HW3

Hanyu Lu

October 2020

Problem 1

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------ tidyverse 1.3.0 --
```

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

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

```r
library(arsenal)
library(ggplot2)
library(ggridges)
set.seed(996)
```

```
## Parsed with column specification:
## cols(
##   Group = col_double(),
##   Age = col_double(),
##   Gender = col_double(),
##   Race = col_double(),
##   HTN = col_double(),
##   T2DM = col_double(),
##   Depression = col_double(),
##   Smokes = col_double(),
##   Systolic_PRE = col_double(),
##   Systolic_POST = col_double()
## )
```

```
##
## Table: Descriptive Statistics
##
## |                           |     Overall (N=72)      |
```

```
## |:-------------------------|:----------------------:|
## |Systolic_POST_intervention |                        |
## |-  Mean (SD)              |     125.06 (15.44)     |
## |-  Median (Q1, Q3)        | 124.00 (116.75, 135.00) |
## |Systolic_POST_control     |                        |
## |-  Mean (SD)              |     130.14 (14.35)     |
## |-  Median (Q1, Q3)        | 127.50 (120.00, 140.00) |
## |Systolic_PRE_intervention  |                        |
## |-  Mean (SD)              |     133.64 (15.11)     |
## |-  Median (Q1, Q3)        | 134.00 (121.50, 144.00) |
## |Systolic_PRE_control      |                        |
## |-  Mean (SD)              |     133.47 (15.94)     |
## |-  Median (Q1, Q3)        | 131.00 (122.50, 143.50) |
## |control_difference        |                        |
## |-  Mean (SD)              |      -3.33 (14.81)     |
## |-  Median (Q1, Q3)        |  -3.50 (-12.25, 8.25)   |
## |intervention_difference    |                        |
## |-  Mean (SD)              |      -8.58 (17.17)     |
## |-  Median (Q1, Q3)        |  -5.50 (-23.00, 3.00)   |
```

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:
b) Intervention group (5p)

Since we don't know true population variance. We are going to use paired t-test because we intend to compare scores on two different variables but on the same group. Additionally, we test for the mean of the differences with unknown variance.

H_0: the Systolic BP at 6 months is equal to the baseline values for intervention group H_1: the Systolic BP at 6 months is significantly different from the baseline values for intervention group

$\bar{d} = \sum_{i=1}^{n} d_i/n =$ -8.58 $s_d = \sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2/(n-1)} = 17.17$

$t = \frac{\bar{d}-0}{s_d/\sqrt{n}} = \frac{-8.58-0}{17.17/\sqrt{36}} =$ -3 ~ t_{36-1}

$t_{36-1,0.975} = 2.03$

Since this t-test is two-sided, |t| $= 3 > t_{36-1,0.975} = 2.03$.

We can reject H_0. We can conclude that the Systolic BP at 6 months is significantly different from the baseline values for intervention group

```
## [1] 17.1687
```

```
## [1] -2.998253
```

```
## [1] 2.030108
```

ii) Control group (5p)

Since we don't know true population variance. We are going to use paired t-test because we intend to compare scores on two different variables but on the same group. Additionally, we test for the mean of the differences with unknown variance.

H_0: the Systolic BP at 6 months is equal to the baseline values for control group H_1: the Systolic BP at 6 months is significantly different from the baseline values for control group

$\bar{d} = \sum_{i=1}^{n} d_i/n$ = -3.33 $s_d = \sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2/(n-1)} = 14.81$

$t = \frac{\bar{d}-0}{s_d/\sqrt{n}} = \frac{-3.33-0}{14.81/\sqrt{36}}$ = -1.35 ~ t_{36-1}

$t_{36-1,0.975} = 2.03$

Since this t-test is two-sided, |t| = 1.35 < $t_{36-1,0.975} = 2.03$.

We cannot reject H_0. We can conclude that the Systolic BP at 6 months is not significantly different from the baseline values for intervention group

