Vanderbilt University Biostatistics Comprehensive Examination

MS Theory Exam/
PhD Theory Exam Series 1

May 20, 2024

Instructions: Please adhere to the following guidelines:

- This exam begins on Monday, May 20 at 9:00am. You will have until 2:00pm to complete it.
- There are four equally weighted problems of varying length and difficulty. Note that not all sub-problems are weighted equally. You are strongly advised not to spend too much time on any one problem.
- Answer each question clearly and to the best of your ability. Partial credit will be awarded for partially correct answers.
- Be as specific as possible, show your work when necessary, and please write legibly.
- This is a closed-everything examination, though you will be permitted to use a scientific calculator.
- This examination is an *individual effort*. Vanderbilt University's academic honor code applies.
- Please address any clarifying questions to the exam proctor.

---

1. $\boxed{\text{25 pts}}$ A study was conducted to characterize the circulation of SARS-CoV-2, the virus responsible for COVID-19. The number of (independently sampled) patients who enroll in the study, N, is distributed as a Poisson(ν) random variable. Each enrolled patient has probability p of testing positive for SARS-CoV-2. Let Y denote the number of enrolled patients in the study who test positive for SARS-CoV-2.

(a) In terms of ν, what is the probability of enrolling exactly two patients in the study?

(b) Given that four patients are enrolled in the study, what is the probability, in terms of p, that at least two test positive for SARS-CoV-2?

(c) State the values of the following quantities (you may simply state your answers without proof).

- $\mathrm{E}[N]$
- $\mathrm{Var}[N]$
- $\mathrm{E}[Y|N]$
- $\mathrm{Var}[Y|N]$

(d) Use the results of part (c) to determine the value of $\mathrm{E}[Y]$.

(e) Use the results of part (c) to determine the value of $\mathrm{Var}[Y]$.

(f) Show that $Y \sim \text{Poisson}(\theta)$ by computing the marginal probability mass function, $p_Y(y;\theta)$, where $\theta = \nu p$.

(g) Derive the moment-generating function (MGF) of Y and use it to compute the values of $\mathrm{E}[Y]$ and $\mathrm{Var}[Y]$ (thereby confirming your calculations of part (d) and (e)).

(h) Devon is a statistician who seeks to estimate θ as a Bayesian based on having enrolled five-hundred patients, ten of whom tested positive for SARS-CoV-2. Devon utilized the marginal probability mass function of part (f) as the likelihood, together with the prior $\theta \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$. Devon reported an estimate of $\widehat{\theta} = 8$ based on the posterior mean. Determine the value of λ that was featured in Devon's prior.

(i) Even without performing the calculation of part (h), briefly (a maximum of three sentences), explain why you could have anticipated that Devon chose a value of $\lambda > 1/8$ without having explicitly solved for it. Your argument can be heuristic.

Key information: The following is information that you will likely find helpful in this problem. You are free to utilize any and all of the results below without providing any further proof; however, note that this may not include information on every probability function you will need in this problem.

(I) $\exp(x) = \sum_{n=0}^{\infty} (x^n/n!)$.

(II) If $X \sim \text{Poisson}(\lambda)$, then:

- X has probability mass function given by:

$$p_X(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}; \quad \lambda > 0, \quad x = 0, 1, 2, \ldots$$

(III) If $X \sim \text{Exponential}(\lambda)$, then:

- X has probability density function given by:

$$f_X(x;\lambda) = \lambda \exp(-\lambda x); \quad \lambda > 0, \quad x > 0.$$

(IV) If $X \sim \text{Gamma}(\alpha, \beta)$, then:

- X has probability density function given by:

$$f_X(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x); \quad \alpha, \beta > 0, \quad x > 0.$$

- $\mathrm{E}[X] = \alpha/\beta$.

2. $\boxed{\text{25 pts}}$ Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ comprise independent samples, with $X_i \sim \text{Exponential}(\lambda_X)$ and $Y_i \sim \text{Exponential}(\lambda_Y)$. Throughout this problem, let $S_n^X = \sum_{i=1}^n X_i$ and let $S_m^Y = \sum_{i=1}^m Y_i$.

(a) Show that $S_n^X \sim \text{Gamma}(n, \lambda_X)$. From this, you can and should conclude an analogous statement about the distribution of S_m^Y without going through the same calculations.

(b) Determine the likelihood ratio statistic, $\Lambda_{n,m}$, for testing $H_0 : \lambda_X = \lambda_Y$ vs. $H_1 : \lambda_X \neq \lambda_Y$. Provide the form of the likelihood ratio test as part of your response.

(c) Argue that the test you derived in part (b) can be based on the value of the statistic:

$$R_{n,m} = \frac{S_n^X}{S_n^X + S_m^Y}.$$

You should try to characterize the rejection region associated with $R_{n,m}$ as part of your response.

(d) Show that when H_0 is true, $R_{n,m} \sim \text{Beta}(n, m)$. *Hint*: Begin by identifying the joint distribution of $U = R_{n,m}$ and $V = S_n^X + S_m^Y$ and factor the joint density accordingly.

(e) Argue that when $n = m$, $R_n \equiv R_{n,m} \xrightarrow{p} 1/2$ (i.e., as $n \to \infty$).

Key information: The following is information that you will likely find helpful in this problem. You are free to utilize any and all of the results below without providing any further proof; however, note that this may not include information on every probability function you will need in this problem.

(I) If $X \sim \text{Exponential}(\lambda)$, then:

- X has probability density function given by:

$$f_X(x; \lambda) = \lambda \exp(-\lambda x); \quad \lambda > 0, \quad x > 0.$$

- X has moment-generating function given by $M_X(t) = \lambda/(\lambda - t)$, for $t < \lambda$.

