p8130_hw3_ps3194

Pangsibo Shen

10/20/2020

Contents

```
library(tidyverse)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.3
```

Problem 1

```
exercise = read_csv("./data/Exercise.csv") %>%
  janitor::clean_names()
```

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 solution: we are using two-sided paired t-test μ_b stands for the mean systolic BP for baseline group and μ_a stands for the mean systolic BP for 6 month group. The null and alternative hypotheeses are:

$$H_0 : \mu_a - \mu_b = 0$$

$$H_1 : \mu_a - \mu_b \neq 0$$

with a pre-specified significance level $\alpha = 0.05$, compute the test statistics:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}, DF : t_{n-1}$$

where $\bar{d}$ stands for mean difference from sample and s_d stands for standard deviation of differences.

```
intervention = exercise %>%
  filter(group == 1)

systolic_pre_1 = pull(intervention, systolic_pre) #baseline systolic data for intervention group

systolic_post_1 = pull(intervention, systolic_post) #6 month systolic data for intervention group

systolic_diff_1 = systolic_post_1 - systolic_pre_1 # difference in systolic BP
```

```r
sd_diff_1 = sd(systolic_diff_1) # standard deviation for difference in systolic BP

test_systolic_1 = mean(systolic_diff_1)/(sd_diff_1/sqrt(length(systolic_diff_1)))

#t.test(systolic_post_1, systolic_pre_1 , paired = T, alternative = "two.sided")
```

```r
abs(test_systolic_1) # absolute value of t statistic
```

```
## [1] 2.999645
```

```r
qt(1 - 0.05/2,length(systolic_diff_1) - 1) # critical value
```

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

$$|t| > t_{n-1,1-\alpha/2}$$

the absolute t statistic equals 2.9996446 which is greater the critical value 2.0301079. Hence we reject the null hypothesis H_0 which means that the Systolic BP at 6 months is significantly different from the baseline values for intervention group.

ii) Control group solution: we are using two-sided paired t-test μ_b stands for the mean systolic BP for baseline group and μ_a stands for the mean systolic BP for 6 month group. The null and alternative hypotheeses are:

$$H_0 : \mu_a - \mu_b = 0$$

$$H_1 : \mu_a - \mu_b \neq 0$$

with a pre-specified significance level $\alpha = 0.05$, compute the test statistics:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}}, DF : t_{n-1}$$

where $\bar{d}$ stands for mean difference from sample and s_d stands for standard deviation of differences.

```r
control = exercise %>%
  filter(group == 0)

systolic_pre_0 = pull(control, systolic_pre) #baseline systolic data for control group

systolic_post_0 = pull(control, systolic_post) #6 month systolic data for control group

systolic_diff_0 = systolic_post_0 - systolic_pre_0 # difference in systolic BP

sd_diff_0 = sd(systolic_diff_0) # standard deviation for difference in systolic BP

test_systolic_0 = mean(systolic_diff_0)/(sd_diff_0/sqrt(length(systolic_diff_0)))

#t.test(systolic_post_0, systolic_pre_0 , paired = T, alternative = "two.sided")
```

```r
abs(test_systolic_0) # absolute value of t statistic
```

```
## [1] 1.350154
```

```r
qt(1 - 0.05/2,length(systolic_diff_0) - 1) # critical value
```

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

$$|t| < t_{n-1,1-\alpha/2}$$

the absolute t statistic equals 1.3501543 which is less the critical value 2.0301079. Hence we fail to reject null hypothesis H_0 which means that there is no difference between systolic BP at 6 month and at baseline for control group.

---

b) Now perform a test and provide the 95% confidence interval to assess the Systolic BP absolute changes between the two groups. solution: We are going to use two-sample independent t-test for the problem. First, we need to test for equality of variances.

$$H_0 : \sigma_c^2 - \sigma_i^2 = 0$$

$$H_1 : \sigma_c^2 - \sigma_i^2 \neq 0$$

with a pre-specified significance level $\alpha = 0.05$, compute the test statistics:

$$F = \frac{s_c^2}{s_i^2}, DF : F_{n_c-1,n_i-1}$$

where s_c^2 stands for sample variance for control group and s_i^2 stands for sample variance for intervention group.

```r
# Testing equality of variances for two independent samples
# drawn from two underlying normal distributions.

# control(1): s1=14.81, n1=36, x1_bar=-3.33
# intervention(2): s2=17.17, n2=36, x2_bar=-8.58

F_test = 14.81^2/17.17^2
F_test
```

```
## [1] 0.7439942
```

```r
F_crit = qf(.975, df1=35, df2=35)
F_crit
```

