# homework3_jn2761

jiaying Ning

10/18/2020

```
rm(list=ls())
library(readxl)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

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

# Problem 1

```
#Import Data
Exercise_df=
   read.csv("./data/Exercise.csv")
```

A study was conducted over a six-month period at a local ambulatory virology clinic. The goal was to test the effect of a structured exercise program for overweight/obese, virally suppressed HIV positive subjects on different parameters. A total of 36 individuals agreed to participate in the intervention group (group 1) and another group of 36 individuals were selected as controls (group 0). The table below shows descriptive statistics: mean(SD), median(Q1, Q3) to summarize the Systolic Blood Pressure (SBP) variable by groups at baseline (pre), at 6 months follow-up (post) and also the absolute changes (Δ=Post-Pre). We want to perform some tests to assess changes in SBP for the two groups (within and between). For each question, make sure to state the formulae for hypotheses, test-statistics, decision rules/p-values, and provide interpretations in the context of the problem. Use a type I error of 0.05 for all tests. Note: The raw dataset 'Exercise.csv' used to generate this table can be found on Canvas.

**a) Perform appropriate tests to assess if the Systolic BP at 6 months is significantly different from the baseline values for each of the groups: i) Intervention group (5p)**

since we are comparing I will use paired-t test to test the hypothesis.

- Null Hypothesis: There is no difference between baseline values and Systolic BP at 6 month for intervention group.

- Alternative Hypothesis: There is difference between baseline values and Systolic BP at 6 month for intervention group

```
#creating intervention group dataframe
Exercise_df_intervention  =
Exercise_df %>%
  filter(Group==1) %>%
  mutate(difference = Systolic_POST - Systolic_PRE)
```

```
#perform paired-t.test
sd_diff<-sd(pull(Exercise_df_intervention,difference))
test_weight<-mean(pull(Exercise_df_intervention,difference))/(sd_diff/sqrt(length(pull(E
xercise_df_intervention,difference))))
#perform paired-t.test using build-in function
t.test(pull(Exercise_df_intervention,Systolic_POST), pull(Exercise_df_intervention,Systo
lic_PRE), paired=T, alternative="two.sided")
```

```
##
##  Paired t-test
##
## data:  pull(Exercise_df_intervention, Systolic_POST) and pull(Exercise_df_interventio
n, Systolic_PRE)
## t = -2.9996, df = 35, p-value = 0.004953
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -14.392386  -2.774281
## sample estimates:
## mean of the differences
##               -8.583333
```

- Using a type I error of 0.05, since we have p-value of 0.004953 for the current test, we reject the null hypothesis and conclude there is sufficient evidence to show that there is difference between baseline values and Systolic BP at 6 month for intervention group

**ii) Control group (5p)**

- Null Hypothesis: There is no difference between baseline values and Systolic BP at 6 month for control group.
- Alternative Hypothesis: There is difference between baseline values and Systolic BP at 6 month for control group

```
#creating control group dataframe
Exercise_df_control  =
Exercise_df %>%
  filter(Group==0) %>%
  mutate(difference = Systolic_POST - Systolic_PRE)
```

```
#perform paired-t.test
sd_diff<-sd(pull(Exercise_df_control,difference))
test_weight<-mean(pull(Exercise_df_control,difference))/(sd_diff/sqrt(length(pull(Exercise_df_control,difference))))
#perform paired-t.test using build-in function
t.test(pull(Exercise_df_control,Systolic_POST), pull(Exercise_df_control,Systolic_PRE),
 paired=T, alternative="two.sided")
```

```
##
##  Paired t-test
##
## data:  pull(Exercise_df_control, Systolic_POST) and pull(Exercise_df_control, Systolic_PRE)
## t = -1.3502, df = 35, p-value = 0.1856
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.345373  1.678706
## sample estimates:
## mean of the differences
##               -3.333333
```

- Using a type I error of 0.05, since we have p-value of 0.1856 for the current test, we fail to reject the null hypothesis and conclude there is no sufficient evidence to show that there is difference between baseline values and Systolic BP at 6 month for intervention group

**b) Now perform a test and provide the 95% confidence interval to assess the Systolic BP absolute changes between the two groups. (12p)**

