

P8130_hw3_zl2974

Jeffrey Liang

10/13/2020

Problem 1 (30p)

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

##				
##				
##	month		0 (N=72)	1 (N=72)
##	:-----	:-----	:-----	:-----
##	6_month	systolic		
##		- Mean (SD)	130.139 (14.345)	125.056 (15.444)
##		- Median (Q1, Q3)	127.500 (120.000, 140.000)	124.000 (116.750, 135.000)
##		- Min - Max	103.000 - 162.000	92.000 - 163.000
##		systolic_diff		
##		- Mean (SD)	3.333 (14.813)	8.583 (17.169)
##		- Median (Q1, Q3)	3.500 (-8.250, 12.250)	5.500 (-3.000, 23.000)
##		- Min - Max	-25.000 - 35.000	-34.000 - 39.000
##	baseline	systolic		
##		- Mean (SD)	133.472 (15.938)	133.639 (15.108)
##		- Median (Q1, Q3)	131.000 (122.500, 143.500)	134.000 (121.500, 144.000)
##		- Min - Max	103.000 - 172.000	106.000 - 162.000
##		systolic_diff		
##		- Mean (SD)	3.333 (14.813)	8.583 (17.169)
##		- Median (Q1, Q3)	3.500 (-8.250, 12.250)	5.500 (-3.000, 23.000)
##		- Min - Max	-25.000 - 35.000	-34.000 - 39.000

- Perform appropriate tests to assess if the Systolic BP at 6 months is significantly different from the baseline values for each of the groups:
 - Intervention group (5p)
 - Control group (5p)
- Now perform a test and provide the 95% confidence interval to assess the Systolic BP absolute changes between the two groups. (12p)
- What are the main underlying assumptions for the tests performed in parts a) and b)? (3p)

- i) Use graphical displays to check the normality assumption and discuss the findings. (3p)
- ii) If normality is questionable, how does this affect the tests validity and what are some possible remedies? (2p)

PROOF

- 1) We use paired test for testing systolic BP difference within group of two measure point. Because the population variance is unknown, and we have no prior knowledge of how's systolic BP going to change in 6-month period so two-side paired t-test is used for this problem

a.

$$\begin{aligned}\bar{d} &= \frac{\sum_{i=1}^n d_i}{n} = 8.583 \\ se &= \sqrt{\sum_{i=1}^n (d_i - \bar{d})^2 / (n - 1)} = 2.861 \\ t &= \frac{\bar{d} - 0}{se} = 3 \sim t_{36-1}\end{aligned}$$

With critical value $t_{36-1, 1-\alpha/2} = 2.03$, we reject the Null hypothesis and conclude that there's difference between baseline systolic BP value and 6-month later in intervention group.

a.

$$\begin{aligned}\bar{d} &= \frac{\sum_{i=1}^n d_i}{n} = 3.333 \\ se &= \sqrt{\sum_{i=1}^n (d_i - \bar{d})^2 / (n - 1)} = 2.469 \\ t &= \frac{\bar{d} - 0}{se} = 1.35 \sim t_{36-1}\end{aligned}$$

With critical value $t_{36-1, 1-\alpha/2} = 2.03$, we can not reject the Null hypothesis and conclude that there's no difference between baseline systolic BP value and 6-month later in control groups.

- 2) Now that we compare two groups, and the population variance is unknown, we first test if the standard deviation of two groups is equal:

$$\begin{aligned}s_{d_1} &= \sum_{i=1}^{n_1} (d_i - \bar{d}_1)^2 / (n_1 - 1) = 17.169 \\ s_{d_0} &= \sum_{j=1}^{n_0} (d_j - \bar{d}_0)^2 / (n_0 - 1) = 14.813 \\ F &= s_1^2 / s_0^2 = 1.343 \sim F_{n_1-1, n_0-1}\end{aligned}$$

With F-test critical value of 1.961, we cannot reject the null hypothesis that two sample variance is different. And we have no idea how's the intervention going to change the systolic value between groups, so two-side equal variance t-test is used.

$$\bar{X}_1 - \bar{X}_0 = 8.583 - 3.333 = 5.25$$

$$\begin{aligned} s_{pool} &= \frac{(n_1 - 1)s_1 + (n_0 - 1)s_0}{n_1 + n_0 - 2} \\ &= \frac{(36 - 1) * 17.1687007578991 + (36 - 1) * 14.8131215963608}{36 + 36 - 2} \\ &= 15.991 \end{aligned}$$