```
## [1] 14.81312
```

```
## [1] -1.349088
```

```
## [1] 2.030108
```

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

First test the hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$

With significance level α pre-specified, compute the test statistic:

$F = \frac{s_1^2}{s_2^2} = F = \frac{17.17^2}{14.81^2}$ =1.34 ~ $F_{36-1,36-1}$

$F_{36-1,36-1,0.975} = 1.96$

$F_{36-1,36-1,0.025} = 0.51$

$F_{36-1,36-1,0.025} = 0.51$ < F = 1.34 < $F_{36-1,36-1,0.975} = 1.96$

Therefore we are unable to reject H_0 that the variances are equal.

Thus we can conduct two-sample independent t-test with equal variances.

H_0 = The Systolic BP absolute changes between the two groups is equal to 0 H_1 = The Systolic BP absolute changes between the two groups is not equal to 0

$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(36-1)17.17^2+(36-1)14.81^2}{36+36-2}} = 16.03$

$t = \frac{\bar{X}1-\bar{X}2}{s\sqrt{\frac{1}{n1}+\frac{1}{n2}}} = \frac{-8.58-(-3.33)}{16.03\sqrt{\frac{1}{36}+\frac{1}{36}}}$ = -1.39 ~ $t_{36+36-2}$

$t_{36+36-2,0.975} = 1.99$

|t| = 1.39 < $t_{36+36-2,0.975} = 1.99$, so we fail to reject H_0, which means that there is no significant Systolic BP absolute changes between the two groups.

95% confidence interval: $(\bar{X}1-\bar{X}2-t_{n_1+n_2-2,1-\alpha/2}*s*\sqrt{\frac{1}{n1}+\frac{1}{n2}}, \bar{X}1-\bar{X}2+t_{n_1+n_2-2,1-\alpha/2}*s*\sqrt{\frac{1}{n1}+\frac{1}{n2}})$ = (-8.58-(-3.33) - 1.99 * 16.03 * $\sqrt{\frac{1}{36}+\frac{1}{36}}$, -8.58-(-3.33) + 1.99 * 16.03 * $\sqrt{\frac{1}{36}+\frac{1}{36}}$) = (-12.77, 2.27)

```
# F test statistic
17.17^2/(14.81^2)
```

```
## [1] 1.344097
```

```r
qf(0.975,35,35)
```

```
## [1] 1.961089
```

```r
qf(0.025,35,35)
```

```
## [1] 0.5099207
```

```r
#two-sample independent t-test with equal variances

sqrt((35*17.17^2+35*14.81^2)/70) #s
```

```
## [1] 16.03348
```

```r
(-8.58-(-3.33))/(16.03*sqrt(1/36+1/36)) #t
```

```
## [1] -1.389511
```

```r
qt(0.975,70)
```

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

```r
# Confidence Interval

-8.58-(-3.33) - 1.99 * 16.03 * sqrt(1/36+1/36)  #lower
```

```
## [1] -12.76883
```

```r
-8.58-(-3.33) + 1.99 * 16.03 * sqrt(1/36+1/36)  #upper
```

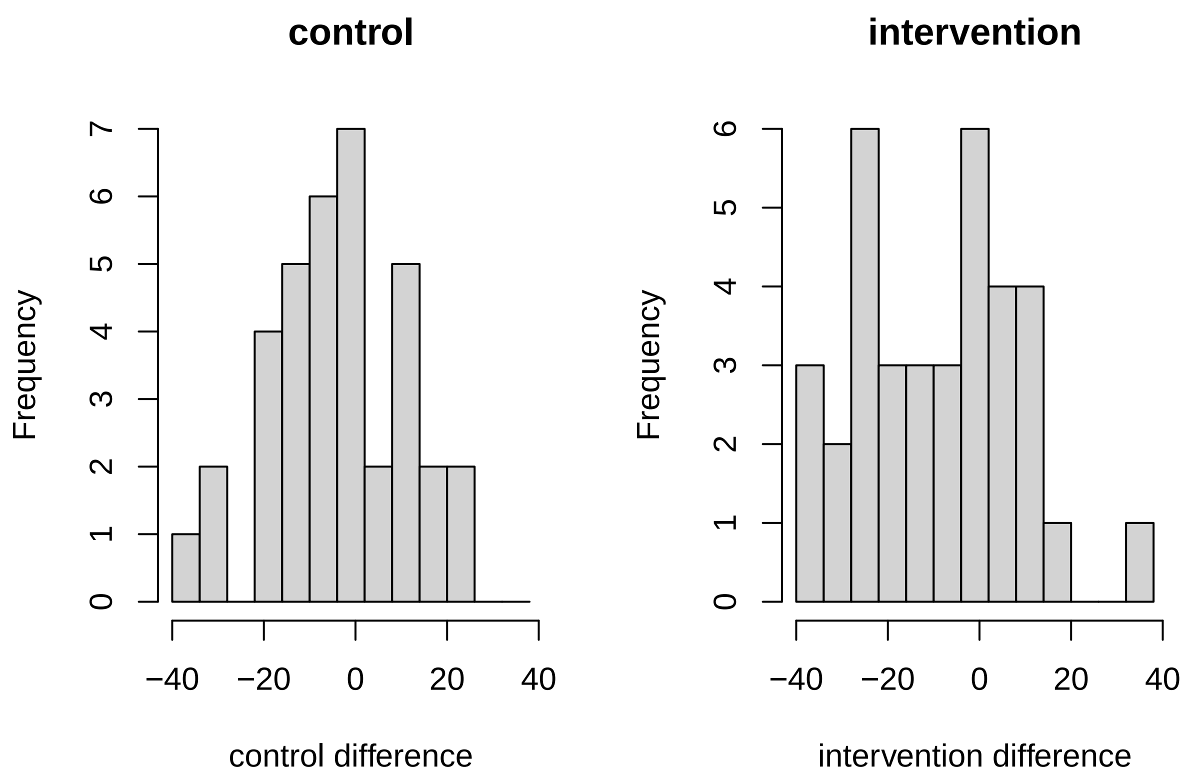
```
## [1] 2.268831
```

c) What are the main underlying assumptions for the tests performed in parts a) and b)?
d) Use graphical displays to check the normality assumption and discuss the findings.

