Department of Applied Mathematics
PROBABILITY AND STATISTICS PRELIMINARY EXAMINATION
August 2019

Instructions:

Do two of three problems in each section (Stat and Prob).
Place an **X** on the lines next to the problem numbers
that you are **NOT** submitting for grading.

Prob
1. ____
2. ____
3. ____

Please do not write your name anywhere on this exam.
You will be identified only by your student number.
Write this number **on each page** submitted for grading.
Show all relevant work.

Stat
4. ____
5. ____
6. ____
Total ____

Student Number _____

Probability Section

1. Probability: Problem 1

   Let $X \sim \text{Uniform}(-1, 1)$ and $Z \sim \text{Uniform}(0, 0.1)$ be two independent random variables. Let $Y = X^2 + Z$.

   (a) What is the conditional probability density function of Y given $X = x$?
   (b) What is the joint probability density function of X and Y?
   (c) Are X and Y independent?
   (d) Are X and Y correlated?
   (e) Compute the correlation coefficient between X and Y (start with the definition for the correlation coefficient).

---

2. Probability: Problem 2

   A machine alternates between the good state (state 2) where it functions properly and the bad state (state 1) where it is out of order. The state of such a machine at time $t \geq 0$, $X(t)$, is modeled as a continuous-time Markov chain with the infinitesimal generator (or rate matrix)

$$G = \begin{bmatrix} -a & a \\ b & -b \end{bmatrix},$$

   where $a > 0$ represents how intensely effort is spent on repairing the machine (in the bad state), and $b > 0$ stands for how intensely the machine is being used (in the good state). Consider the corresponding transition probability matrix

$$P(t) = \begin{bmatrix} P_{11}(t) & P_{12}(t) \\ P_{21}(t) & P_{22}(t) \end{bmatrix}, \quad \forall t \geq 0. \tag{1}$$

(a) For any $t \geq 0$, derive the forward equation $P'(t) = P(t)G$, or equivalently,

$$P'_{ij}(t) = \sum_{k \in \{1,2\}} P_{ik}(t) g_{kj}, \quad \forall i, j \in \{1, 2\}.$$

(**Hint:** Use the Chapman-Kolmogorov equation for $P_{ij}(t)$.)

(b) Use part (a) to find explicit formulas of $P_{11}(t)$ and $P_{21}(t)$ in (**??**), in terms of $a, b > 0$.

(c) Suppose that $X(0) = 1$. The limiting fractions of time the chain X spends at state 1 and state 2, denoted by π_1 and π_2 respectively, are defined by

$$\pi_i := \lim_{t \to \infty} \frac{1}{t} \int_0^t 1_{\{X(s)=i | X(0)=1\}} ds, \quad i = 1, 2. \tag{2}$$

Show that $\pi_1 = \frac{b}{a+b}$ and $\pi_2 = \frac{a}{a+b}$.
(**Hint:** Take expectation on both sides of (**??**), so as to use the explicit formula in part (b).)

---

3. Probability: Problem 3

Let $\{Z_n\}_{n \in \mathbb{N}}$ be i.i.d. random variables with $Z_n \sim Exponential(1)$, and $\alpha > 0$ be a given constant. Consider the sequence of random times

$$\tau_0 := 0, \quad \tau_n := \inf \{t \geq \tau_{n-1} : \alpha \cdot (t - \tau_{n-1}) \geq Z_n\} \quad \forall n \in \mathbb{N}.$$

Define the counting process N by $N(t) := n$ for $t \in [\tau_n, \tau_{n+1})$.

(a) Show that for any $t > 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}(N(t) = n \mid \tau_n) = e^{-\alpha \cdot (t - \tau_{n-1})} 1_{\{\tau_n < t\}}.$$

(b) For each $n \in \mathbb{N}$, find the distribution of the random variable $\tau_n - \tau_{n-1}$. Based on this, what is the distribution of τ_n for any $n \in \mathbb{N}$?

(c) Use the distribution of τ_n and part (a) to derive $\mathbb{P}[N(t) = n]$ for all $n \in \mathbb{N}$ and $t > 0$.

Statistics Section

4. Statistics: Problem 4

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Normal distribution with mean θ and variance 1, with $\theta \in \mathbb{R}$.

(a) Show that the best unbiased estimator of θ^2 is given as $\bar{X}^2 - 1/n$.

(b) Calculate the variance of the estimator $\bar{X}^2 - 1/n$. (Recall that for $Y \sim \chi^2(k)$, $E(Y) = k$, and $V(Y) = 2k$.)

(c) Is the estimator $\bar{X}^2 - 1/n$ efficient? Explain.

(d) Find the maximum likelihood estimator (MLE) for θ^2, and find its bias and variance.

(e) Which estimator, $\bar{X}^2 - 1/n$ or the MLE estimator, has a lower mean square error (MSE)? Recall that the MSE of an estimator is defined as $\text{MSE}(\hat{\theta}) = B(\hat{\theta})^2 + V(\hat{\theta})$ (bias squared plus variance).

---

5. Statistics: Problem 5

Let $X_1, X_2, \ldots, X_n$ be *iid* from the distribution with density Beta(μ,1), and let $Y_1, Y_2, \ldots, Y_m$ be *iid* from the distribution with density Beta(θ,1). Assume that the Xs are independent of the Ys.

(a) Find an LRT (likelihood ratio test) of $H_0 : \theta = \mu$ vs $H_1 : \theta \neq \mu$

(b) Show that the test in part (a) can be based on the statistic:

$$T = \frac{\sum \log X_i}{\sum \log X_i + \sum \log Y_i}$$

(c) Find the distribution of T when H_0 is true, and then based on that, show how you would design a test of size $\alpha = 0.10$. (No need to solve for the actual rejection region thresholds.)

---

6. Statistics: Problem 6

Two scientists recently emerged from the rainforests of northeastern Australia with some alarming findings. They discovered two new diseases that have infected the n nearby communities over the past month. The number of individuals infected by the first disease (call it disease X) was $X_1, X_2, \ldots, X_n$. The number of individuals infected by the second disease Y was $Y_1, Y_2, \ldots, Y_n$. The scientists have observed that the disease is not spread person to person. Rather, individuals are infected independently of each other. Furthermore, the scientists believe that the rates of infection for both diseases in each community are proportional to a community-specific vulnerability index v_i for $i = 1, \ldots, n$, such that a community with a higher vulnerability index v_j will likely have a higher number of infected individuals X_j and Y_j. The scientists believe that, given the v_i's, the X_i's and Y_i's are all mutually independent. (Note that the v_i's are theoretical quantities that must be inferred from the data.) Finally, the scientists believe that disease X is λ times more infectious than disease Y.

(a) The scientists need the help of a statistician to explain the observed counts of infected individuals $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. Assume that the number of infected individuals (for both diseases) is MUCH smaller than the total population of each community. Using the information above, write down a model for the X_i's and Y_i's and explain your thinking.

(b) Given your model, write down the likelihood function for the parameters $\lambda, v_1, \ldots, v_n$.

(c) Find the MLE for the parameters $\lambda, v_1, \ldots, v_n$.

(d) A third scientist emerged from the rainforest with new data! She recorded X_{n+1}, the number of individuals from community $n+1$ infected by disease X over the same time period as all the other observations. She unfortunately was not able to record Y_{n+1}. How will this new information affect your estimates of $\lambda, v_1, \ldots, v_n$? What is $E(Y_{n+1}|X_1, \ldots, X_{n+1}, Y_1, \ldots, Y_n)$?