STA255: Assignment 2

Shahriar Shams

Winter, 2020

Submission deadline: April 18, 2020, 11.59pm (Toronto time) (Late submissions will not be accepted)

This entire assignment should take some where between 2 to 6 hours (max). I am giving you almost 15 days to complete it in order to make it inclusive to students in different time zones and students with accommodations and also considering the fact that you have other assignments for other courses. **So please start early and do not request for an extension**.

Instructions on completing the assignment The goal of this assignment is not to test you. Rather it's my try to "force" you to study the materials that we covered after the midterm. If you haven't studied the materials of last few weeks, please study them first at which point this assignment will seem like bunch of exercises.

Instructions on creating documents for submission

- Please create 4 separate pdfs (one for each question).
- I recommend using R-markdown(if you are familiar with it). If you are not familiar with R-markdown, you can write your answers using Microsoft word and in the end save it as pdfs. **Pdf is the only acceptable format of files**.
- You should type all your answers. Handwritten answers(even electronic) are not acceptable.
- Use your judgement on formatting your answers.
- We will use crowdmark for submission and grading. You will have to upload 4 separate documents as your answers to 4 separate questions. Crowdmark links to upload your documents will be emailed to you in couple of days.

Academic Integrity

Each student will work alone. If you need clarification on any of these questions, you are allowed to **ask questions on Piazza**. Don't ask for solutions to anyone. Do not share your codes or answers on any platform.

Question 1 [6 points] (Power/Sample size calculation)

Suppose we have $X_1, X_2, ..., X_n$ independently taken from a $N(\mu, \sigma^2)$. σ^2 is known to be 16 and μ is the unknown parameter.

We want to test $H_0 : \mu = 5$ vs $H_a : \mu = 4$ at 5% level of significance.

You will do two separate calculations here. In part(a), you will calculate Power for a given sample size. In part(b), you will calculate required sample size for a given power.

(a) Suppose we decide to collect 30 observations ($n = 30$). Calculate Power of this test.

(b) Suppose we want to ensure that the power is 80%. What should be the sample size? (n=?)

Data for Question 2,3 and 4

For questions 2,3 and 4, you are not allowed to use the t.test() or lm() commands in R or the equivalent of t.test() or lm() in python or excel or other software. I want you to "manually" calculate everything and show every calculation.

The data sets will be different for each of you. The following little R code will generate the data for the rest of the questions. Copy this code, paste it in R. **Change only the line where it says "student_id=255"**. Remove the number 255 and write your student id there. Now highlight and run the whole thing and it will give you four sets of numbers under the names *Group*1, *Group*2, x and y.

```
# Only change this following line, remove 261 and put your student id
student_id=255

# do not change anything below
set.seed(student_id)
Group1= round(rnorm(15,mean=10,sd=4),2)
Group2= round(rnorm(12,mean=7,sd=4),2)

x= round(rnorm(15,mean=18,sd=4),2)
y= round(50+1.5*x+rnorm(15, mean=0, sd=5),2)

#Data for Quesiton 2 and 3
Group1
Group2

#Data for Quesiton 4
x
y
```

Use the numbers printed under *Group*1 and *Group*2 for answering questions 2 and 3.

Use the numbers printed under x and y for answering question 4.

Graders will rerun this code with your student id and check your numbers, you will get a zero for questions 2, 3 and 4 if the numbers you used do not match with theirs. So don't change anything else(except the student_id) in the code and don't make up numbers.

Question-2 [6 points] (Two sample t-test)

Show all calculations, do not use t.test()

Suppose we have two different sets of samples drawn from two different populations: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ independently of each other.

Suppose the 15 numbers generated under the name $Group1$ are from the $N(\mu_1, \sigma_1^2)$ and the 12 numbers generated under the name $Group2$ are from $N(\mu_2, \sigma_2^2)$.

(a) Assuming $\sigma_1^2 = \sigma_2^2$,

(i) construct a 95% confidence interval for $\mu_1 - \mu_2$ and interpret.
(ii) at 5% level of significance, test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$

(b) Assuming $\sigma_1^2 \neq \sigma_2^2$,

(i) construct a 90% confidence interval for $\mu_1 - \mu_2$ and interpret.
(ii) at 10% level of significance, test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 > \mu_2$

Question-3 [8 points] (Regression with categorical predictor)

For this question you are not allowed to use the lm() command in R or the equivalent of lm() in python or other software. If you want to use a software, only use it as a calculator. So if you are using R, you may only use *mean*(), *sum*(), *qt*() and of course +,-,*,/. (Similar restrictions for Python, excel or others).

Consider all the 27 numbers printed under $Group1$ and $Group2$ as your y values, and the two group indicators as a categorical variable x (indicating Group1 vs Group2).

(a) Fit a least square regression line and calculate the intercept and the slope.

(b) At 5% level of significance, test that the true slope parameter is zero.

(c) Match your answer from part (b) to your answers from Question 2. Describe briefly any similarity that you see.

Question-4 [10 points] (Regression Analysis by hand)

For this question you are not allowed to use the lm() command in R or the equivalent of lm() in python or other software. If you want to use a software, only use it as a calculator. So if you are using R, you may only use *mean*(), *sum*(), *qt*() and of course +,-,*,/. (Similar restrictions for Python, excel or others).

Suppose we are fitting a regression model

$$Y_i|X = x_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$$

X represents number of hours studied during the week before the final exam of STA255. Y represents the score on the final exam.

Y_i's are independent. We have 15 observations (the data that we generated under the name x and y). Think we have observed data from 15 students.

Show **detailed** calculation for each of the followings

(a) Calculate the maximum likelihood estimates of β_1 and β_2 (Let's call them $b1$ and $b2$)

(b) Interpret $b1$ and $b2$

(c) Construct a 95% confidence interval for β_2 and interpret.

(d) At 5% level of significance, test $H_0 : \beta_2 = 1.5$ vs $H_a : \beta_2 \neq 1.5$ and write your conclusion in plain English.

(e) Calculate an estimate of σ^2 using an unbiased estimator.

(f) Compute and interpret the coefficient of determination (R^2).