedTime==0,]
shape_untimed <- shape_untime[shape_untime$timeUsed!=0,]
```

We also changed some numerical variables to factors since they have only a fixed number of levels. We also changed the unit of the time data from milliseconds to seconds.

```
shape_untimed$numShapes <- as.factor(shape_untimed$numShapes)
shape_untimed$matchingScheme <- as.factor(shape_untimed$matchingScheme)
shape_untimed$requestedTime <- as.factor(shape_untimed$requestedTime)
```

```
shape_untimed$timeUsed <- as.numeric(shape_untimed$timeUsed)
shape_untimed$timerDisplay <- as.factor(shape_untimed$timerDisplay)
shape_untimed$numErrors <- as.numeric(shape_untimed$numErrors)

shape_untimed <- mutate(shape_untimed, TimeUsedSec = shape_untimed$timeUsed/1000)
```

We further trimed the data to include a varible column named Gender.

```
gender <- filter(shape_untimed, tolower(strtrim(shape_untimed$v1label,3))=="gen" |
                   tolower(strtrim(shape_untimed$v1label,3))=="sex" |
                   tolower(strtrim(shape_untimed$v1label,4))=="male" |
                   tolower(strtrim(shape_untimed$v1label,1))=="f")

case <- (tolower(strtrim(gender$v1value,1)) == "m") & (tolower(strtrim(gender$v2value,1)) =="f")
gender1 <- gender[(strtrim(gender$v1value,1) != "1") &
                    (gender$v1value != "2") &
                    (gender$v1label != "Gender\rOrder") &
                    (gender$v1label != "female\rorder") &
                    (gender$v1value != "0") &
                    (tolower(gender$v1value) != "morf") &
                    (gender$studentID != "mb") &
                    (gender$studentID != "31207") &
                    (gender$v1value != "cat") &
                    (gender$studentID != "3659") &
                    (tolower(gender$v1value) != "attempt") &
                    (!case),]

gender1 <- mutate(gender1,
                  gender = as.factor(ifelse(
                    pmax((tolower(strtrim(gender1$v1label,1)) == "m"),
                         (tolower(strtrim(gender1$v1value,1)) == "m")),
                    1,
                    ifelse(pmax(tolower(strtrim(gender1$v1label,1)) == "f",
                                (tolower(strtrim(gender1$v1value,1)) =="f")),
                           0,
                           NA))))

gender2 <- na.omit(gender1)
```

Extract GroupID Information

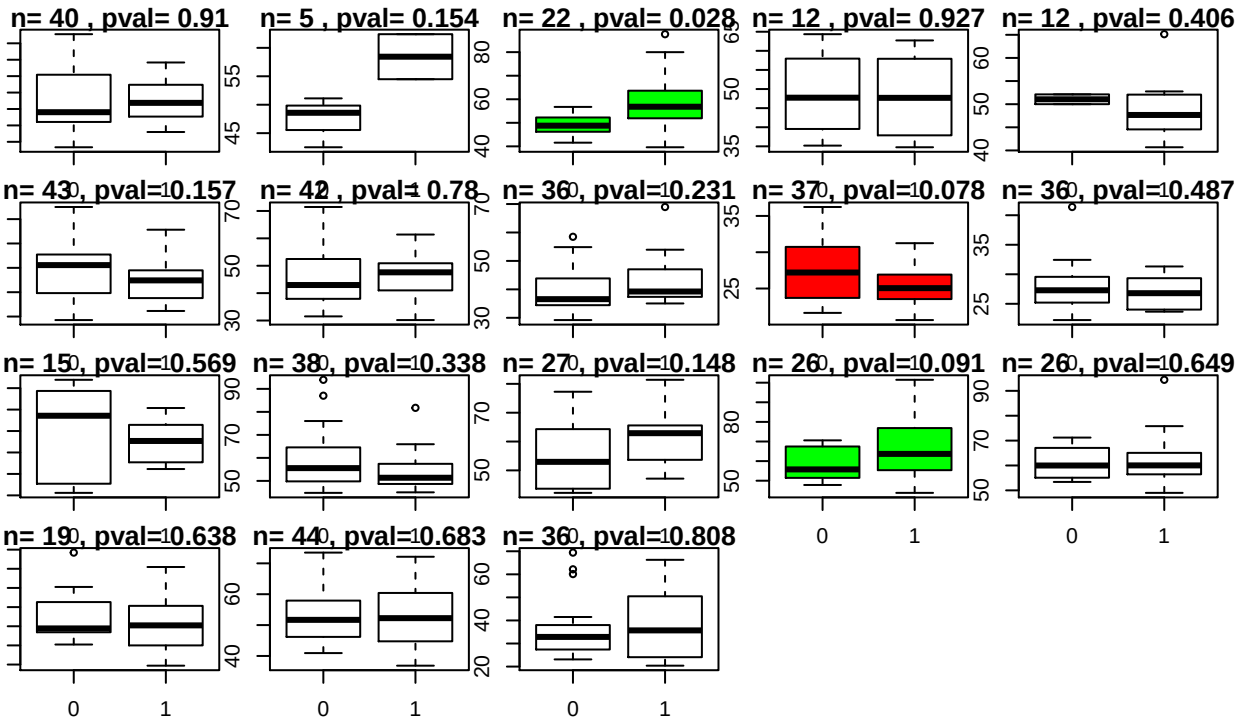
Then, within the gender data, we identified the group names with sample size between 5 and 50. We make a vector of groupID of our desired groups.