```
#calculate the absolute difference for both group
Exercise_df_control  =
Exercise_df_control %>%
  mutate(abs_difference = abs(Systolic_POST - Systolic_PRE))

Exercise_df_intervention  =
Exercise_df_intervention %>%
  mutate(abs_difference = abs(Systolic_POST - Systolic_PRE))
```

first we need to test for equality of variance

```
# Test the equality of variances
F_stats<-sd(pull(Exercise_df_control,abs_difference))^2/sd(pull(Exercise_df_intervention,abs_difference))^2
F_stats
```

```
## [1] 0.6239293
```

```
#since we know we have 36 observation, n-1=35 n-2=34
F_crit<-qf(.975, df1=35, df2=34)
F_crit
```

```
## [1] 1.974435
```

- when comparing the test statistics,Because F_stats:0.6239293 < F_crit:1.9744352, we fail to reject H0 and conclude that the variances are not significantly different.
- Then we can perform the two-sample t-test with equal variances

```
#X1bar-X2bar
mean_diff=mean(pull(Exercise_df_intervention,abs_difference))-mean(pull(Exercise_df_cont
rol,abs_difference))
#pooled estimate of the variance
s_pool = sqrt(((35*sd(pull(Exercise_df_control,abs_difference))^2)+(35*sd(pull(Exercise_
df_intervention,abs_difference))^2))/(36+36-2))
s_pool
```

```
## [1] 10.46231
```

```
t_crit <- qt(0.975,70)
t_crit
```

```
## [1] 1.994437
```

```
lower=mean_diff-(t_crit*s_pool*sqrt((1/36)+(1/36)))
upper=mean_diff+(t_crit*s_pool*sqrt((1/36)+(1/36)))
lower
```

