P8130 HW3

10/16/2020

Problem 1

```r
# data import
exercise = read_csv(file = "./data/Exercise.csv") %>%
  janitor::clean_names() %>%
  select(group, systolic_pre, systolic_post) %>%
  mutate(systolic_diff = systolic_post - systolic_pre) # difference
```

```
## 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()
## )
```

(a) Since we are comparing the changes of Systolic BP for same patients at two different timepoints, We will employ 'Two-Sided Paired t-test' to assess whether the Systolic BP at 6 months is significantly from the baseline values for each of the groups: intervention, control. (assumptions are checked in part (c)) *intervention group*

$H_0 : \mu_{post} - \mu_{pre} = 0$

$H_1 : \mu_{post} - \mu_{pre} \neq 0$

With a pre-specified significance level $\alpha = 0.05$, the test statistics

$t_{stat} = \frac{\bar{d}-0}{s_d/\sqrt{n}}$

Reject H_0: if $|t_{stat}| > t_{n-1,1-\alpha/2}$

Fail to reject H_0: if $|t_{stat}| \leq t_{n-1,1-\alpha/2}$

We find the following:

$\bar{d} = -8.583333$

$s_d = 17.1687$

$n = 36$

$t_{stat} = \frac{\bar{d}-0}{s_d/\sqrt{n}} = \frac{-8.583333-0}{17.1687/\sqrt{36}} \cong -2.999645$

t_crit = `qt(0.975,35)` = 2.030108

- $|t_{stat}| > t_{35,\ \ 0.975}$
- Conclusion: Reject H_0, and conclude that the Systolic BP at 6 month is significant different from the baseline values for the intervention group @ $\alpha = 0.05$.

control group

$H_0 : \mu_{post} - \mu_{pre} = 0$

$H_1 : \mu_{post} - \mu_{pre} \neq 0$

With a pre-specified significance level $\alpha = 0.05$, the test statistics

$t_{stat} = \frac{\bar{d}-0}{s_d/\sqrt{n}}$

Reject H_0: if $|t_{stat}| > t_{n-1,1-\alpha/2}$

Fail to reject H_0: if $|t_{stat}| \leq t_{n-1,1-\alpha/2}$

We find the following:

$\bar{d} = -3.333333$

$s_d = 14.81312$

$n = 36$

$t_{stat} = \frac{\bar{d}-0}{s_d/\sqrt{n}} = \frac{-3.333333-0}{14.81312/\sqrt{36}} \cong -1.350154$

t_crit = `qt(0.975,35)` = 2.030108

- $|t_{stat}| < t_{35,\ \ 0.975}$
- Conclusion: Fail to reject H_0, and conclude that the Systolic BP at 6 month is not significant different from the baseline values for the control group @ $\alpha = 0.05$.

```r
## intervention group:
intervention = exercise %>% filter(group == 1)
mean_diff_int = mean(pull(intervention, systolic_diff)) # -8.583333
sd_diff_int = sd(pull(intervention, systolic_diff)) #17.1687
n_int = nrow(intervention) #36
t_stat_int = (mean_diff_int - 0)/(sd_diff_int/sqrt(n_int)) #-2.999645
t_crit = qt(0.975, 35) #2.030108
## summary:
t.test(intervention$systolic_post, intervention$systolic_pre, paired = T)
```

```
##
##  Paired t-test
##
## data:  intervention$systolic_post and intervention$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
```

```r
## control group:
control = exercise %>% filter(group == 0)
mean_diff_con = mean(control$systolic_diff) # -3.333333
sd_diff_con = sd(control$systolic_diff) #14.81312
n_con = nrow(control) #36
t_stat_con = (mean_diff_con - 0)/(sd_diff_con/sqrt(n_con)) #-1.350154
t_crit = qt(0.975, 35) #2.030108
## summary:
t.test(control$systolic_post, control$systolic_pre, paired = T)
```

```
##
##  Paired t-test
##
## data:  control$systolic_post and control$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
```

(b) Since patients in the two groups are totally different, we will assess the Systolic BP absolute changes between the two groups by using an 'Two-sample Test for Independent samples' Then, the first step is to check whether the underlying variances of the two groups are equal or not. In this case, yes. Test for Equality of Variance

$H_0 : \sigma^2_{intervention} = \sigma^2_{control}$

$H_1 : \sigma^2_{intervention} \neq \sigma^2_{control}$

Compute the F test statistics and critical value: $F_{stat} = \frac{s^2_{intervention}}{s^2_{control}} = \frac{17.1687^2}{14.81312^2} \cong 1.3433268$

$F_{crit} = \texttt{qf(0.975, 35, 35)} = 1.9610894$

- Decision: At 0.05 significance level, because $F_{stat} < F_{crit}$, we fail to reject H_0 and conclude that it is safe to assume variances are equal.