```
tb <- as.data.frame(table(gender2$groupID))

##Select groups that are under size 50
tb2 <- tb[tb$Freq >= 5 & tb$Freq <= 50,]

##Create a vector of groupID's whose size is between 25 and 50
selected_groupID <- as.character(tb2$Var1)
```

T-test and Boxplots

We will write a function to conduct a two-sided two sample t-test. We use the Student's t.test. Our Null Hypothesis is that consider only gender, female and male use same time. Our Alternative Hypothesis is that female and male use different times. Since t.test does not work when either sample has less than 2 observations, we further filter the groupID so that each group has at least two female and two male observations. We will show the boxplot of all groups using gender as the independet variable and time used (seconds) as the dependent variable. In the title of each graph, we will specify individual sample size and p-value. Choose alpha = 0.1, we color the those plots whose group reject the null hypothesis of no difference in green.

```
par(mar=c(1,1,1,1))
par(mfrow = c(5,5))
groupName <- c()
pvalues <- c()
for (i in 1:length(selected_groupID)) {
  female <- gender2[gender2$groupID == selected_groupID[i] & gender2$gender == 0,]$TimeUsedSec
  male <- gender2[gender2$groupID == selected_groupID[i] & gender2$gender == 1,]$TimeUsedSec
  if (length(female) > 1 & length(male) > 1) {
    groupName <- cbind(groupName, selected_groupID[i])
    p <- round(t.test(female, male)$p.value, digits = 3)
    pvalues <- cbind(pvalues, p)
    data1 <- gender2[gender2$groupID == selected_groupID[i],]
    if (p < 0.1) {
      if (mean(female) < mean(male)) {
      boxplot(TimeUsedSec ~ gender,data=data1, col="green", main=paste("n=", dim(data1)[1], ", pval=", p
    } else {
      boxplot(TimeUsedSec ~ gender,data=data1, col="red", main=paste("n=", dim(data1)[1], ", pval=", p)
    }
  } else {
    boxplot(TimeUsedSec ~ gender,data=data1, main=paste("n=", dim(data1)[1], ", pval=", p), xlab="Gende
  }

