Vanderbilt University Biostatistics Comprehensive Examination

MS Theory Exam/
PhD Theory Exam Series 1

May 19, 2025

Instructions: Please adhere to the following guidelines:

- This exam begins on Monday, May 19 at 9:00am. You will have until 2:00pm to complete it.
- There are four problems of varying length and difficulty. Note that not all problems and subproblems 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. $\boxed{\text{30 pts}}$ Suppose the number of offspring produced by an organism follows a Poisson distribution with mean μ. Each offspring from that generation (if any) then has its own independent number of offspring, each also Poisson-distributed with mean μ. This process continues until all offspring in a generation produce no further offspring (extinction). To facilitate notation, let $G_0 \equiv 1$ denote the number in the "zeroth" generation (the original organism), G_1 denote the number in the first generation (the number of direct offspring of the original organism), G_2 denote the number in the second generation (the combined number of offspring produced by all of the offspring in G_1), etc. The total number of individuals in this process (including the original organism), is given by $Y = G_0 + G_1 + \cdots$, and follows a Borel(μ) distribution with mass function given by:

$$p_Y(y;\mu) = \frac{\exp(-\mu y)(\mu y)^{y-1}}{y!}; \quad y = 1, 2, 3, \ldots \text{ if } 0 < \mu < 1.$$

If $\mu = 0$, then $Y \equiv 1$. **You need not derive this mass function and may take it as a given.** Further, assume for this problem that $0 \leq \mu < 1$. For convenience, if $G_j = 0$, you can let $G_{j'} = 0$ for all $j' > j$.

(a) Determine the value of $\mathrm{E}[G_2|G_1]$ in terms of μ.

(b) Determine the value of $\mathrm{E}[G_2]$ in terms of μ using the law of iterated expectation (also called the law of total expectation).

(c) Argue that the expected number of offspring in the j^{th} generation is given by μ^j.

(d) Argue that $\mathrm{E}[Y] = 1/(1-\mu)$.

(e) Use Markov's inequality to argue that the probability of eventual extinction is one. *Note*: Markov's inequality asserts that for any nonnegative random variable, X, $\mathrm{P}(X \geq a) \leq \mathrm{E}[X]/a$ for all $a > 0$.

(f) Derive the maximum likelihood estimator for μ based on n i.i.d. Borel(μ) observations, and show that it is equal to the method-of-moments estimator. Call this estimator $\widehat{\mu}_n$.

(g) Justifying your answer by naming and/or stating (but not proving) any theorems you invoke, argue that $\widehat{\mu}_n$ is consistent for μ.

For the remainder of the problem, suppose you obtain data from a single organism and find that it has no offspring (i.e., $Y = 1$). On the basis of this single observation, you seek to estimate μ.

(h) Determine the value of $\widehat{\mu}$ (i.e., the maximum likelihood estimate).

(i) Suppose you seek to estimate μ as a Bayesian under the prior $\mu \sim \text{Uniform}(0,1)$. Derive the posterior density (you need not identify it by name, but you should fully express the density and its support).

(j) Consider estimating μ as the posterior mean. Call this estimator $\widetilde{\mu}$, and determine its value (please round your answer to three significant digits). In your response, include a graph of the parameter space; mark both $\widehat{\mu}$ and $\widetilde{\mu}$ on that graph.

(k) Your colleague asserts that a frequentist analysis can be thought of as merely a Bayesian analysis with a flat prior on the unknown parameter. Briefly explain how this problem directly conflicts with your colleague's line of thinking. How might you explain the proper interpretation of a maximum likelihood estimate and a posterior mean to your colleague?

(l) Suppose you wanted to check your math in part (j) using computing capabilities. Outline the steps of an algorithm based on the inverse-CDF method that could be used to generate draws from the posterior using software (from which you could then ostensibly take the sample mean of a large number of draws to numerically determine $\widetilde{\mu}$). You do not need to present your answer in accordance with syntax of any specific software package.

---

2. $\boxed{\text{25 pts}}$ Suppose $Y_1, \ldots, Y_n$ are independent random variables, each with common density function given by:

$$f_Y(y; \lambda) = \frac{\lambda}{y^2}; \quad 0 < \lambda < y < \infty,$$

where $\lambda > 0$ is an unknown parameter.

(a) Determine the maximum likelihood estimator, $\widehat{\lambda}_n$, of λ.

(b) Determine the exact distribution of $\widehat{\lambda}_n$ (showing your work, express both the CDF and the PDF).

(c) Determine $\mathrm{E}[\widehat{\lambda}_n]$ and $\mathrm{Var}[\widehat{\lambda}_n]$.

(d) Argue that the minimum order statistic, $Y_{(1)}$, is a sufficient statistic for λ.

(e) It happens that $Y_{(1)}$ is minimal sufficient. State what this means (i.e., by definition), but you need not prove this. In your response, provide another example of a sufficient statistic for λ that is not minimal sufficient.

(f) It happens that $Y_{(1)}$ is complete. State what this means (i.e., by definition), but you need not prove this.

(g) Determine the unique minimum-variance unbiased estimator (MVUE) for λ; call this estimator $\widetilde{\lambda}_n$. Justify your answer by naming/stating any major theorems you invoke.

(h) Determine the mean squared error (MSE) for each of $\widehat{\lambda}_n$ and $\widetilde{\lambda}_n$. Although the MSE for each estimator goes to zero, show that the *ratio* of the two MSEs does not approach one as $n \longrightarrow \infty$. Based on your finding, which estimator, $\widehat{\lambda}_n$ or $\widetilde{\lambda}_n$, would you recommend?