```
## [1] -1.723818
```

```
upper
```

```
## [1] 8.112706
```

- Therefore, the 95% confidence interval to assess the Systolic BP absolute changes is (-1.7238175,8.1127064), Since the interval contain 0, we fail to reject a null hypothesis and we do not have sufficient evidence to conclude that there is difference in the absolute changes between the two groups.

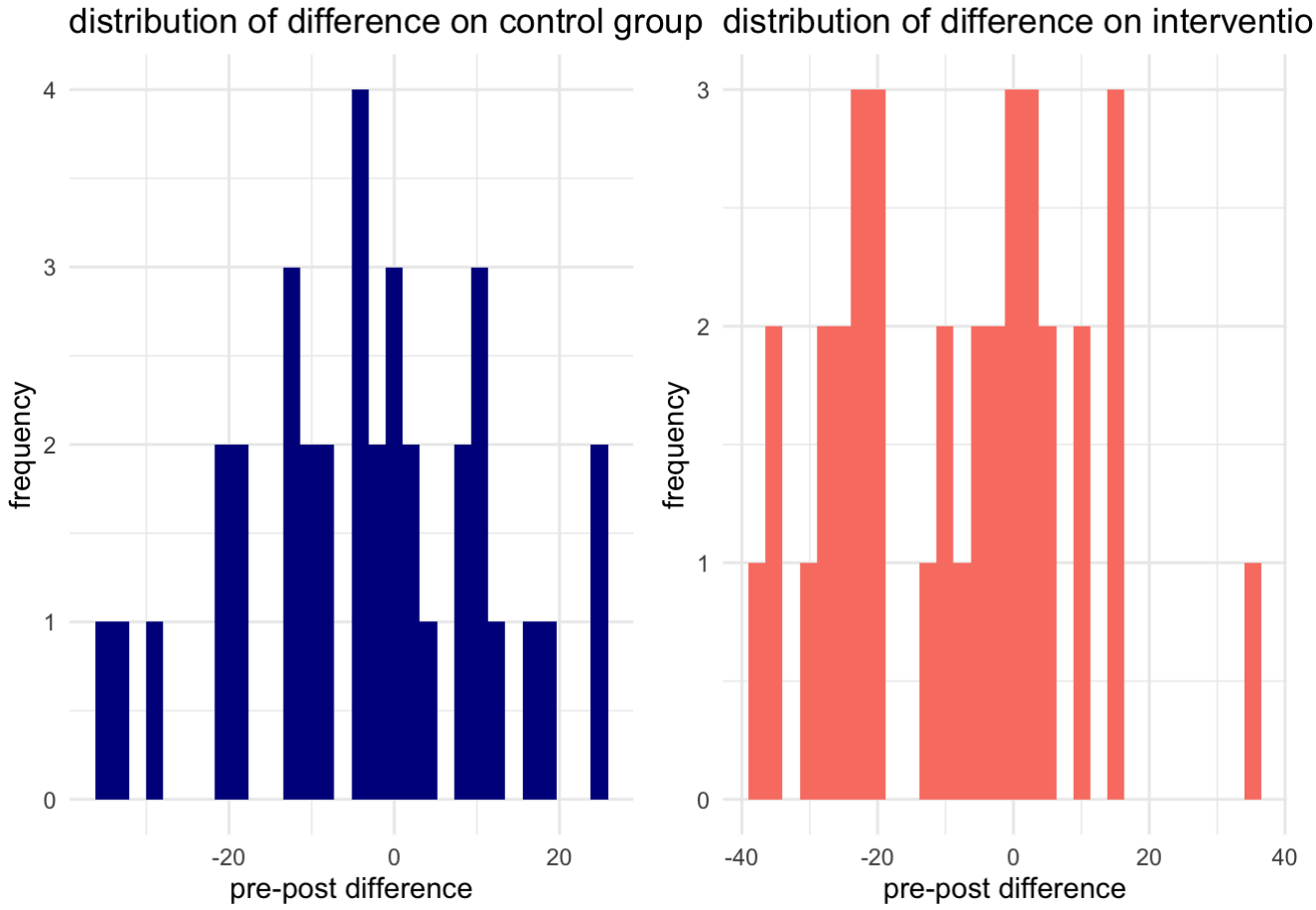
**c) What are the main underlying assumptions for the tests performed in parts a) and b)? (3p)**

- The main assumpetion is that the the observed differences constutue a random sample from a normally distributed population of difference

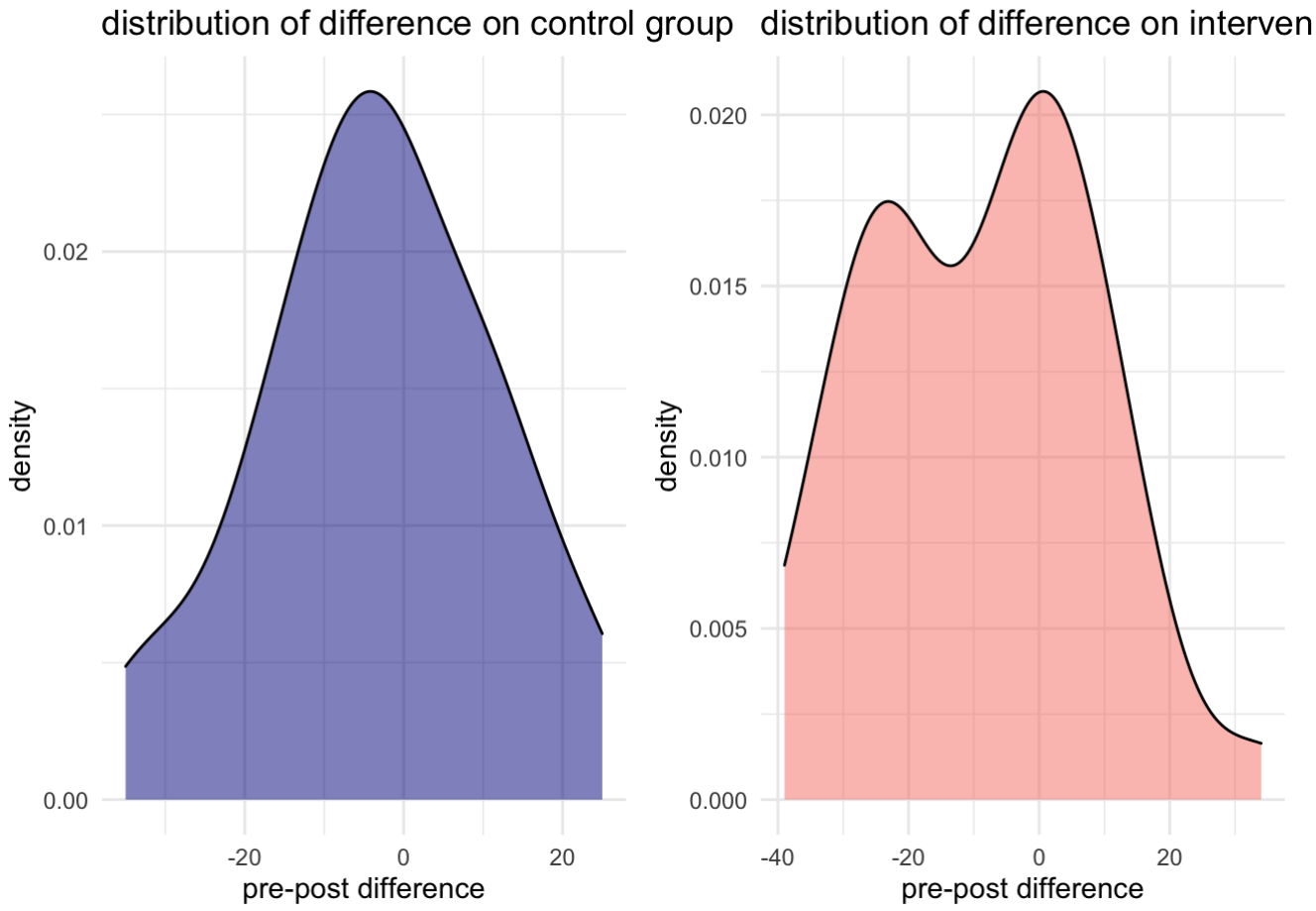
**i) Use graphical displays to check the normality assumption and discuss the findings. (3p)**