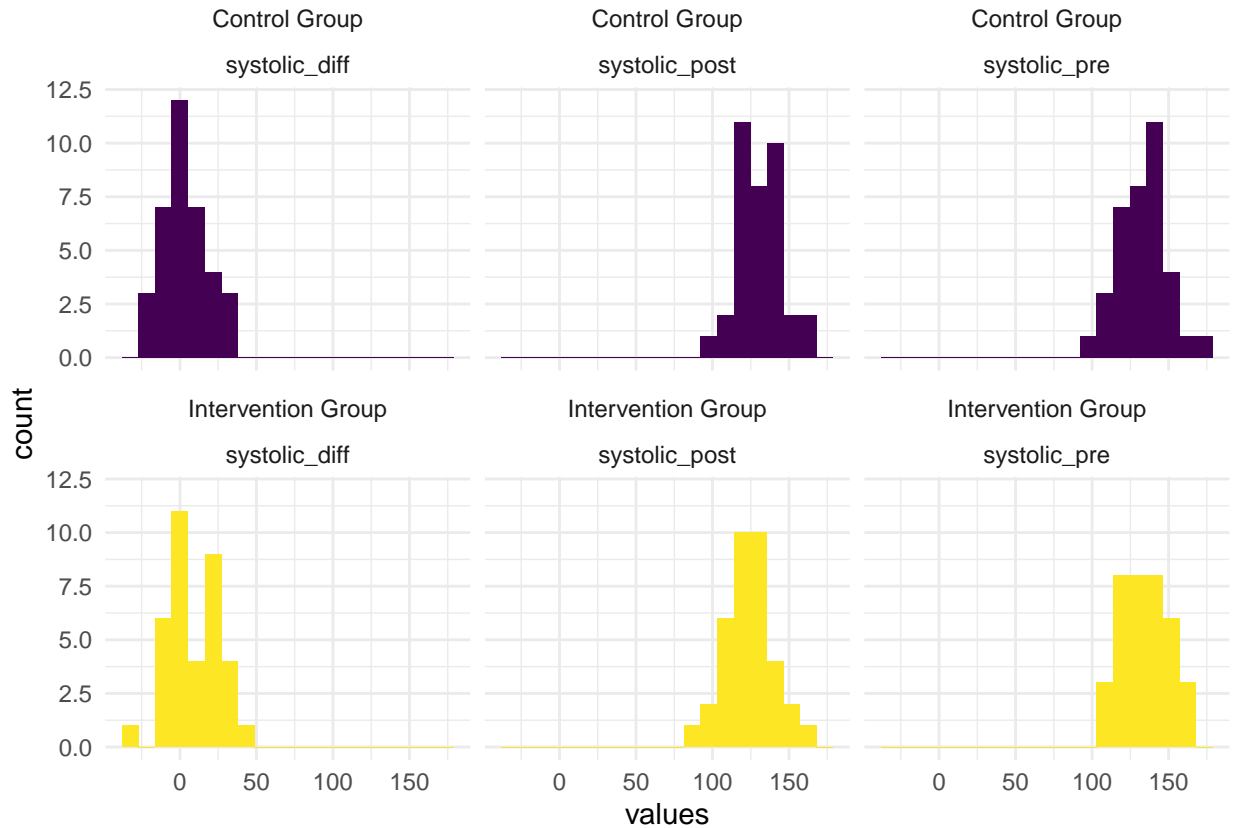
$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_0}{s_{pool} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \\ &= 1.389 \sim t_{36+36-2} \end{aligned}$$

According to the critical value we choose(1.994), we cannot reject H_0 that there's difference between the difference of baseline systolic BP and 6 month later in intervention group and control group.

- 1) The testing in 1),2) are based on that sample distributions fit normality. As plotted as follow, the sample distribution statisfied normality requirement for testing.

But if the sample size is skewed, there're:

- By central limit theorem, if our sample size is large enough, we will have a sample distribution approximate normal distribution
- Use normal transformation: we have methods(eg. logarithm or square root) to transform our sample distribution to approximate normal.
- Use bayesian inference or other statistics inference methods that don't rely on Normality assumption of the distribution.



Problem 2 (25p)

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 H_0 , when H_0 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!

- 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 reject H_0 . (5p) Hint: use `rnorm(20, mean = 120, sd = 15)`
- 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) Suggestions: use a for loop to loop over the 100 samples and create a variable using the function `ifelse()` to keep track of your 1’s and 0’s.

- 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)
- 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)

```
## Reject_H0 n percent
##      FALSE 1      1

## Reject_H0 n percent
##      FALSE 92    0.92
##      TRUE  8     0.08

## Reject_H0 n percent
##      FALSE 951   0.951
##      TRUE  49    0.049
```