McGill University

Faculty of Science

Department of Mathematics and Statistics

Part A Examination

Statistics: Theory Paper

Date: Tuesday May 7th, 2019 Time: 1pm-5pm

Instructions

- Answer only **two** questions from Section P. If you answer more than two questions, then only the **FIRST TWO questions will be marked.**
- Answer only **four** questions from Section S. If you answer more than four questions, then only the **FIRST FOUR questions will be marked.**

Questions	Marks
P1	
P2	
P3	
S1	
S2	
S3	
S4	
S5	
S6	

This exam comprises the cover page and seven pages of questions.

- ***You may use any result that is known to you, but you must state the name of the result (law/theorem/lemma/formula/inequality) that you are using, and show the work of verifying the condition(s) for that result to apply.***
- ***For the problems with multiple parts, you are allowed to assume the conclusion from the previous part in order to solve the next part, whether or not you have completed the previous part.***

P1. Let $\{X_n : n \geq 1\}$ be a sequence of i.i.d. $\mathbb{R}-$valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that X_n's are NOT a.e. constant, i.e., for every $c \in \mathbb{R}$, $\mathbb{P}(X_n = c) < 1$.

(i) Prove that $\mathbb{P}(X_n > X_{n+1}) > 0$. 10 MARKS

(ii) Prove that $\mathbb{P}(X_n > X_{n+1} \text{ i.o. }) = 1$. 10 MARKS

P2. Let $\{X_n : n \geq 1\}$, $\{Y_n : n \geq 1\}$, X and Y be $\mathbb{R}$-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that $X_n \to X$ in probability and $Y_n \to Y$ in probability as $n \to \infty$.

(i) Prove that $X_n + Y_n \to X + Y$ in probability and $X_n \cdot Y_n \to X \cdot Y$ in probability as $n \to \infty$. 10 MARKS

(ii) Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a continuous and bounded function. Prove that $f(X_n) \to f(X)$ in L^1 as $n \to \infty$. 10 MARKS

P3. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose that $\{Y_n : n \geq 1\}$ is a sequence of i.i.d. random variables with the common distribution being $N(0,1)$. Let $\{a_n : n \geq 1\}$ be a sequence of real numbers. Set $X_0 \equiv 1$, and for $n \geq 1$,

$$X_n := \exp\left(\sum_{j=1}^{n} a_j Y_j - \frac{1}{2}\sum_{j=1}^{n} a_j^2\right).$$

(i) Prove that there exists $X \in L^1$ such that $X_n \to X$ a.s. as $n \to \infty$. 10 MARKS

(ii) Let X be the same as in (i). Prove that, if $\sum_{n=1}^{\infty} a_n^2 < \infty$, then $\mathbb{E}[X] = 1$; if $\sum_{n=1}^{\infty} a_n^2 = \infty$, then $\mathbb{E}[X] = 0$. 10 MARKS

S1. Three prisoners A, B, and C with apparently equally good records have applied for parole. The parole board has decided to release two, but not all three. A warder knows which two are to be released, and one of the prisoners, say A, asks the warder for the name of the one prisoner other than himself who is to be released. While his chances of being released before asking are 2/3, he thinks his chances after asking and being told "B will be released" are reduced to 1/2, since now either A and B or B and C are to be released. He is, however, mistaken. Explain the fallacy.

20 MARKS

S2. It is known that 5% of the members of a population have disease A, which can be discovered by a blood test. Suppose that we want to ascertain the disease status of N (a large number) people. This can be done in two ways: (1) Each person is tested separately; or (2) The N people are divided into n groups, each of size k (assume that $N = nk$ with n and k be integers). Then the blood samples of each group of size k people are combined and analyzed. If the test is negative, all of the people in the group are healthy, that is, just this one test is needed for that group. If the test is positive, each of the k persons in the group must be tested separately, that is, a total of $k + 1$ tests are needed for that group.

a) Find the expected number of tests needed to ascertain the disease status of these N people using scheme (2)? 5 MARKS

b) Find the k that minimizes the expected number of tests required in scheme (2)? 9 MARKS

c) If k is selected as in (b), on average how many tests does scheme (2) save in comparison with scheme (1)? 6 MARKS

S3. Suppose that insect i of a random number, N, of insects lays X_i eggs where $N \sim \text{Poisson}(\lambda)$ and $X_i \overset{iid}{\sim} Ls(p)$ (the Logarithmic series distribution), i.e.

$$P(X_i = t) = -\frac{(1-p)^t}{t\log(p)}, \quad t = 1, 2, \ldots, \quad \text{for } i = 1, 2, \ldots$$

Assume that N is independent of $(X_1, \ldots, X_n)$. Derive the distribution of $H_N = \sum_{i=1}^{N} X_i$, the total number of eggs laid by the insects.