(II) If $X \sim \text{Gamma}(\alpha, \beta)$, then:

- X has probability density function given by:

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x); \quad \alpha, \beta > 0, \quad x > 0.$$

- X has moment-generating function given by $M_X(t) = (1 - t/\beta)^{-\alpha}$, for $t < \beta$.

(III) If $X \sim \text{Beta}(\alpha, \beta)$, then:

- X has probability density function given by:

$$f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

- $\text{E}[X] = \alpha/(\alpha + \beta)$.
- $\text{Var}[X] = (\alpha\beta)/[(\alpha + \beta)^2(\alpha + \beta + 1)]$.

3. $\boxed{\text{25 pts}}$ Let $X_1, \ldots, X_n$ denote i.i.d. Exponential(θ) random variables, each having density function given by:

$$f_X(x;\theta) = \theta \exp(-\theta x); \quad \theta > 0, \quad x > 0.$$

(a) Determine the maximum likelihood estimator (MLE) of θ; call it $\widehat{\theta}_n$.

(b) Show that $\mathrm{E}[\widehat{\theta}_n] = n\theta/(n-1)$. You may freely use the result of Problem 2(a) in your response (re-stated below for your convenience), but you should use the definition of expectation to specifically determine the value of $\mathrm{E}[\widehat{\theta}_n]$.

(c) Suggest an unbiased estimator based on $\widehat{\theta}_n$; call it $\widetilde{\theta}_n$. Argue that $\widetilde{\theta}_n$ is the unique uniformly minimum-variance unbiased estimator (UMVUE) for θ, citing any theorems you invoke and justifying why they apply.

(d) Determine whether $\widetilde{\theta}_n$ achieves the Cramér-Rao bound for the variance of an unbiased estimator for θ in finite samples.

(e) Now consider estimators having the form $\overline{\theta}_n = c\widetilde{\theta}_n$, where $c > 0$ may depend upon n. Determine the value of c that minimizes the quantity $\mathrm{MSE}(\overline{\theta}_n) = \mathrm{E}[(\overline{\theta}_n - \theta)^2]$.

Key information: The following is information that you will likely find helpful in this problem. You are free to utilize any and all of the results below without providing any further proof; however, note that this may not include information on every probability function you will need in this problem.

(I) If $X_1, \ldots, X_n$ are i.i.d. Exponential(λ) random variables, then $S_n^X = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$. Note that this is simply a restatement of what you are asked to show in Problem 2(a), but you may use this fact in Problem 3(b) without showing it.

(II) If $X \sim \text{Gamma}(\alpha, \beta)$, then:

- X has probability density function given by:

$$f_X(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x); \quad \alpha, \beta > 0, \quad x > 0.$$

4. $\boxed{\text{25 pts}}$ Suppose that Y is a random variable that takes on non-negative integer values. You are interested in estimating the quantity $\theta = \mathrm{P}(Y = 0)$ based on the i.i.d. random variables $Y_1, \ldots, Y_n$, each having the same probability mass function as Y.

(a) Let $S_n = Z_1 + \cdots + Z_n$, where $Z_i = I(Y_i = 0)$. Then, $\widehat{\theta}_n = S_n/n$ is a very sensible estimator of θ that does not presume that the Y_i's conform to a particular known distribution. Determine $\mathrm{E}[\widehat{\theta}_n]$ and $\mathrm{Var}[\widehat{\theta}_n]$.

For parts (b)-(f), suppose you (correctly) assume that the $Y_1, \ldots, Y_n$ are distributed as independent Poisson(λ) random variables.

(b) Determine the maximum likelihood estimator, $\widetilde{\theta}_n$, for θ.

(c) Use Jensen's inequality to argue that $\widetilde{\theta}_n$ is "not downwardly biased" in the sense that $\mathrm{E}[\widetilde{\theta}_n] \geq \theta$.

(d) Using the definition of convergence in probability, define what it means for $\widetilde{\theta}_n$ to be consistent for θ; argue that $\widetilde{\theta}_n$ is consistent for θ, naming any theorems you invoke.

(e) Use the delta method to determine an asymptotically valid expression for $\mathrm{Var}[\widetilde{\theta}_n]$.

(f) Show that, asymptotically, $\mathrm{Var}[\widetilde{\theta}_n] < \mathrm{Var}[\widehat{\theta}_n]$.

(g) Now suppose that in reality (and unbeknownst to you), $Y_1, \ldots, Y_n$ are distributed as independent Geometric(ϕ) random variables despite your assumption that they are distributed as Poisson(λ) random variables. Show that $\widetilde{\theta}_n \xrightarrow{p} \exp(1 - 1/\theta)$, where $\widetilde{\theta}_n$ is the estimator you derived in part (b).

(h) Very briefly (no more than four sentences), describe how your findings from parts (a)-(g) are aligned with the following statement: "*A statistician is rewarded for making correct assumptions, but penalized for making incorrect assumptions.*"

Key information: The following is information that you will likely find helpful in this problem. You are free to utilize any and all of the results below without providing any further proof; however, note that this may not include information on every probability function you will need in this problem.

(I) If $X \sim$ Poisson(λ), then:

- X has probability mass function given by:

$$p_X(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}; \quad \lambda > 0, \quad x = 0, 1, 2, \ldots$$

(II) If $X \sim$ Geometric(ϕ), then:

- X has probability mass function given by:

$$p_X(x; \phi) = (1 - \phi)^x \phi; \quad 0 < \phi < 1, \quad x = 0, 1, 2, \ldots$$

- $\mathrm{E}[X] = (1 - \phi)/\phi$.

(III) $x + 1 < \exp(x)$ for all $x \neq 0$ [this is straightforward but possibly time consuming to show].