P8130 Fall 2020: Biostatistical Methods I

Homework 3

Due, Oct 21 @ 5:00pm

P8130 Guidelines for Submitting Homework

Your homework should be submitted only through CourseWorks. No email submissions!

All derivations, graphs, output and interpretations to each section of the problem(s) must be included in the PDF (not the code), otherwise it will not be graded.

Only 1 PDF file should be submitted. When derivations were required and handwriting was allowed, scan the derivations and merge ALL PDF files (http://www.pdfmerge.com/) into a single one.

We are encouraged to use R for calculations, but you still have to show the mathematical formulae. Also, make sure to also submit your commented code as a separate R/RMD file.

DO NOT FORGET:

You are encouraged to collectively look for answers, explain things to each other, and use questions to test each other's knowledge.

But

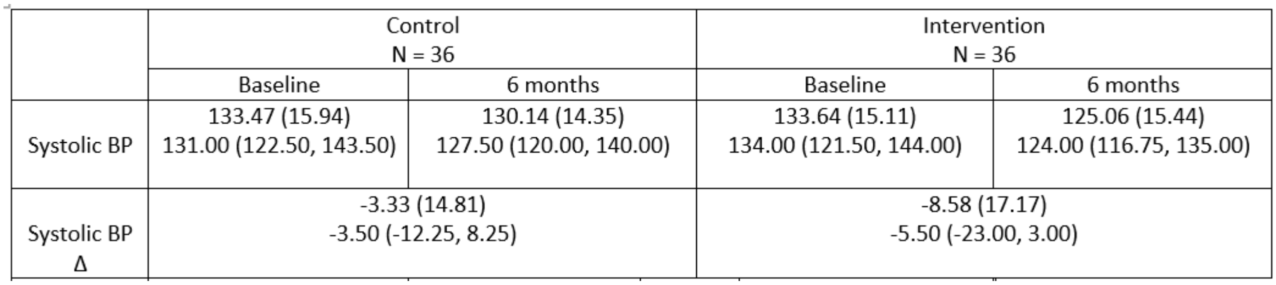
Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

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.

Note: The raw dataset ‘Exercise.csv’ used to generate this table can be found on Canvas.



	Control N = 36		Intervention N = 36	
	Baseline	6 months	Baseline	6 months
Systolic BP	133.47 (15.94) 131.00 (122.50, 143.50)	130.14 (14.35) 127.50 (120.00, 140.00)	133.64 (15.11) 134.00 (121.50, 144.00)	125.06 (15.44) 124.00 (116.75, 135.00)
Systolic BP Δ	-3.33 (14.81) -3.50 (-12.25, 8.25)		-8.58 (17.17) -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:

i) Intervention group (5p)

ii) Control group (5p)

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

c) 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)

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!

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. (5p)
Hint: use rnorm(20, mean = 120, sd = 15)

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)
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)

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 the main 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.