The pooled estimate of the variance from the two independent groups is given by:

$s^2 = \frac{(n_{int}-1)s^2_{int}-(n_{con}-1)s^2_{con}}{n_{int}+n_{con}-2} = \frac{(36-1)17.1687+(36-1)14.81312}{36+36-2} \cong 257.0964$ Where $s = \sqrt{s^2} = 16.03423$

Therefore, we have the following hypothesis test:

$H_0 : \mu_{intervention} = \mu_{control}$

$H_1 : \mu_{intervention} \neq \mu_{control}$

$t_{stat} = \frac{\bar{\mu}_{intervention}-\bar{\mu}_{intervention}}{s/\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$ under the H_0

$t_{stat} = \frac{(-8.583333-(-3.333333))}{16.03423/\sqrt{1/36+1/36}} \cong -1.389145$

t_{crit}= $\texttt{qt(0.975,70)}$= $1.994437 > |t_{stat}|$=1.389145

- Conclusion: Fail to reject H_0, and conclude the difference in mean systolic blood pressure change (post-pre)is not significantly different between the intervention group and control group @ $\alpha = 0.05$.

95% Confidence Interval:

$\overline{X}_{intervention} - \overline{X}_{control} \quad \pm t_{n_{intervention}+n_{control}-2,\ 1-\alpha/2} \times \frac{s}{\sqrt{1/n_1+1/n_2}}$

$(-8.583333 - (-3.333333)) \quad \pm\ 1.995469 \times \sqrt{\frac{257.0964}{2/72}}$

$(-197.1255, 186.6255)$

- Conclusion: we are 95% confident that the true Systolic BP absolute changes for the two groups is between -12.78758 and 2.287583.

```
# variance test #
var.test(intervention$systolic_diff, control$systolic_diff, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  intervention$systolic_diff and control$systolic_diff
## F = 1.3433, num df = 35, denom df = 35, p-value = 0.3869
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6849901 2.6343840
## sample estimates:
## ratio of variances
##           1.343327
```

```
## Alternatively:
# F_stat
F_stat = (sd_diff_int)^2/(sd_diff_con)^2#1.343327
# F_crit
F_crit = qf(0.975,35,35)#1.961089
##################################################################################

# Independent two samples test #
t.test(intervention$systolic_diff, control$systolic_diff, var.equal = TRUE, paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  intervention$systolic_diff and control$systolic_diff
## t = -1.3891, df = 70, p-value = 0.1692
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.787583   2.287583
## sample estimates:
## mean of x mean of y
## -8.583333 -3.333333
```

```
## Alternatively:
# s^2, s
s_2 = (35*sd_diff_int^2 + 35*sd_diff_con^2)/(36 + 36 - 2)#257.0964
s = sqrt(s_2)#16.03423
# test statistics
(mean_diff_int - mean_diff_con)/(s*sqrt(1/36 + 1/36))# |-1.389145| --> 1.389145
```

```
## [1] -1.389145
```

```
# test critical
qt(0.975,70)#1.994437 > test statistics --> fail to reject H0
```

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

```
# 95% C.I
mean_diff_int - mean_diff_con #-5.25
```

```
## [1] -5.25
```

```
lower_bound = (mean_diff_int - mean_diff_con) - qt(0.975,70)*s*sqrt(2/36)#-12.78758
upper_bound = (mean_diff_int - mean_diff_con) + qt(0.975,70)*s*sqrt(2/36)#2.287583
```

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

- By CLT if you are sampled from a underlying normal distribution, the sample mean also follows a normal distribution, and the difference if the sample means are also follows a normal distribution.
- For part a), we use a paired t-test, and we assume the SBP measurements(post and pre) for both intervention and control groups are normally distributed with mean $\mu_{int_{post}} \quad and \quad \mu_{int_{pre}}$, and $\mu_{con_{post}} \quad and \quad \mu_{con_{pre}}$, respectively. If follows that the differences $d_{i_{int}} \sim N(\Delta_{int},\ \sigma^2_{d_{int}})$, $d_{i_{con}} \sim N(\Delta_{con},\ \sigma^2_{d_{con}})$, where i=1,2,3,...n. Under the null hypothesis, the test statistics follows a t_{n-1}.
- For part b), we consider the the two groups are independent samples, and that they are normally distributed: $X_{intervention_\Delta} \sim N(\mu_{int_\Delta},\sigma^2_{int_\Delta})$ and $X_{control_\Delta} \sim N(\mu_{int_\Delta},\sigma^2_{int_\Delta})$. And we found the underlying variances of the two samples are equal.

