STA261: 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 1 hour 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 accomodations 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 of last few weeks (specially Week-9 to Week-12). 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 3 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**.
- Question 1(a) is the only part where you can submit a hand written answer. If you are writing it by hand, take a picture of your answer paste it inside you word document or attach it in R-markdown report. For any of the other questions, you can not submit anything hand written, you need to type up.
- Use your judgement on formatting your answers. Make sure to attach any R-code that you use either as part of appendices or as part of your actual report.
- We will use crowdmark for submission and grading. You will have to upload 3 separate documents as your answers to 3 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 [12 points] (Relates to Likelihood Ratio Test)

(a) Suppose $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ where σ^2 is known and μ is unknown.

We want to test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ at the level of significance, α.

Here, we know that $\bar{X}$ is the maximum likelihood estimator of μ (ie. $\hat{\mu} = \bar{X}$).

Option-1: we can do a z-test with the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Option-2: we can do a likelihood ratio test with the test statistic (lets call it W)

$$W = -2log\frac{L(\mu_0)}{L(\hat{\mu})}$$

Show that $W = Z^2$

The goal of part (b) and (c) (below) is to "see" the distribution of the test statistic, W (defined in part(a)) under H_0. In other words, we want to see, if H_0 is really true, what will be the distribution of W. We will do this under two scenarios.

(b) Suppose $X_1, X_2, ..., X_{10} \overset{iid}{\sim} N(\mu, \sigma^2 = 9)$. Treat $\sigma^2 = 9$ as the known constant.

We want to test $H_0 : \mu = 5$ vs $H_1 : \mu \neq 5$ at level of significance, α.

(i) Write a function in R that

- generates 10 samples from a $N(\mu = 5, \sigma^2 = 9)$ distribution
- evaluates the likelihood function at $\mu = 5$ (save it under the name L_theta0)
- evaluates the likelihood function at $\mu = \bar{x}$ (save it under the name L_theta1)
- calculates and returns $-2 * log(L\_theta0/L\_theta1)$

(ii) Run this function (100000 times) using the *replicate*() command (or something similar) and save the output under the name LRT_vec.

(iii) Plot a density histogram using LRT_vec.

code hint: use *hist()* with options *freq=FALSE, breaks=100*.

(iv) Overlay a $\chi^2_{(df=1)}$ density curve on top of this histogram.

code hint: generate 100000 random samples from a $\chi^2_{(df=1)}$, use *density()* and *lines()*

(c) (we will repeat the process of part(b) but with a different distribution here)

Suppose $X_1, X_2, ..., X_{10} \overset{iid}{\sim} Pois(\lambda)$.

We want to test $H_0 : \lambda = 5$ vs $H_1 : \lambda \neq 5$ at level of significance, α.

(i) Write a function in R that

- generates 10 samples from a $Pois(\lambda = 5)$ distribution
- evaluates the likelihood function at $\lambda = 5$ (save it under the name L_theta0)
- evaluates the likelihood function at $\lambda = \bar{x}$ (save it under the name L_theta1)
- calculates and returns $-2 * log(L\_theta0/L\_theta1)$

(ii) Run this function (100000 times) using the $replicate()$ command (or something similar) and save the output under the name LRT_vec.

(iii) Plot a density histogram using LRT_vec.

(iv) Overlay a $\chi^2_{(df=1)}$ density curve on top of this histogram.

(d) In both part (b) and (c), your histograms should match(almost if not completely) with the $\chi^2_{(df=1)}$ density. Make a brief comment on what role you expect the sample size to play in the closeness of the histograms and the density. (In other words, do you expect these type of closeness irrespective of the value of n?)

Question-2 [6 points] (Bayesian Inference)

Data: According to Public Health Canada, as of April 03, there are 295,065 Canadians who have been tested for COVID-19 and 12,784 of them tested positive.

Prior: Public health officials believe that there is a 95% chance that the proportion of COVID-19 cases in Canada (θ) is between 0.011 and and 0.065.

A little try and error approach in R (using function *qbeta()*) suggests that the 2.5th and 97.5th percentiles of a Beta(5,150) distribution are approximately 0.011 and 0.065.

(The outputs of (a), (b) and (c) are numeric numbers)

(a) Calculate the mean of θ (expected proportion) only using the prior distribution. (Let's call it the prior mean)

(b) Calculate the observed proportion only using the data. (Let's call it the sample mean)

(c) Calculate the mean of θ using the posterior distribution of θ. (Let's call it the posterior mean)

(d) As a statistician if you are asked to pick any of these three numbers and forward it to the health ministry to help with decision making (though the numbers are very close), which one will you pick? Why do you think your pick is better than the other two options.

(There is no right or wrong answer here, you can pick any of these you like, briefly justify your preference)

Question-3 [12 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. I want you to "manually" calculate everything and show every calculation. 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()*, *qchisq()* and of course +,-,*,/. (Similar restrictions for Python, excel or others).

The data set will be different for each of you. The following little R code will generate the data for you. Copy this code, paste it in R. **Change only the line where it says "student_id=261"**. Remove the number 261 and write your student id there. Now highlight and run the whole thing and it will give you two sets of numbers under the names x and y.

Those numbers will be the one that you will use to answer the questions on the next page. (The grader will rerun the code with your student id and check your numbers, you will get a zero for the entire question if they don't match).

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

# do not change anything below
set.seed(student_id)
x= round(rnorm(15,mean=18,sd=4),2)
y= round(50+1.5*x+rnorm(15, mean=0, sd=5),2)

x
y
```

Congrats! You have successfully generated your data.

Note: It is possible to complete the rest of the questions (a-h) without using any software.

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 STA261. Y represents the score on the final exam.

Y_i's are independent. We have 15 observations (the data that we generated on the previous page). 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

(d) At 5% level of significance, test $H_0 : \beta_2 = 1.5$

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

(f) Complete the following ANOVA table

Source	df	Sum of Square (SS)	Mean SS = SS/df	F= $\frac{\text{Mean SS for X}}{\text{Mean SS for error}}$
X	?	?	?	?
Error	?	?	?	-
Total	?	?	-	-

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

(h) At 5% level of significance, test $H_0 : \sigma^2 = 25$