20 MARKS

S4. a) Suppose $X_1, X_2, \ldots, X_n$ are independent random variables from a Gaussian mixture model with probability density function

$$f(x_i; t_i, \boldsymbol{\theta}) = \sum_{j=1}^{K} \pi_j \, \phi(x_i; \alpha_j + \beta_j t_i, \sigma_j^2), \;\; i = 1, 2, \ldots, n,$$

where $t_1, t_2, \ldots, t_n$ are known constants, π_j's and K are known, the vector of unknown parameters is $\boldsymbol{\theta} = (\alpha_1, \beta_1, \sigma_1^2, \ldots, \alpha_K, \beta_K, \sigma_K^2)^\top$, and $\phi(\cdot; \mu, \sigma^2)$ is the pdf of a Gaussian distribution with mean μ and variance σ^2. Devise an EM algorithm to obtain (an approximation to) the MLE of $\boldsymbol{\theta}$. Describe the missing data, complete and incomplete data. Write down the E- and M-steps of the algorithm including the parameter updates in the M-step. 10 MARKS

b) Let Y be a real-valued continuous random variable with the true probability density function (pdf) g. We propose to use a parametric family $\mathcal{F} = \{f(x; \theta) : \theta \in \Theta \subset \mathbb{R}\}$ as an approximation to g. Let θ_0 be a unique value of θ that minimizes the Kullback–Leibler distance between $f(\cdot; \theta)$ and g such that

$$\int_{\mathcal{X}} \left. \frac{d \log f(x; \theta)}{d\theta} \right|_{\theta=\theta_0} g(x) \, dx = 0.$$

Suppose $Y_1, Y_2, \ldots, Y_n$ is an iid sample from g, and we consider the log-likelihood function

$$l_n(\theta) = \sum_{i=1}^{n} \log f(x; \theta).$$

Let $\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}} l_n(\theta)$ be the MLE of θ_0 that satisfies the likelihood equation

$$\left. \frac{dl_n(\theta)}{d\theta} \right|_{\theta=\hat{\theta}_n} \equiv l_n'(\hat{\theta}_n) = 0.$$

Assuming the consistency of $\hat{\theta}_n$, under the regularity conditions discussed in Class for $\mathcal{F}$, **derive** the asymptotic distribution of $\hat{\theta}_n$. 10 MARKS

S5. Let $X_1, X_2, \ldots, X_m$ be i.i.d. from $N(\mu_1, \sigma_1^2)$, and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. from $N(\mu_2, \sigma_2^2)$. The two samples are independent. The unknown parameters are $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$. Note that the Gaussian distribution satisfies the regularity conditions R1-R4 discussed in class.

(a) Find the maximum likelihood estimator (MLE) of $\eta = \frac{\sigma_2^2}{\sigma_1^2}$. Call it $\hat{\eta}_{m,n}$. State the property that is used to obtain this MLE. 4 MARKS

(b) Using an appropriate pivotal quantity, construct an exact $100(1-\alpha)\%$ confidence interval for η. 4 MARKS

(c) Write down the large sample properties of $\hat{\eta}_{m,n}$, as $m, n \to \infty$ such that $m/n \to 1$. (Convergence in probability and asymptotic distribution). 4 MARKS

(d) Using the results in part (c) and Wald statistic, construct an approximate $100(1-\alpha)\%$ confidence interval for η. Write down one potential draw back of this interval, other than being approximate, compared to the one in (b). 4 MARKS

(e) Using the likelihood ratio statistic, design a test for testing

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = \eta_0 \ , \ \ H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq \eta_0$$

at a significance level $0 < \alpha < 1$, for some known η_0. 4 MARKS

S6. a) Suppose that the random variables $Y_1, Y_2, \ldots, Y_n$ satisfy the equation

$$Y_i = \beta x_i + \varepsilon_i, \;\; i = 1, 2, \ldots, n,$$

where $x_1, x_2, \ldots, x_n$ are fixed constants, the ε_i are iid from $N(0, \sigma^2)$, and the unknown parameters are (β, σ^2).

(i) Find the MLE of both parameters (β, σ^2). 3 MARKS

(ii) Find the UMVUE of β, and write down its distribution. 3 MARKS

(iii) Find the Cramer-Rao lower bound (CRLB) for the variance of any unbiased estimator of β. Compare the variance of the estimator in part (b) with the CRLB. Comment on your finding. 3 MARKS

(iv) Find the moment estimator of β, and compare its variance with the CRLB. 3 MARKS

b) Let $X_1, X_2, \ldots, X_n$ be independent Poisson random variables each with mean

$$\mu(t_i) = E(X_i; t_i) = \exp\{\beta_0 + \beta t_i\} \;\; ; \;\; i = 1, 2, \ldots, n.$$

The vector of unknown parameters is $\boldsymbol{\theta} = (\beta_0, \beta_1)^\top$, and $t_1, t_2, \ldots, t_n$ are known constants. Assume the regularity conditions for this parametric family.

(i) Find (if possible) a minimal sufficient statistic for $\boldsymbol{\theta}$. 2 MARKS

(ii) It is known that there is no closed form for the MLE of $\boldsymbol{\theta}$. Explain, in full details, how the Newton-Raphson algorithm can be use to find an approximation to the MLE of $\boldsymbol{\theta}$. 3 MARKS

(iii) Using your choice of approximation, for large n: **(i)** construct an approximate $100(1-\alpha)\%$ confidence interval for β_1; **(ii)** explain how to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, at a significance level α. 3 MARKS