MTH 102: Probability and Statistics

Quiz 7; Post (*a Light*) Lunch Assignment

27/05/2020

Sanjit K. Kaul

Explain your answers. Show your steps. You can use all available online resources. If you use a resource other than class notes and the textbook, you must explicitly mention the resource. Anyone should be able to access the said resource. Using a resource but not mentioning it will amount to plagiarism. Help from your classmates or any other person is prohibited and will amount to plagiarism. You may be asked to explain your work post evaluation of your submission.

Your submission must be a single PDF. Your work should be easy to read from the PDF.

Question 1. [30 marks] Your startup has designed a new sensor. You would like to raise money for its production and distribution. You have compared your sensor with one that is the current market leader. You have shown that the average measurement error (sample mean estimate) obtained over 100 measurements using your sensor is smaller by 0.3 than the average error obtained using the market leader.

You are delighted by the reduction in error and arrange a meeting with your prospective clients. They, however, are unimpressed and tell you that the number 0.3 doesn't say much about whether your sensor is truly better than the competitor. They say that the difference may just be a happy coincidence thanks to randomness that is inherent in the process of measurement.

Making measurements is very costly and you can't repeat them. Disappointed, you approach your statistician friend to help better understand the difference of 0.3 in average error. Your friend, of course, reminds you of your favorite book on probability and statistics. Motivated, you dive deep into making sense of the number 0.3.

You know that the measurement errors of both your sensor and the market leader can be modeled as Gaussian distributions with variance 1. You don't have the true means of the distributions but, as mentioned above, you have experimentally measured the difference between sample mean estimates of the mean measurement error obtained using the market leader and your own sensor to be 0.3. Your starting hypothesis is that your new sensor is the same as (has identically distributed measurement error) the market leader. Note that we are interested in the difference in error. Write down the distribution that corresponds to your starting hypothesis.

What can you claim about your sensor and whether it is truly better than the competitor in light of the above? Substantiate your claim with appropriate math.

How would you revise your claim if the difference of 0.3 in average error had been obtained over 1000 instead of 100 measurements? Would you be able to make a stronger claim? Why or why not? Explain. Again, all explanations must be substantiated using math.

Question 2. [25 marks] The news of increase in the use of the next generation of wireless networks (called *NG*) is accompanied by a certain section of social media trend the claim that use of *NG* increases

diastolic blood pressure levels. While the social media platforms try and handle the unexpected increase in traffic, scientists are summoned to verify the claim.

Scientists know the distribution of blood pressure levels in the absence of *NG*. They start with the hypothesis that *NG* doesn't impact blood pressure levels. They measure the blood pressure of n randomly selected people from those who are using *NG* and design a sample mean based significance test with significance $\alpha = 0.05$. As per the designed test, the hypothesis must be rejected when the sample mean is larger than the expected value of blood pressure by a certain level $c > 0$.

The blood pressure levels of the n selected people are measured and this results in a sample mean of $c - 1$. The scientists thus declare with significance 0.05 that *NG* doesn't impact blood pressure levels.

A month later a study by a group named NG-AWARE claims that the earlier scientific study arrived at a false conclusion and that in fact their own study shows that *NG* increases blood pressure levels. The group reveals how it arrived at their conclusion. They do exactly what the scientists did (same choices of n and c) but repeat the experiment of collecting blood pressure levels of n people a total of 50 times. Each time a different group of randomly selected n people is chosen. Out of the 50 experiments, two result in sample means larger than c. This, says the group, makes them conclude that *NG* increases blood pressure levels.

Do you agree with the group NG-AWARE's conclusions? What about the conclusions that the scientists arrived at? Explain your answers. Are the experimental results obtained by the group in contradiction of the scientists observations?

Suppose the blood pressure levels in the absence of *NG* are distributed as a Gaussian with variance 1 and mean 80. Suppose $n = 1000$. Derive the minimum value of c for a significance level of 0.05.

For the calculated value of c, now suppose we only measure a sample size $n = 20$. What is the significance level of your test?

Question 3. [45 marks] A DJ plays songs on request. A new request may arrive when the DJ isn't playing any song. In this case, the DJ plays the new request immediately.

If, on the other hand, a new request arrives while the DJ is playing a song, the request is placed on hold. Assume that requests arrive one at a time. If more than one new request arrives while the DJ is playing a song, only the latest request is placed on hold while all others are discarded. The request that is on hold when the current song finishes playing, is played next by the DJ.

At any given time instant, the length of time to the arrival of the next new song request is given by an exponential random variable with mean 1. The length of time that the DJ plays a song is distributed as an exponential random variable with mean $1/2$ (rate 2). These times are mutually independent.

Suppose you are told that a certain new song request arrived when the DJ was playing a song. **Derive** the probability that the new song request will get played after the current song finishes playing. To do so, **first derive** the distribution of the play time *remaining* for the currently playing song at the instant the new song request arrives.

