UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE AND ENGINEERING

Final Examination, April 18, 2020

ECE 286S — Probability and Statistics

Examination Type: D
Examiner: Raviraj Adve

Instructions

- You are allowed 3 *one-sided* 8.5×11 handwritten sheets of notes and a non-programmable calculator.
- Please make sure to write with a pen on white paper to make it easy to read online.
- Make sure to submit the right answer sheet to the various questions.
- You can solve the exam on the single pdf I have uploaded. **Please submit an answer for each question even if it is the same file every time.** Given the time you have, we would appreciate your submitting just the relevant page - but do submit something for every question.
- Show intermediate steps for partial credit. *Answers without justification will not be accepted.*
- **On Quercus, please submit a short paragraph (2-4 sentences) explaining how you approached and solved each question - what concept did you use? In your own words. You must write an explanation to get any marks.** This explanation will factor into your grades.
- Note that this justification must be entered on Quecus, so I can use Turnitin.
- You have to scan in your formula sheet(s) and submit to crowdmark
- This exam is nominally out of 100 marks.
- Take your time - read the information on pg. 2.

Some (potentially) useful information...

- A random variable is referred to as an RV, the probability density function as a pdf, the probability mass function as a pmf.
- The phrase "independent and identically distributed" is written as i.i.d.
- A Poisson RV X with parameter α has mean $E(X) = \alpha$ and variance $\mathrm{var}(X) = \alpha$.
- A binomial RV X over n Bernoulli trials and success probability p has mean $E(X) = np$ and variance $\mathrm{var}(X) = np(1-p)$.
- For any x,

$$e^x = \sum_{k=0}^{\infty} x^k/k!.$$

- The $Q(\cdot)$ function is defined as

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-x^2/2} dx.$$

- Note the tables provided for the area under the Gaussian pdf and for the t-distribution.

Checklist

Make sure you have the submitted the following items

- Your declaration that you will adhere to the code of conduct.
- Your solution to each solution- each question/sub-question requires a submission. There are 16 submissions required for the technical questions.
- A scan of your formula sheet(s) into crowdmark - this is a 17th submission.
- Your typed explanations of how you answered each question - into Quercus.

1 This question builds from the article I posted on pool testing for the virus: we are testing a large population for Covid19. The probability of any person in the population being ***not*** infected is p. Assume infections are independent from person to person.

We combine the samples of n test subjects and test the combined sample. If the test comes out negative, all n are declared as not infected. If the test comes out positive, we test each of the n subjects individually.

Let X denote the random variable measuring the number of tests required to test n subjects. What is the pmf of X? What is the average number of tests required to test n subjects? **6 marks**

2 The pdf of a continuous RV X is non-zero in the range $x \in (0, 1]$. Its CDF $F(x)$ is given by $F(x) = ax^2 + b$ in the range $0 \leq x \leq 1$.

True or False; This given information implies that $b = 0$ and $a = 1$. **5 marks**

3 True or False: If $f(x)$ is a valid pdf, so is $[f(x) \star f(x)]$ (here $\star$ denotes convolution). **5 marks**

4 Let X and Y be jointly continuous RVs with joint pdf $f(x, y)$ non-zero only in a rectangle $x_1 < x < x_2$ and $y_1 < y < y_2$. We are told that we can factor $f(x, y)$ as $\underline{f(x, y) = f_1(x)f_2(y)}$. Here, $f_1(x)$ and $f_2(y)$ are <u>not necessarily</u> the marginal pdfs of X and Y. Prove that X and Y are independent.

<u>6 marks</u>

5 A large train station has a capacity of $C = 100$ passengers. Passengers arrive at the train station independently at an average rate of 10 passengers per minute. With equal probability the inter-arrival time between trains is either 9 minutes or 11 minutes. **11 marks total**

You can assume that every train is can always take all the passengers present in the station.

This question has two parts. See next page as well.

(a) In terms of the $Q(\cdot)$ function, what is the probability that the station reaches its capacity before a train arrives? Justify any assumptions/approximations you make. **6 marks**

Repeating the relevant information from the previous page:

A large train station has a capacity of $C = 100$ passengers. Passengers arrive at the train station independently at an average rate of 10 passengers per minute. With equal probability the inter-arrival time between trains is either 9 minutes or 11 minutes.

You can assume that every train is can always take all the passengers present in the station.

(b) *Question*: Let $n(x; \mu, \sigma)$ denote the Gaussian pdf with mean μ and standard deviation σ. You are told that when a particular train arrived, there were 100 passengers in the station. What is the probability that it had been 9 minutes since the previous train? **5 marks**

You can leave your answer in terms of $n(x; \mu, \sigma)$ with the appropriate values for x, μ and σ.

6 RVs X and Y have ranges $S_X = S_Y = \{-3, -1, 1, 3\}$ with joint pmf

$$f(x, y = \begin{cases} \frac{1}{8} & x \in \{-3, -1\}, y \in \{-3, -1\} \\ \frac{1}{8} & x \in \{1, 3\}, y \in \{1, 3\} \end{cases}$$

Find the correlation coefficient of X and Y. Start by justifying that X and Y are uniform RVs. Then justify that $E(X) = E(Y) = 0$. **6 marks**

7 This question is related to the Poisson RV. **<u>11 marks total</u>**

(a) Show that a Poisson RV, X, with mean α has MGF $M_X(t) = e^{\alpha(e^t-1)}$. **<u>6 marks</u>**

(b) X_1 and X_2 are two independent Poisson RVs with means α_1 and α_2 respectively. What is $h(y)$, the pmf of $Y = X_1 + X_2$?

