HW6_Frank_Jiang

Frank Jiang

11/12/2019

```
#6.1
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Afterschool_data<- read.csv('C:/HW/HW6/AfterSchool.csv')
#examine the data frame object
head(Afterschool_data)
```

```
##   ID Treatment  Aggress   Delinq   Victim
## 1  1         0 63.16264 44.46308 64.42996
## 2  2         0 51.82728 76.81361 64.42996
## 3  3         0 74.49800 50.93319 41.54106
## 4  4         0 40.49192 44.46308 41.54106
## 5  5         0 56.36143 44.46308 52.98551
## 6  6         0 62.70923 50.93319 70.15219
```

```
tail(Afterschool_data)
```

```
##      ID Treatment  Aggress   Delinq   Victim
## 351 351         1 38.22485 44.46308 41.54106
## 352 352         1 45.02607 70.34351 64.42996
## 353 353         1 67.69678 57.40329 52.98551
## 354 354         1 38.22485 44.46308 41.54106
## 355 355         1 49.56021 44.46308 41.54106
## 356 356         1 65.42971 44.46308 41.54106
```

```
str(Afterschool_data)
```

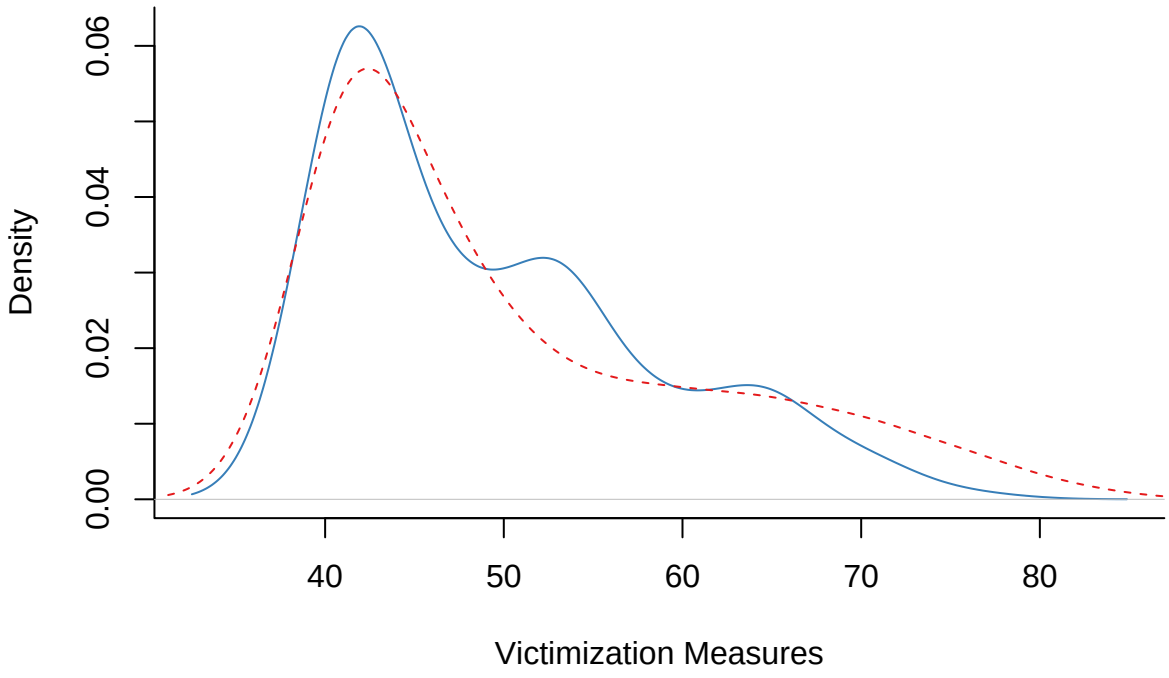
```
## 'data.frame':    356 obs. of  5 variables:
```

```
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Treatment: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Aggress  : num  63.2 51.8 74.5 40.5 56.4 ...
##  $ Delinq   : num  44.5 76.8 50.9 44.5 44.5 ...
##  $ Victim   : num  64.4 64.4 41.5 41.5 53 ...
```

```r
summary(Afterschool_data)
```

```
##        ID           Treatment         Aggress          Delinq
##  Min.   :  1.00   Min.   :0.0000   Min.   :38.22   Min.   :44.46
##  1st Qu.: 89.75   1st Qu.:0.0000   1st Qu.:42.76   1st Qu.:44.46
##  Median :178.50   Median :0.0000   Median :47.29   Median :44.46
##  Mean   :178.50   Mean   :0.4747   Mean   :50.05   Mean   :49.92
##  3rd Qu.:267.25   3rd Qu.:1.0000   3rd Qu.:56.36   3rd Qu.:50.93
##  Max.   :356.00   Max.   :1.0000   Max.   :79.03   Max.   :89.75
##      Victim
##  Min.   :41.54
##  1st Qu.:41.54
##  Median :47.26
##  Mean   :49.98
##  3rd Qu.:52.99
##  Max.   :81.60
```

```r
#Exploratory analysis (denisty plot)
#for treatment group
plot(density(Afterschool_data$Victim[Afterschool_data$Treatment==1],bw=3),
     main=" ",xlab="Victimization Measures", bty="l",col="#377EB8",lty="solid")
#for control group
lines(density(Afterschool_data$Victim[Afterschool_data$Treatment==0]),
      col="#E41A1C",lty="dashed")
#Add legend
legend(x = 50, y = 0.085, legend = c('Control Group', 'Treatment Group'),
         col = c("#E41A1C","#377EB8"), lty= c("dashed", "solid"), bty = "n")
```




```
#Conditional Means
conditional_means<- tapply(X=Afterschool_data$Victim,INDEX = Afterschool_data$Treatment,FUN = mean)
#Conditional standard deviations
conditional_sd<- tapply(X=Afterschool_data$Victim, INDEX= Afterschool_data$Treatment, FUN = sd)
#sample sizes
sample_sizes<- table(Afterschool_data$Treatment)
#summary table
summary<- rbind(conditional_means,conditional_sd,sample_sizes)

#randomization
#randomly permute the victimization measures and assign them to an object
#called random_victim
random_victim<- sample(Afterschool_data$Victim)
summary(random_victim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   41.54   41.54   47.26   49.98   52.99   81.60
```

```
#mean difference between treatment group and control group with randomized data
mean(random_victim[1:187])-mean(random_victim[188:356])
```

```
## [1] 0.917099
```

```
#repeat the randomization for p-value
#repeat 4999 times
victim_permuted<- replicate(n=4999, expr=sample(Afterschool_data$Victim))
#create function to calculate the mean difference of two groups
mean.diff <- function(data) {
```

```
  mean(data [1:187]) - mean(data[188:356])
}
mean.diff(Afterschool_data$Victim)
```

```
## [1] 1.303856
```

```
#calculate mean difference of each column with different groups
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
#count groups of permuted mean difference higher than 1.3
r<- length(diffs[abs(diffs)>= 1.3])
#calculate Monte Carlo P-value
p_value<- (r+1)/(4999+1)
p_value
```

```
## [1] 0.228
```