[Hint 1: Suppose the current song had started playing at time 0 and that the new song request arrives at $t > 0$. At time 0, the playing time of the song is given to be exponentially distributed with mean $1/2$. The song is still playing at time t. What is the probability that the current song will play for at least

an additional amount of time x? Is it a function of t? Does this help calculating the distribution of the remaining play time?]

[Hint 2: Recall that at any given time instant, the length of time to a new song request is exponentially distributed with mean 1. Once you have calculated the distribution of the remaining play time, think of the two competing distributions that decide whether the new song request will be played. You should have both the distributions by now.]

Derive the expected time a song request, which arrives while the DJ is playing a song and is played next by the DJ, stays on hold.

Derive the expected time between the arrival of such a request and the next new request. Note that the next new request must arrive only after the current song finishes playing.

Derive using MGF(s) the distribution of the above time. Use relevant tables in the book to map distributions to MGF(s) and vice versa.

Derive the distribution of the time for which a song request, which arrives while the DJ is playing a song and gets discarded by a newer request, stays on hold. Note that the request stays on hold till the next new request arrives.

Derive the expected time for which a song request, which arrives while the DJ is playing a song and gets discarded by a newer request, stays on hold.

(Q1) The result of your experiment on the basis of which you must either accept or reject your starting hypothesis is the difference between the sample mean of errors obtained using the market leader and your sensor.

Let X be the RV that governs the measurements made by the market leader and let Y be the RV that governs the measurements made by your sensor.

$M_n(X) - M_n(Y)$ is the difference in errors.

Note that both $M_n(Y)$ & $M_n(X)$ have variance $\frac{1}{n}$. This is simply because X and Y have their variances given to be 1.

Defining RV(s), stating sample means, stating involved variances.

10/30

Your starting hypothesis is that the sensors are identical. That is X and Y are identical. Since both are Gaussian with variance 1, if your hypothesis holds, $E[X] = E[Y]$.

Correctly identifying the starting hypothesis

5/10

You have already measured an estimate of the difference to be 0.3. You would like to reject your hypothesis. That is you would like to have 0.3 in the reject set. Of course, you must now provide an associated significance.

Suppose we say that differences $\geq c$ belong to the reject set. We can write

$$p\left[M_n(X) - M_n(Y) \geq c \,\middle|\, H_0\right] = \alpha.$$

$$p\left[\frac{M_n(X) - M_n(Y)}{\sqrt{2/n}} \geq \frac{c}{\sqrt{2/n}} \,\middle|\, H_0\right] = \alpha$$

[Note that $\mathrm{Var}[M_n(X) - M_n(Y)] = \frac{1}{n} + \frac{1}{n} = \frac{2}{n}$]

$\rightarrow n = 100 \Rightarrow c = 0.3.$

$\frac{M_n(X) - M_n(Y)}{\sqrt{2/n}}$ is $N(0,1)$

[Note that $M_n(X)$ and $M_n(Y)$, given H_0, are Gaussian $(\mu, \frac{1}{\sqrt{n}})$, where μ is unknown]

$$Q\left(\frac{c}{\sqrt{2/n}}\right) = \alpha$$

$$\Rightarrow Q\left(\frac{0.3}{\sqrt{2/100}}\right) = \alpha \Rightarrow Q\left(\frac{3}{\sqrt{2}}\right) = \alpha.$$

$$\alpha = 0.0169.$$

15/30

You can go back to your investors and tell them that your measurements have shown that your sensor is better by 0.3 and you say this with $100 - 1.69 = 98.31\%$ confidence. In other words, the significance of your test was 0.0169. So your investors know that the prob of Type-I error is 0.0169.

The prob that an experiment of the kind you performed declares your sensor to be better, when in fact both sensors are similar is 0.0169.

If $n = 1000$,

$$\alpha = Q\left(\frac{0.3}{\sqrt{2/1000}}\right)$$

$$\Rightarrow \alpha = 9.8 \times 10^{-12}$$

A larger number of samples will improve your sample estimates $M_n(X)$ and $M_n(Y)$. You can therefore trust their difference even more.//

Q2

The experiment outcome on the basis of which the scientists decide whether to accept or reject their hypothesis that NG doesn't impact BP is the sample mean of BP levels calculated over a population of n people using NG.

H_0: NG doesn't impact BP.

Reject the hypothesis if

$$\bar{x} - \mu > c.$$

Significance

$$\alpha = P\left[M_n(X) - \mu > c \mid H_0\right]$$

Note that X is the RV that governs the BP levels, given H_0. We know the mean and variance of X.

We are given $\alpha = 0.05$. This implies that the test designed by the scientists will yield a value in the reject set, when H_0 is true, with probability 0.05 (a 5% chance).

Type-I error.

NG-Aware performs 50 experiments out of which 2 give outcomes that belong to the reject set. That is 4% of their experiments. Given the test designed by the scientists has a Type-I error of 5%, it is not surprising that NG-Aware had 2 out of 50 tests give outcomes in the reject set. While the truth is far from known, it is clear that NG-Aware can't claim that NG causes BP based on their experimental results.