---

3. $\boxed{\text{20 pts}}$ Consider a sample of independent random variables, $X_1, \ldots, X_n$, each with density given by:

$$f_X(x;\theta) = \left(\frac{x}{\theta}\right)^{\theta-1}; \quad 0 < x < \theta < \infty,$$

where $\theta > 0$ is an unknown parameter.

(a) Argue that the maximum likelihood estimator for θ is given by the maximum order statistic: $\widehat{\theta}_n = X_{(n)}$.

(b) Determine the bias of $\widehat{\theta}_n$ as a function of θ and n.

(c) Show that

$$n(\theta - \widehat{\theta}_n) \xrightarrow{d} \text{Exponential}(1).$$

(d) Consider testing the hypothesis $H_0 : \theta = 1$ vs. $H_1 : \theta \neq 1$. In accordance with the convergence result of part (c), let $T_n = n(1 - \widehat{\theta}_n)$ denote a test statistic for this hypothesis. Determine the rejection region associated with a level-α test of this hypothesis based on T_n. *Hint*: In forming a suitable rejection region, consider and account for the fact that certain values of T_n supply evidence in to support the assertion that $\theta < 1$, while other values of T_n prove beyond a shadow of a doubt that $\theta > 1$.

(e) Determine the approximate power of a 0.05-level test of the hypothesis in part (d) when $n = 50$ and $\theta = 1.02$. *Hint*: Begin by writing $T_n = n(1 - \theta) + n(\theta - \hat{\theta})$ and argue that T_n follows an approximate shifted exponential distribution under fixed alternatives.

---

4. $\boxed{\text{25 pts}}$ **Background**: The receiver operating characteristic (ROC) curve is a useful way to visually summarize the utility of a continuous measure as a classifier between two groups (0 = healthy and 1 = diseased). Let Y_0 and Y_1 denote randomly sampled values of the measure for individuals from the healthy and diseased populations (respectively). Assume without loss of generality that higher values of the outcome are associated with a diseased state (e.g., systolic blood pressure for hypertension or prostate-specific antigen for prostate cancer). In this way, the true positive rate (TPR) at a particular cut-off point, c, is given by $S_1(c) = \text{P}(Y_1 > c)$ and the false positive rate (FPR) is given by $S_0(c) = \text{P}(Y_0 > c)$. The ROC curve is a graph that features the FPR on the x-axis and the TPR on the y-axis across all possible cut-off points, c.

For this problem, suppose you sample observations $Y_{0,1}, \ldots, Y_{0,n_0}$ from the healthy population, each following an Exponential(λ_0) distribution, and you sample observations $Y_{1,1}, \ldots, Y_{1,n_1}$ from the diseased population, each following an Exponential(λ_1) distribution (all observations are independent).

(a) Determine the maximum likelihood estimator for λ_0. By analogy, write down (but do not re-derive) the maximum likelihood estimator for λ_1.

(b) Use the delta method to derive the form of a $100 \times (1-\alpha)\%$ confidence interval for FPR(c). An analogous $100 \times (1-\alpha)\%$ confidence interval for TPR(c) takes the same form (but you need not write it down).

(c) If you recall one of the key assumptions of the delta method, you should be able to identify a key problem with the confidence intervals you formed in part (b). Briefly state the key problem (*Hint*: consider what happens in a finite sample if c is very small or very large).

(d) Show that $\sqrt{n_0}(\log(\widehat{\lambda}_0) - \log(\lambda_0)) \xrightarrow{d} \mathcal{N}(0, 1)$. An analogous statement about the asymptotic distribution of $\log(\widehat{\lambda}_1)$ holds (but you need not write it down).

(e) Based on the result of part (d), suggest a $100 \times (1-\alpha)\%$ confidence interval for FPR(c) that does not suffer from the problem you identified in part (c). An analogous $100 \times (1-\alpha)\%$ confidence interval for TPR(c) takes the same form (but you need not write it down).

(f) Now suppose you wish to form a confidence region for a point on the ROC curve (i.e., at a fixed, designated cut-off point, c). One way to do this would be to form $100 \times (1-\alpha)\%$ confidence intervals for each of FPR(c) and TPR(c) as you have in part (d) and taking their Cartesian product (i.e., to form a confidence rectangle). The confidence rectangle contains the point if and only if the confidence intervals for both FPR(c) *and* TPR(c) contain the true values. Keeping in mind that both the FPR and TPR are estimated independently, determine an appropriate value for α so that such a confidence rectangle would be expected to have 95% coverage.

We are not always interested in estimating the ROC curve at a specific cut-off point. The curve itself, and in particular the area under it, is an aggregate measure of the extent to which the distributions of Y_0 and Y_1 differ. It turns out that the theoretical ROC curve can be expressed as $\text{ROC}(p) = S_1(S_0^{-1}(p))$, $0 < p < 1$; the FPR (p, on the x-axis) is the input of this function and the TPR (ROC(p), on the y-axis) is the output.

(g) Under the assumption of exponentially distributed data, show that $\text{ROC}(p) = p^{\lambda_1/\lambda_0}$.

(h) The area under the ROC curve (AUC) is a useful summary measure of the ROC curve. Under the assumption of exponentially distributed data, derive an expression for the AUC and provide the form of a $100 \times (1-\alpha)\%$ confidence interval for the AUC.