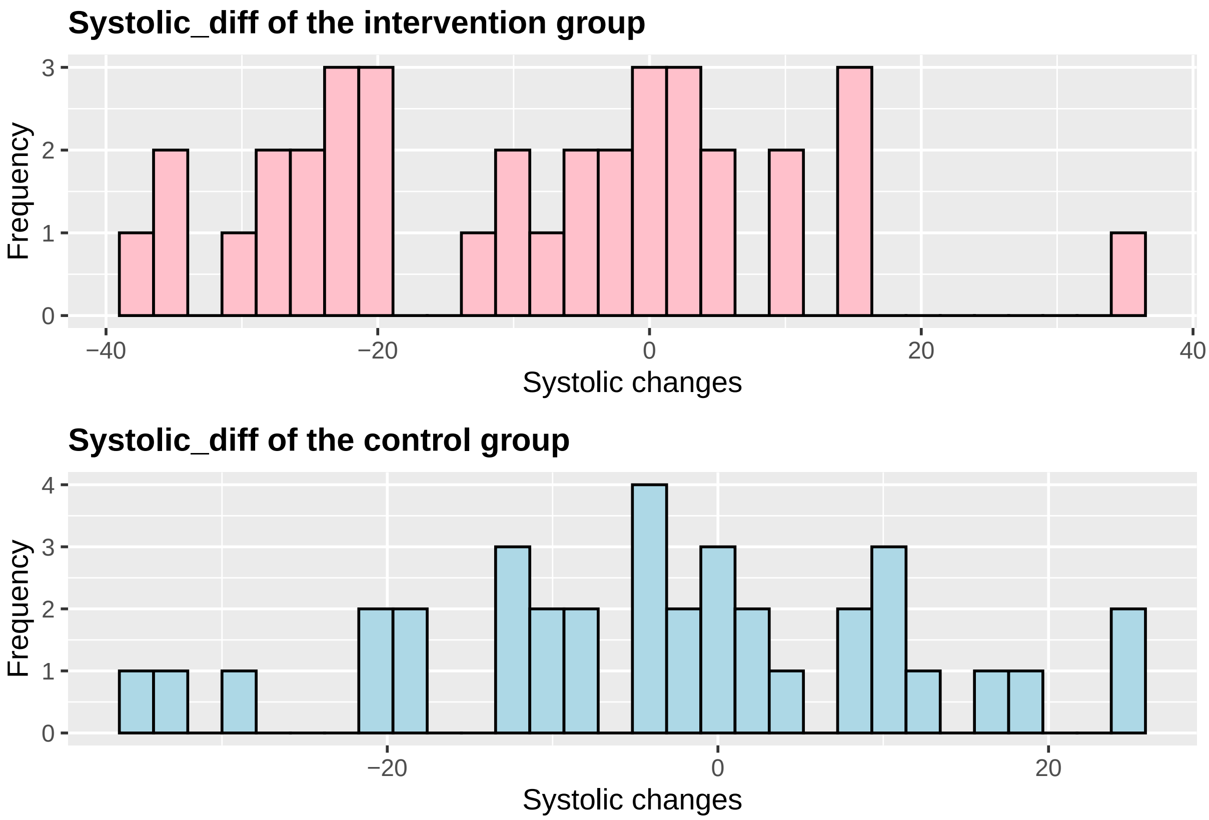
As plotted below, the distribution of the control group looks like pretty much a normal distribution with some sort of symmetric bell-shape. The distribution of the intervention group is slightly off with more than one peaks but that's fine because t-test are pretty robust to some deviation of normality. Even if the distribution is not quite symmetric, the t-test will still generate valid results. Therefore, we don't need to doubt the normality assumption.