```r
par(mfrow = c(1, 2))

hist(exercise_df2$control_difference, breaks=seq(-40,40,6),xlab="control difference", main ="control")


hist(exercise_df2$intervention_difference, breaks=seq(-40,40,6),xlab="intervention difference", main ="
```



Problem 2

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 0 or 0, if you fail to rejected 0.

H_0 = The average IQ score of Ivy League colleges is equal to 120 H_1 = The average IQ score of Ivy League colleges is less than 120

$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ = -0.89

$z_{0.05}$ = -1.64

$z = -0.89 > z_{0.05} = -1.64$. We fail to reject H_0. We conclude that the average IQ score of Ivy League colleges is equal to 120.

```
random_sample = rnorm(20, mean = 120, sd = 15)

z = (mean(random_sample)-120)/(sd(random_sample)/sqrt(20))

qnorm(0.05)
```

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

```r
print(ifelse(z < -1.64, 1, 0))
```

```
## [1] 0
```

By using the code ifelse(z < -1.64, 1, 0), the output is 1.

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.

```r
sample_means_rep1 = rep(NA, 100)
sample_z1 = rep(NA, 100)
random_sample_2 <- list(mode="vector",length=100)

for(i in 1:100){
  random_sample_2[[i]] = rnorm(20, mean = 120, sd = 15)
  #Generate 100 sample z-scores
  sample_z1[i] = (mean(random_sample_2[[i]])-120)/(sd(random_sample_2[[i]])/sqrt(20))
  sample_means_rep1[i] = ifelse(sample_z1[i] < -1.64, 1, 0)
  #Report 1 or 0
}

count1_1 =sum(sample_means_rep1==1)
count0_1 = sum(sample_means_rep1==0)

print(as.data.frame(table(sample_means_rep1)))
```

```
##   sample_means_rep1 Freq
## 1                 0   94
## 2                 1    6
```

The percentage of 0's in the 100 samples is 94% while the percentage of 1's in the 100 samples is 6%.

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.

```r
sample_means_rep2 = rep(NA, 1000)
sample_z2 = rep(NA, 1000)
random_sample_3 <- list(mode="vector",length=1000)

for(i in 1:1000){
  random_sample_3[[i]] = rnorm(20, mean = 120, sd = 15)
  #Generate 1000 sample z-scores
  sample_z2[i] = (mean(random_sample_3[[i]])-120)/(sd(random_sample_3[[i]])/sqrt(20))
  sample_means_rep2[i] = ifelse(sample_z2[i] < -1.64, 1, 0)
  #Report 1 or 0
}

count1_2 =sum(sample_means_rep2==1)
count0_2 = sum(sample_means_rep2==0)
print(as.data.frame(table(sample_means_rep2)))
```

```
##   sample_means_rep2 Freq
## 1                 0  945
## 2                 1   55
```

The percentage of 0's in the 1000 samples is 94.5% while the percentage of 1's in the 1000 samples is 5.5%.

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.

Type I error, i.e., P(reject H_0 | H_0 is true) of 100 samples is 0.06, while that of 1000 samples is 0.055. Since α level is 0.05, we may infer that as times of sample increase, the type I error of sampling distribution becomes closer to the significance level we imposed.