Write-up

Three-hundred fifty-six-middle-school students were randomly assigned to participating in an after-school program(n=169), or were given 'normal' treatment(n=187), which is that they were invited to attend one after-school activity per month.The treatment group(mean=50.6,sd=10.9) have slightly higher victimization measures with larger variance than the control group(mean=49.3,sd=8.8).Based on the density plot, we can conclude that the density distribution of victimization measures for the treatment group and the control group are similar. However, there are some variation when the victimazation measures reaches from 55 to 70. A randomization test were used to determine whether if there was a statistical reliable difference in the effect of victimization between students in these two groups. A Monte Carlo P-value was computed by permuting the data 4999 times, using the correction p-value 0.225. This is weak evidence against null hypothesis of no treatment effect, and may suggest that after-school programs do not contribute to differences in victimization between students who fully participate and those who don't.

```
#6.2
#create function to calculate the mean difference of two groups
mean.diff <- function(data) {
  mean(data [1:187]) - mean(data[188:356])
}
mean.diff(Afterschool_data$Victim)
```

```
## [1] 1.303856
```

```
#a)permute 100 times
victim_permuted<- replicate(n=100, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value<- (r+1)/(100+1)
cat("The p value for permuted 100 times is",p_value,". ")
```

```
## The p value for permuted 100 times is 0.2376238 .
```

```
#b)500 times
victim_permuted<- replicate(n=500, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value<- (r+1)/(500+1)
cat("The p value for permuted 500 times is",p_value,". ")
```

```
## The p value for permuted 500 times is 0.2295409 .
```

```r
#c)1000 times
victim_permuted<- replicate(n=1000, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value<- (r+1)/(1000+1)
cat("The p value for permuted 1000 times is",p_value,". ")
```

```
## The p value for permuted 1000 times is 0.2107892 .
```

```r
#d)5000 times
victim_permuted<- replicate(n=5000, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value <- (r+1)/(5000+1)
cat("The p value for permuted 5000 times is",p_value,". ")
```

```
## The p value for permuted 5000 times is 0.2173565 .
```

```r
#e)10000 times
victim_permuted<- replicate(n=10000, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value<- (r+1)/(10000+1)
cat("The p value for permuted 10000 times is",p_value,". ")
```

```
## The p value for permuted 10000 times is 0.2276772 .
```

```r
#f)100,000 times
victim_permuted<- replicate(n=100000, expr=sample(Afterschool_data$Victim))
diffs <- apply(X = victim_permuted, MARGIN = 2, FUN = mean.diff)
r<- length(diffs[abs(diffs)>= 1.3])
p_value<- (r+1)/(100000+1)
cat("The p value for permuted 100000 times is",p_value, ". ")
```

```
## The p value for permuted 100000 times is 0.2206778 .
```

<p align="left> The p-value for the randomization test decreases when the times of permutation increases from 100 times to 1000 times, and then the p-value increases as the permutation times keep increasing. The permutation times we choose for the randomization test could possibly affect the conclusion.Therefore, my suggestion is that we need to find the appropriate times of permutation for our randomization test. It is not neccessarily always better to randomize a large amount of times.