Answer for part (a):

Repeating the relevant information from previous page:

7 This question is related to the Poisson RV.

(a) Show that a Poisson RV, X, with mean α has MGF $M_X(t) = e^{\alpha(e^t-1)}$.

(b) X_1 and X_2 are two independent Poisson RVs with means α_1 and α_2 respectively. What is $h(y)$, the pmf of $Y = X_1 + X_2$? **5 marks**

Answer for part (b):

8 X_1 and X_2 are i.i.d. zero-mean Gaussian RVs with variance σ^2. Define $\underline{Y_1 = X_1^2 + X_2^2}$ and $\underline{Y_2 = \tan^{-1}(X_2/X_1)}$. Find, $h(y_1, y_2)$, the joint pdf of Y_1 and Y_2. **8 marks**

9 This question relates to the CRLB **<u>17 marks total</u>**

Definition: An estimator is said to be <u>efficient</u> if it is unbiased and its variance meets the Cramer-Rao Lower Bound (CRLB).

In class, when discussing estimation, we always assumed that the samples $X_i, i = 1, \ldots, n$ were i.i.d. Specifically, we focused on Gaussian RVs and, to estimate the mean, used $\hat{\mu} = (1/n) \sum_{i=1}^{n} X_i$. We showed that this is an efficient estimator.

In this question, we will use independent, but not identically distributed, RVs. Specifically, for $i = 1, \ldots, n$, the samples $X_i \sim n(x; \mu_i, \sigma)$ (the means are different, but the variance is the same). Further, we have $\underline{\mu_i = A\alpha^i, i = 1, \ldots, n}$ where α is known. We wish to estimate A.

(a) Show that the <u>joint</u> pdf of all the samples is given by

$$f(\mathbf{x}; A) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[- \sum_{i=1}^{n} \frac{(x_i - A\alpha^i)^2}{2\sigma^2} \right],$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ denotes the vector covering the variables x_i. **<u>4 marks</u>**

(b) Show that this pdf satisfies the regularity condition.

(c) Show that an efficient estimator exists for A.

(d) Find the variance of this estimator.

Answer for only part (a):

Repeating the relevant information

9 (a) Show that the joint pdf of all the samples is given by

$$f(\mathbf{x}; A) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{(x_i - A\alpha^i)^2}{2\sigma^2}\right],$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ denotes the vector covering the variables x_i.

(b) Show that this pdf satisfies the regularity condition. **4 marks**

(c) Show that an efficient estimator exists for A.

(d) Find the variance of this estimator.

Answer for only part (b):

Repeating the relevant information

9 (a) Show that the joint pdf of all the samples is given by

$$f(\mathbf{x}; A) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{(x_i - A\alpha^i)^2}{2\sigma^2}\right],$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ denotes the vector covering the variables x_i.

(b) Show that this pdf satisfies the regularity condition.

(c) Show that an efficient estimator exists for A. **6 marks**

(d) Find the variance of this estimator.

Answer for only part (c):

Repeating the relevant information

9 (a) Show that the joint pdf of all the samples is given by

$$f(\mathbf{x}; A) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{(x_i - A\alpha^i)^2}{2\sigma^2}\right],$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ denotes the vector covering the variables x_i.

(b) Show that this pdf satisfies the regularity condition.

(c) Show that an efficient estimator exists for A.

(d) Find the variance of this estimator. **3 marks**

Answer for only part (d):

10 This question is related to hypothesis testing. **<u>16 marks total</u>**

In “class” we tested the hypothesis that HCQ is a useful treatment for Covid19 patients. There we considered only one class of patients - those that were given HCQ. However, a more realistic trial would be to have two groups of patients: one group who gets the drug (the HCQ group)and a second, control, group, that does not get the drug. There are $n = 100$ patients enrolled in each group.

If in the control group, the probability of recovery is p_C; if in the HCQ group, the probability of recovery is p_H. You can assume that recoveries are independent. We are told $\underline{p_C = 0.25}$.

The hypotheses are:

H_0: HCQ does not provide any benefit for Covid19 patients : $p_H = p_C$
H_1: HCQ does provide a health benefit to patients : $p_H > p_C$.

Let X_C denote the number of patients in the control group that recover and X_H denote the number patients in the HCQ group that recover. <u>We accept H_1 if $(X_H - X_C) \geq 10$</u>.

(a) What is the significance of this test? Justify any approximations you make. **<u>8 marks</u>**

Answer for part (a):

Repeating the question from previous page:

10 A more realistic trial would be to have two groups of patients: one group who gets the drug (the HCQ group)and a second, control, group, that does not get the drug. There are $n = 100$ patients enrolled in each group.

If in the control group, the probability of recovery is p_C; if in the HCQ group, the probability of recovery is p_H. You can assume that recoveries are independent. We are told $\underline{p_C = 0.25}$.

The hypotheses are:

H_0: HCQ does not provide any benefit for Covid19 patients : $p_H = p_C$
H_1: HCQ does provide a health benefit to patients : $p_H > p_C$.

Let X_C denote the number of patients in the control group that recover and X_H denote the number patients in the HCQ group that recover. <u>We accept H_1 if $(X_H - X_C) \geq 10$</u>.

(b) On running a trial with $n = 100$ patients in each group, we get $X_C = 25$ and $X_H = 33$. What is the corresponding p-value? **<u>8 marks</u>**

Answer for part (b):

11 I am attempting to estimate the mean time it takes me to get to office from home (measured in minutes). I measure this time over 5 days and get the following measurements: 31mins, 27mins, 31mins, 32mins and 29mins. Assuming the time taken is Gaussian, find the 95% confidence interval for the mean. **9 marks**