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

$$F < F_{n_c-1,n_i-1,1-\alpha/2}$$

F statistic equals 0.7439942 which is less the critical value 1.9610894. Hence we fail to reject the null hypothesis H_0 which means that there is no significant difference in pop. variance between control and intervention group.Therefore, We are going to use two-sample independent t-test with equal variances.

$$H_0 : \mu_c = \mu_i$$

$$H_1 : \mu_c \neq \mu_i$$

with a pre-specified significance level $\alpha = 0.05$, compute the test statistics:

$$t = \frac{\bar{x_c} - \bar{x_i}}{s\sqrt{\frac{1}{n_c} + \frac{1}{n_i}}}, DF : t_{n_c+n_i-2}$$

where $\bar{x_c}$ stands for Systolic BP absolute changes for control group and $\bar{x_i}$ stands for Systolic BP absolute changes for intervention group.s stands for pooled estimate of standard deviation which can be calculated by the formula below.

$$s = \sqrt{\frac{(n_c-1)s_c^2 + (n_i-1)s_i^2}{n_c+n_i-2}}$$

```
std_pooled = sqrt(((14.81^2*35)+(17.17^2*35))/70)

t_stats = (-3.33+8.58)/(std_pooled*sqrt((1/36)+(1/36)))
t_stats
```

```
## [1] 1.38921
```

```
# Compare t_stats to the critical value: t with 18 df

qt(0.975,70) # 2.10
```

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

$$|t| < t_{n_c+n_i-2,1-\alpha/2}$$

t statistic equals 1.3892095 which is less the critical value 1.9944371. Hence we fail to reject the null hypothesis H_0 which means that there is no significant difference in the Systolic BP absolute changes between two groups.

The formula for two sided 95% confidence interval for two-independent samples with equal variance:

$$(\bar{x_c} - \bar{x_i} - t_{n_c+n_i-2,1-\alpha/2} * s * \sqrt{\frac{1}{n_c} + \frac{1}{n_i}}, \bar{x_c} - \bar{x_i} + t_{n_c+n_i-2,1-\alpha/2} * s * \sqrt{\frac{1}{n_c} + \frac{1}{n_i}})$$

```
CI_lower = -3.33+8.58 - qt(0.975,70)*std_pooled*sqrt((1/36) + (1/36))
CI_lower
```

```
## [1] -2.287232
```

```r
CI_upper = -3.33+8.58 + qt(0.975,70)*std_pooled*sqrt((1/36) + (1/36))
CI_upper
```

```
## [1] 12.78723
```

Hence the two sided 95% confidence interval is (-2.2872324,12.7872324)