```
knitr::opts_chunk$set(
  fig.width = 6,
    fig.asp = .6,
  out.width = "90%"
)
theme_set(theme_minimal()+theme(legend.position = "bottom"))

display_control=ggplot(Exercise_df_control, aes(x=difference)) +
  geom_histogram(fill="dark blue")+
  labs(titles = "distribution of difference on control group",
       x= "pre-post difference",
       y = "frequency"  )
display_intervention=ggplot(Exercise_df_intervention, aes(x=difference)) +
  geom_histogram(fill="salmon") +
labs(titles = "distribution of difference on intervention group",
       x= "pre-post difference",
       y = "frequency"  )

library(patchwork)

display_control+display_intervention
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
display_control=ggplot(Exercise_df_control, aes(x=difference)) +
  geom_density(fill="dark blue",alpha=.5)+
  labs(titles = "distribution of difference on control group",
       x= "pre-post difference",
       y = "density"  )
display_intervention=ggplot(Exercise_df_intervention, aes(x=difference)) +
   geom_density(fill="salmon ",alpha=.5) +
labs(titles = "distribution of difference on intervention group",
       x= "pre-post difference",
       y = "density"  )
display_control+display_intervention
```




- by looking at the two ggplot, the normality assumption seem to hold for control group but not so much for intervention group.(seems like there are two major peak in the distribution of difference in intervention group)

**ii) If normality is questionable, how does this affect the tests validity and what are some possible remedies? (2p)**

- If normality is questionable, our conclusion made based on p-value will also be questionable, because we ware making our inference based on the assumpetion that the difference follow a normal distribution. Some of the remedies can be to increase sample size, because even if the populaiton distribution is not normal, when the sample size is big enough,the sampling distribution will be approximately normal. Also, we can use a more conservative p-value so that we are more cautious about the conclusion we made.

# Problem 2

We have discussed the fact that we are not guaranteed to make the correct decision by the process of hypothesis testing and there is always some level of uncertainty in statistics. The two main errors that we are trying to minimize/control are type I and type II. A type I error occurs when we reject the null hypothesis $H0$, when $H0$ is true. When we set the significance level at 5%, we are saying that we will allow ourselves to make a type I error less than 5% of the time. In practice we can only calculate this probability using a series of "what if" calculations, because we do not really know the truth. In this exercise you learn how to create your own 'true' scenario, simulate corresponding data, and quantify the type I error over many repetitions.

*Scenario: The average IQ score of Ivy League colleges is 120. We will assume this to be the null hypothesis (true mean is 120) with a standard deviation of 15 and a significance level of 5%. For the alternative hypothesis we will consider that the 'true mean is less than 120'.*

Most of the time (95%) when we generate a sample from the underlying true distribution, we should fail to reject the null hypothesis since the null hypothesis is true. Let us test it!

**a) Generate one random sample of size n=20 from the underlying (null) true distribution. Calculate the test statistic, compare to the critical value and report the conclusion: 1, if you reject $H0$ or 0, if you fail to rejected $H0$. (5p) Hint: use rnorm(20, mean = 120, sd = 15)**