  write.csv(gender2[gender2$groupID == selected_groupID[i],], file =paste(selected_groupID[i], "group.cs
  }
}
```

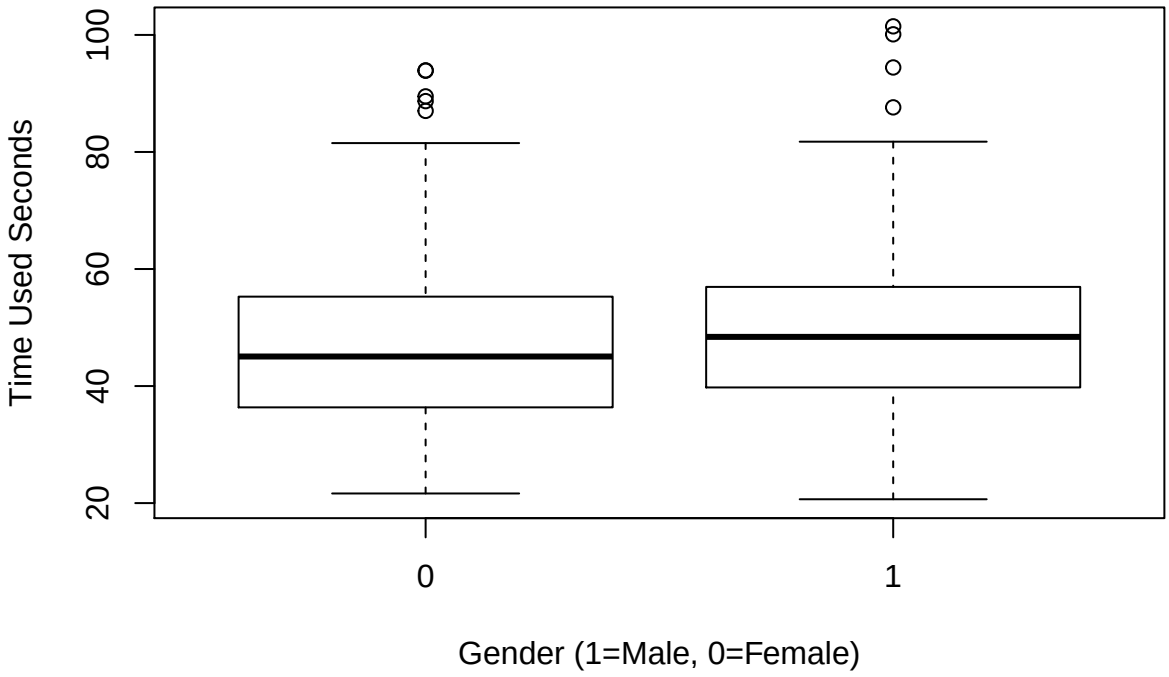


We can also draw a boxplot for the over gender2 data and calculate a p-value as comparison to the group-wise plots and p-values.

```
temp <- list.files(pattern = '*group.csv')
data <- rbind(read.csv(temp[1])[-c(1)], read.csv(temp[2])[-c(1)])
data <- rbind(data, read.csv(temp[3])[-c(1)])
data <- rbind(data, read.csv(temp[4])[-c(1)])
data <- rbind(data, read.csv(temp[5])[-c(1)])
data <- rbind(data, read.csv(temp[6])[-c(1)])
data <- rbind(data, read.csv(temp[7])[-c(1)])
data <- rbind(data, read.csv(temp[8])[-c(1)])
data <- rbind(data, read.csv(temp[9])[-c(1)])
data <- rbind(data, read.csv(temp[10])[-c(1)])
data <- rbind(data, read.csv(temp[11])[-c(1)])
data <- rbind(data, read.csv(temp[12])[-c(1)])
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = structure(c(8L, 8L, 6L, 8L, :
## invalid factor level, NA generated
```

```
data <- rbind(data, read.csv(temp[13])[-c(1)])
data <- rbind(data, read.csv(temp[14])[-c(1)])
data <- rbind(data, read.csv(temp[15])[-c(1)])
data <- rbind(data, read.csv(temp[16])[-c(1)])
data <- rbind(data, read.csv(temp[17])[-c(1)])
data <- rbind(data, read.csv(temp[18])[-c(1)])


female <- data[data$gender == 0,]$TimeUsedSec
male <- data[data$gender == 1,]$TimeUsedSec
pvl2 <- round(t.test(female, male)$p.value, digits = 3)
boxplot(TimeUsedSec ~ gender,data=data, main=paste("n=", dim(data)[1], ", two sided pval=", pvl2), xlab=
```


n= 604 , two sided pval= 0.029
Time Used Seconds
20
40
60
80
100
0
1
Gender (1=Male, 0=Female)