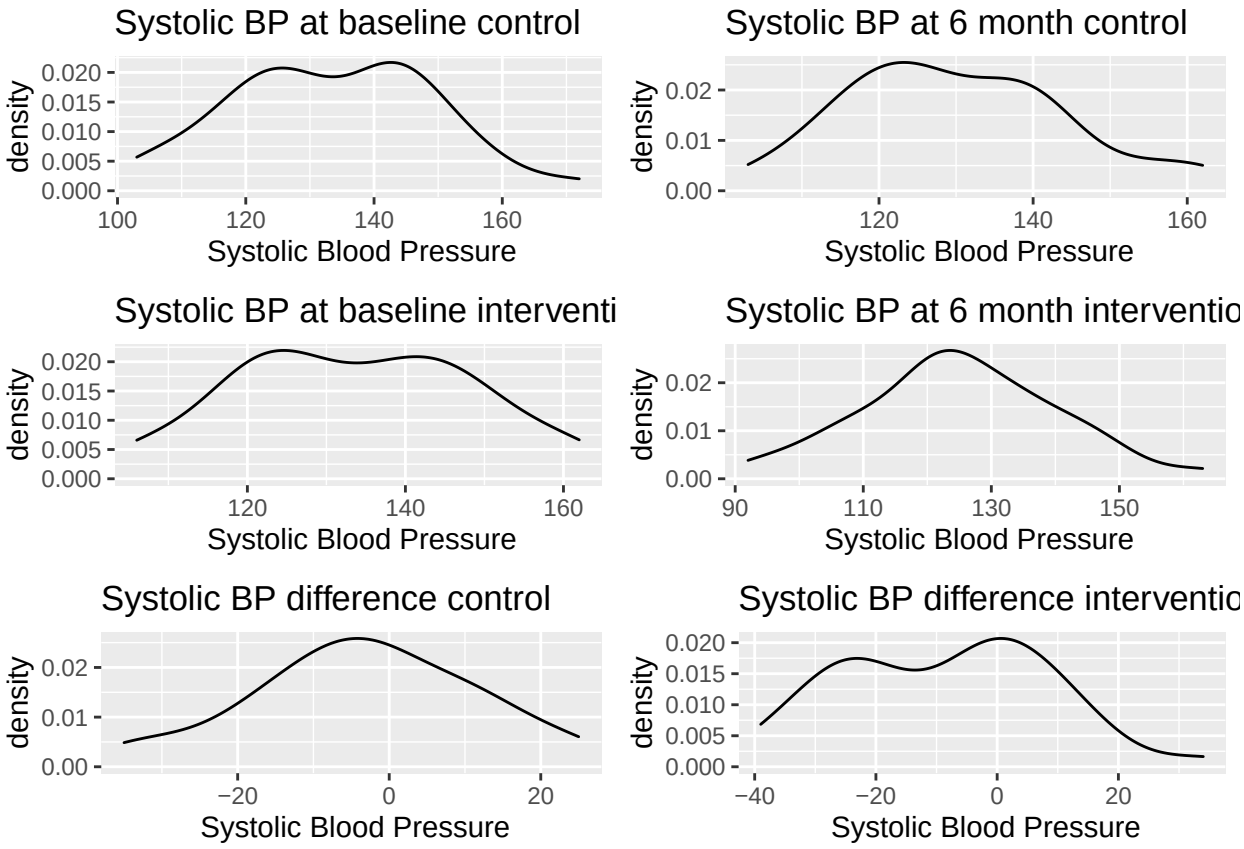
---

c) What are the main underlying assumptions for the tests performed in parts a) and b)? solution: The underlying assumption for parts a and b is that two samples are normally distributed. Also, we assume that two samples are independent for part b.

i)) Use graphical displays to check the normality assumption and discuss the findings. solution:

```r
plot_control_pre = control %>%
  ggplot(aes(x = systolic_pre)) +
  geom_density() +
  ggtitle("Systolic BP at baseline control") +
  labs(x = " Systolic Blood Pressure")


plot_control_post = control %>%
  ggplot(aes(x = systolic_post)) +
  geom_density() +
  ggtitle("Systolic BP at 6 month control") +
  labs(x = " Systolic Blood Pressure")

plot_control_diff = control %>%
  mutate(systolic_diff = systolic_post - systolic_pre) %>%
  ggplot(aes(x = systolic_diff)) +
  geom_density() +
  ggtitle("Systolic BP difference control") +
  labs(x = " Systolic Blood Pressure")


plot_intervention_pre = intervention %>%
  ggplot(aes(x = systolic_pre)) +
  geom_density() +
  ggtitle("Systolic BP at baseline intervention") +
  labs(x = " Systolic Blood Pressure")

plot_intervention_post = intervention %>%
  ggplot(aes(x = systolic_post)) +
  geom_density() +
  ggtitle("Systolic BP at 6 month intervention") +
  labs(x = " Systolic Blood Pressure")

plot_intervention_diff = intervention %>%
  mutate(systolic_diff = systolic_post - systolic_pre) %>%
  ggplot(aes(x = systolic_diff)) +
```

```
  geom_density() +
  ggtitle("Systolic BP difference intervention") +
  labs(x = " Systolic Blood Pressure")

ggarrange(plot_control_pre,plot_control_post,plot_intervention_pre,plot_intervention_post,plot_control_
```