Possible methods to reinforce the normality assumption would be increase the total sample sizes for both groups. Therefore, with a large sample size, the normality assumption holds approximately due to CLT.

```
#normality_intervention_diff
int_diff =
  ggplot(intervention, aes(x = systolic_diff)) +
  geom_histogram(color = "black", fill = "pink", bins = 30) +
  labs(
    title = "Systolic_diff of the intervention group",
    x = "Systolic changes",
    y = "Frequency") +
  theme(
    plot.title = element_text(lineheight = 3, face = "bold", size = 12)
  )

#normality_control_diff
con_diff =
  ggplot(control, aes(x = systolic_diff)) +
  geom_histogram(color = "black", fill = "light blue", bins = 30) +
  labs(
    title = "Systolic_diff of the control group",
    x = "Systolic changes",
    y = "Frequency") +
  theme(
    plot.title = element_text(lineheight = 3, face = "bold", size = 12)
  )

(int_diff)/(con_diff)
```



Problem 2

One_tailed Hypothesis test:

$H_0 : \mu = \mu_0 = 120$ v.s. $H_1 : \mu < 120$

Where $\sigma = 15, \quad \alpha = 0.05$

$Z_{stat} = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$

$Z_{crit} =` qnorm(0.05)` \cong -1.65$

Reject $H_0 : if \quad Z_{stat} < Z_{crit}$ Fail to reject $H_0 : if \quad Z_{stat} \geq Z_{crit}$

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

(a) one random sample from the underlying (null) true distribution We find: |Z_stat| = 0.01776018 > Z_crit = -1.644854, fail to reject H0, and conclude that the true mean of average IQ score of IVY League is 270. The conclusion is 1 in this one random sample case.

```r
set.seed(88)
sample_1 = rnorm(20,120,15)
Z_stat = (mean(sample_1) - 120) / (15*sqrt(20));Z_stat
```

```
## [1] 0.01776018
```

```r
Z_crit = qnorm(0.05);Z_crit
```

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

(b) 100 random samples from the underlying (null) true distribution We find the percentage of 1s to be 0.02, and percentage of 0s to be 0.98. Among the 100 random samples, 2% of them reject the H0, 98% of them fail to reject H0. This is what we would expect to see since they are sampled from the underlying normal distribution with alpha = 0.05. The Type I error is 2%, which is less than our default alpha = 0.05.

```r
set.seed(88)
Z_stat = rep(NA, 100)
decision = rep(NA,100)

for (i in 1:100) {
  Z_stat[i] = (mean(rnorm(20, 120, 15) - 120) / (15/sqrt(20)))
  Z_crit = qnorm(0.05)
  if (Z_stat[i] < Z_crit) {decision[i] = 1}
  else(decision[i] = 0)

}

reject = 0
fail_to_reject = 0

for (i in 1:100) {
   if (decision[i] == 1) (reject = reject + 1)
   if (decision[i] == 0) (fail_to_reject = fail_to_reject + 1)
}

percent_of_reject = reject/100; percent_of_reject
```

```
## [1] 0.06
```

```r
percent_of_fail_to_reject = fail_to_reject/100; percent_of_fail_to_reject
```

```
## [1] 0.94
```

(c) 1000 random samples from the underlying (null) true distribution We find the percentage of 1s to be 0.045, and percentage of 0s to be 0.955. Among the 100 random samples, 4.5% of them reject the H0, 95.5% of them fail to reject H0. This is what we would expect to see since they are sampled from the underlying normal distribution with alpha = 0.05. The Type I error is 4.5%, which is still less than our default alpha = 0.05. However, as we increase the number of random samples, we find the Type I error (0.045) is much more closer to the default alpha = 0.05.

```r
set.seed(88)
Z_stat = rep(NA, 1000)
decision = rep(NA,1000)

for (i in 1:1000) {
  Z_stat[i] = (mean(rnorm(20, 120, 15) - 120) / (15/sqrt(20)))
  Z_crit = qnorm(0.05)
  if (Z_stat[i] < Z_crit) {decision[i] = 1}
  else(decision[i] = 0)

}

reject = 0
fail_to_reject = 0

for (i in 1:1000) {
   if (decision[i] == 1) (reject = reject + 1)
   if (decision[i] == 0) (fail_to_reject = fail_to_reject + 1)
}

percent_of_reject = reject/1000; percent_of_reject
```

```
## [1] 0.045
```

```r
percent_of_fail_to_reject = fail_to_reject/1000; percent_of_fail_to_reject
```

```
## [1] 0.955
```

(d) In part(b), we get a Type I error of 0.06, in part (c), we get a Type I error of 0.045. As we increase the number of random samples from the underlying true distribution, we find the Type I error are more likely to get closer to our default setting alpha of 0.05.