```
#generae a randome sample with size=20, mean=120, sd=15
set.seed(2761)
rs1=rnorm(20, mean = 120, sd = 15)

#calculate the criticle value
(mean(rs1)-120)/(15/sqrt(20))
```

```
## [1] 0.6303571
```

```
qnorm(0.05, mean = 0, sd = 1)
```

```
## [1] -1.644854
```

- The conclusion is 0 since our test statistic is 0.63 which is larger than -1.64, and we fail to reject the H0.

**b) Now generate 100 random samples of size n = 20 from the underlying (null) true distribution and repeat the process in part (a) for each sample (calculate the test statistic, compare to the critical value, and record 1 or 0 based on criteria above). Report the percentage of 1s and 0s respectively across the 100 samples. The percentage of 1s represents the type I error. (7.5p)**

```
#creat two empty list
  l <- list()
result=list()

#generae 100 random sample with size=20, mean=120, sd=15 and store them in newly creatd
 list L, then record the decision of each random sample in list "result"
set.seed(2761)
for (i in 1:100){
l[i] <- list(rnorm(20, mean = 120, sd = 15))
}

for (i in 1:100){
result[i] = ifelse((mean(l[[i]])-120)/(15/sqrt(20)) < -1.645,1,0)
}
#convert list into columns
result_df = c(unlist(result))

#show the frequency of each decision
table(result_df)
```

```
## result_df
##  0  1
## 97  3
```

- the percentage of 1s and 0s for the current cases with 100 sample size are 0.03 for 1s and 0.97 for 0s. The precentage of type I error is 0.03 in the current cause.

**c) Now generate 1000 random samples of size n = 20 from the underlying (null) true distribution, repeat the same process, and report the percentage of 1s and 0s across the 1000 samples. (7.5p)**

```
#create two empty list
  l <- list()
result=list()

#generae 1000 random sample with size=20, mean=120, sd=15 and store them in newly creatd
list L, then record the decision of each random sample in list "result"
set.seed(2761)
for (i in 1:1000){
l[i] <- list(rnorm(20, mean = 120, sd = 15))
result[i] = ifelse((mean(l[[i]])-120)/(15/sqrt(20)) < (-1.645),1,0)
}

#convert list into columns
result_df = c(unlist(result))

#show the frequency of each decision
table(result_df)
```

```
## result_df
##   0   1
## 947  53
```

- the percentage of 1s and 0s for the current cases are 0.053 for 1s and 0.947 for 0s. The precentage of type I error is 0.053 in the current cases.*

**d) Final conclusions: compare the type I errors (percentage of 1s) from part b) and c). How do they compare to the level that we initially imposed (i.e. 0.05)? Comment on your findings. (5p)**

- The frequency of Type I error I generated are 0.03 for 100 sample size and 0.053 for 1000 sample size. The type I error approches to the initially imposed type I error as we increase the sample size. From this obervation, my insight is that we need to be extra cautious when making inference for smaller sample size. Because when we are making inference we are assuming the underlying distribution is normal, but there are many natural variation that can confound the result when the sample size is small, however if the sample size is big enough, we can be more confidence when making inferenece.

Notes: For this problem you are encouraged to use R for all calculations/simulations. You can follow the hints or feel free to use other functions – there are several ways to tackle these simulations. You do not need to write the test statistics, critical values, etc., but please include themain results (percentage of correct and incorrect decisions) for each part and conclusions in the main homework document. Make sure to comment your R code and don’t forget to set the seed for replicability.