For the distribution of Systolic BP at baseline for control group, there are two peaks in the center. For the distribution of Systolic BP at 6 month for control group, there are two peaks in the center and the first peak is higher than the second peak. For the distribution of Systolic BP at baseline for intervention group, there are two peaks in the center and the first peak is higher than the second peak. For the distribution of Systolic BP at 6 month for intervention group, the graph is fairly normal distributed. For the distribution of Systolic BP difference for control group, the graph is fairly normal distributed. For the distribution of Systolic BP difference for intervention group, there are two peaks in the center and the second peak is higher than the first peak and the curve is right - skewed.

ii) If normality is questionable, how does this affect the tests validity and what are some possible remedies? solution: We often assume that independent t-test is robust against violation of the normality assumption, especially when the sample size is greater than 30.We can still be fairly confidence about our test results. As for possible remedies, we can increase the sample size for both group. With larger sample sizes, the distributions will approximate toward normal distribution. Furthermore, we could also use non parametric t-test which does not assume anything about the underlying distribution.

Problem 2

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

where μ_o is the average IQ score of Ivy League colleges which is 120. μ stands for the true mean.The standard deviation is 15 with a pre-specified significance level $\alpha = 0.05$.

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 H_0 or 0, if you fail to rejected H_0. solution:

```
set.seed(123)

sample1  = rnorm(20, mean = 120, sd = 15)

z_test_left = function(sample, mu, sd, cr_z){
  z = (mean(sample) - mu)/(sd/sqrt(length(sample)))

  if(z > cr_z){
    print(1) #print 1 when we reject null hypothesis
  } else {
    print(0) #print 0 when we fail to reject null hypothesis
  }
  print(z) #print the z statistic
}

z_test_left(sample1, 120, 15, qnorm(1 - 0.05))
```

```
## [1] 0
## [1] 0.6333609
```

In this case, z statistic (0.633) is less than the critical value (1.64) hence we fail to reject the null hypothesis. The result is 0.

---

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. solution:

In order to stimulate this model 100 times, I first write a function called z_test_repeat which takes four inputs: the sample, the true mean μ, standard deviation, and critical value for $\alpha = 5$. z_test_repeat function will return 1 when we reject null hypothesis and 0 when we fail to reject null hypothesis. Then, we are going to iterate the z_test_repeat function 100 times and record each output into an empty list. The percentage of 1's can be calculated by taking the average of the list.

```r
set.seed(321)

z_test_repeat = function(sample, mu, sd, cr_z){
  z = (mean(sample) - mu)/(sd/sqrt(length(sample)))
  if(z > cr_z){
    return(1) #return 1 when we reject null hypothesis
  } else {
    return(0) #return 0 when we fail to reject null hypothesis
  }
}

list_z_100 = list() #initialize an empty list

# using while loop to iterate 100 times
i = 1
while (i <= 100) {
  list_z_100[i] =
    z_test_repeat(rnorm(20, mean = 120, sd = 15), 120, 15, qnorm(1 - 0.05)) #add 1s and 0s into the emp
  i = i + 1
}

mean(as.numeric(list_z_100))*100 #type I error percentage
```

```
## [1] 3
```

After we repeat the process in part a for 100 times, the percentage of 1s is 3% and the percentage of 0s is 97%. Hence the type I error is 3%.

---

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.
solution:

In order to stimulate the model 1000 times,we can use z_test_repeat function from part b and iterate it 1000 times and record each output into an empty list. The percentage of 1's can be calculated by taking the average of the list.

```r
set.seed(3)

list_z_1000 = list() #initialize an empty list

# using while loop to iterate 1000 times
i = 1
while (i <= 1000) {
  list_z_1000[i] =
    z_test_repeat(rnorm(20, mean = 120, sd = 15), 120, 15, qnorm(1 - 0.05)) #add 1s and 0s into the emp
  i = i + 1
}

mean(as.numeric(list_z_1000))*100 #type I error percentage
```

```
## [1] 4.1
```

After we repeat the process in part a for 1000 times, the percentage of 1s is 4.1% and the percentage of 0s is 95.9%. Hence the type I error is 4.1%.

**** 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.**** solution:

From part b with 100 trials, we had type I error of 3% and from part c with 1000 trials, we had type I error of 4.1%. As we increased the number of repeated trials (from 100 times to 1000 times), we got a closer approximate of type I error to the initially imposed value of 5%.