We require

$$P\left[M_n(X) - \mu_X > c\right] = 0.05.$$

That is

$$P\left[\frac{M_n(X) - \mu_X}{\sqrt{1/n}} > \frac{c}{\sqrt{1/n}}\right] = 0.05.$$

$$Q\left(c\sqrt{n}\right) = 0.05$$

$$Q\left(c\sqrt{1000}\right) = 0.05$$

$$\boxed{c \approx 0.05}$$

If $c = 0.05$ & $n = 20$.

Significance is $Q(0.05\sqrt{20})$

$= 0.4115$!

This is understandable. With 20 samples your sample mean is likely to be a much noisier estimate. With $c = 0.05$, there is a high probability of Type-I error.

12-8/25

Arguments that are close to correct are Ok. Must not be fundamentally wrong!

10/28

Some may have started with

$P\left(M_n(X) > c\right) = 0.05.$

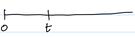
This is okay as long as the rest is correct! Specifically

$$P\left(M_n(X) > c\right) = P\left[M_n(X) - \mu > c - \mu\right]$$

$$= P\left[\frac{M_n(X) - \mu}{(1/\sqrt{n})} > \frac{c - \mu}{(1/\sqrt{n})}\right] \ldots$$

(3)

Hint 1:



Let X be the random length of time a song plays

$$f_X(x) = \begin{cases} 2e^{-2x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

$P[X > t+x \mid X > t]$ is the probability that the current song that is playing t seconds later, will play for at least x more.

$$P[X > t+x \mid X > t] = \frac{P[X > t+x, X > t]}{P[X > t]}$$

$$= \frac{P[X > t+x]}{P[X > t]} = \frac{e^{-2(t+x)}}{e^{-2t}}$$

$$= e^{-2x}, \quad x \geq 0.$$

If the remaining time is Z:

$$P[X > t+x \mid X > t] = P[Z + t > t + x \mid X > t]$$

$$= P[Z > x \mid X > t]$$

$$= e^{-2x}, \quad x \geq 0.$$

This is independent of t. Therefore:

$$f_Z(z) = 2e^{-2z}, \quad z \geq 0.$$

5/45

(a) Derive the prob that the new song request will get played after the current song finishes playing.

Note that this requires that a newer song request doesn't arrive during Z.

The time interval for a newer song request is distributed as an exponential RV with mean 1. Let Y be this interval.

At time t: We require the current song to finish playing before the newer song request arrives. That is we want the event $\{Z < Y\}$ to take place.

Note that Z is an Exp RV with mean $1/2$, since $f_Z(z) = 2e^{-2z}$, $z \geq 0$.

Also, Y is an Exp RV with mean 1.

$$P[Z < Y] = \frac{2}{1+2} = \frac{2}{3}.$$

Identifying the RV(s) involved. Stating the event of interest clearly.

Calculating the prob. Either deriving it or quoting the book

8/45

2/45

(b) Derive the expected time a song request, which arrives while the DJ is playing a song and is played next by the DJ, stays on hold.

The distribution of the time is simply the conditional distribution of Z, given $Z < Y$.

We want $E[Z \mid Z < Y]$. Note that Z and Y are independent. So you know their joint PDF.

$$f_{Z|Z<Y}(z) = \int_z^\infty \frac{f_{Z,Y}(z,y)}{P[Z<Y]}\, dy = \frac{1}{P[Z<Y]} \int_z^\infty f_Z(z) f_Y(y)\, dy$$

$$= \frac{f_Z(z)}{P[Z<Y]} \int_z^\infty e^{-y}\, dy$$

$$= \frac{f_Z(z)\, e^{-z}}{(2/3)}$$

$$= \frac{2e^{-2z}e^{-z}}{2/3}$$

$$= 3e^{-3z}, \quad z \geq 0.$$

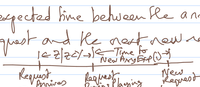
Exp(3).

$$\therefore E[Z \mid Z < Y] = \frac{1}{3}.$$

7/45

3/45

(c) Derive the expected time between the arrival of such a request and the next new request.



The time is: Time the request stays on hold + Time for new request starting at the instant when the request begins playing

The Expected time is

$$E[Z \mid Z < Y] + E[\text{Time for Arrival}]$$

$$\frac{1}{3} + 1 = \frac{4}{3}//.$$

1+1

3/45

(d)

We have the sum of two exponential RVs that are independent of each other.

The RV $Z \mid Z < Y$ is Exp(3) $\longrightarrow \frac{3}{3-s}$

The RV, which is time for new arrival is Exp(1). $\longrightarrow \frac{1}{1-s}$

MGF of the desired RV is $\left(\frac{3}{3-s}\right)\left(\frac{1}{1-s}\right)$

5/45

(e)

If the song request got discarded, then the event $\{Y < Z\}$ took place.

The time such a song stays on hold is the time to a new request, given that $\{Y < Z\}$

We require $f_{X \mid Y<Z}(y)$.

This is simply $2e^{-2y}$, $y \geq 0$, which is Exp(2).

5

(f) The expected time is $\frac{1}{2}//$

5