Biostatistics MS Comprehensive Exam: Theory

June 1 - 2, 2021

Instructions: Please adhere to the following guidelines:

- The MS Theory Comprehensive Exam will be administered on Tuesday, June 1 at 9:00am (central time); you have until Wednesday, June 2 at 5:00pm (central time) to complete the exams and place your responses into your respective Box folder. You may (should) place draft solutions in your Box folder throughout the examination period; the latest version submitted prior to the deadline will be considered the final version. In addition, please also email your final version to Drs. Andrew Spieker (**andrew.spieker@vumc.org**) and Robert Greevy (**robert.greevy@vumc.org**) prior to the deadline (dual submission helps ensure the exam is received).
- There are six equally weighted problems of varying length and difficulty. Note that not all subproblems are weighted equally. You are 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 for any handwritten responses.
- This is an open-book and open-notes examination, but it is an *individual effort*; do not discuss any part of this exam with anyone. Vanderbilt University's academic honor code applies.
- Please email any clarifying questions to:
  Dr. Andrew Spieker (andrew.spieker@vumc.org), and
  Dr. Amber Hackstadt (amber.hackstadt@vumc.org).

1. $\boxed{\textbf{25 pts}}$ Consider the following probability density function (PDF) for a random variable X:

$$f(x;\alpha) \quad = \quad c \times \frac{\alpha^3}{x^4}; \qquad x > \alpha.$$

Here, $\alpha > 0$ is an unknown parameter and c is the real-valued constant such that $f(x;\alpha)$ is a valid PDF.

(a) Determine the value of c.

(b) Determine $\mathrm{E}[X]$ and $\mathrm{Var}[X]$ using the definitions of expectation and variance.

(c) Determine the cumulative distribution function (CDF) of X.

For the remainder of the problem, let $X_1, \ldots, X_n$ denote independent random variables, each having PDF $f(x;\alpha)$ with $\alpha = 1$. Further, let $X_{(1)} = \min_{1 \leq i \leq n} X_i$.

(d) Appealing to your response to part (c), implement the inverse-CDF approach using R in order to approximate the expected value and variance of $X_{(1)}$ under the following sample sizes:

- $n = 5$
- $n = 50$
- $n = 500$
- $n = 5000$

Please use 100,000 simulation replicates for each sample size. Comment on the results. To what value does $X_{(1)}$ appear to converge as $n \longrightarrow \infty$? **Please include your R code as an appendix**.

(e) Prove that in large samples, $\mathrm{P}(X_{(1)} > 1 + 1/n) \approx e^{-3}$.

(f) Prove that as $n \longrightarrow \infty$,

$$\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} X_i X_j \xrightarrow{p} r.$$

Specifically determine the value of r as part of your response to this problem.

---

2. $\boxed{\textbf{25 pts}}$ Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables, each following a standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

(a) Let $Y_1 = X_1$ and let $Y_2 = X_1 + X_2$. Determine the following values:

- $\mathrm{E}[Y_2|Y_1]$
- $\mathrm{E}[Y_1|Y_2]$
- $\mathrm{Var}[Y_2|Y_1]$
- $\mathrm{Var}[Y_1|Y_2]$

(b) Let $Y_j = \sum_{i=1}^n a_{ji} X_i$ for real-valued constants a_{ji}. Show that

$$\mathrm{E}[Y_j|Y_k] \quad = \quad \frac{\sum_{i=1}^n a_{ji} a_{ki}}{\sum_{i=1}^n a_{ki}^2} Y_k,$$

and determine an expression for $\mathrm{Var}[Y_j|Y_k]$.

---

3. $\boxed{\textbf{25 pts}}$ Suppose $X_1, \ldots, X_n$ are i.i.d. random variables each with density function given by

$$f(x; \alpha, \beta) \quad = \quad \frac{\alpha\beta^\alpha}{x^{1+\alpha}} \text{ with } x > \beta \qquad (\alpha, \beta > 0)$$

(a) Show that $Y_i = \log(X_i/\beta)$ follows an exponential distribution with rate parameter α.

(b) Let $W_{(i)} = \log X_{(i)}$ denote the i^{th} order statistic of the log-transformed variables, and let $Z_i = W_{(i)} - W_{(1)}$ for $i = 2, \ldots, n$. Show that $(Z_2, \ldots, Z_n)$ is distributed as the order statistics of $n - 1$ i.i.d. exponential random variables.

Hint: You may utilize the following fact without proof:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{pmatrix} \Longrightarrow \mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

(c) Derive the maximum likelihood estimators, $\widehat{\alpha}_n$ and $\widehat{\beta}_n$.

(d) Consider the hypothesis test $H_0 : \alpha = 1$ vs. $H_1 : \alpha \neq 1$, together with the test statistic

$$T = \log\left(\frac{\prod_{i=1}^n X_i}{\min_{1 \leq i \leq n} X_i^n}\right).$$

Show that under the null hypothesis, $2T$ follows an exact χ^2 distribution with a certain number of degrees of freedom (determine the degrees of freedom as part of your response to this problem).

---

4. $\boxed{\textbf{25 pts}}$ Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ for $0 < \theta < 1$.

(a) Name the distribution of $Y_n = \sum_{i=1}^n X_i$, and state its expectation and variance.

(b) Suppose we estimate θ via $\widehat{\theta}_n = Y_n/n$. Determine the asymptotic distribution of $\sqrt{n}\Big(\widehat{\theta}_n - \theta\Big)$, justifying your response by naming any theorems you invoke.

(c) Use the delta method to show that

$$\sqrt{n}\Big(\widehat{\theta}_n(1-\widehat{\theta}_n) - \theta(1-\theta)\Big) \xrightarrow{d} \mathcal{N}(0, v(\theta))$$

for some $v(\theta)$; specifically determine $v(\theta)$ as part of your response. In no more than two sentences, explain why this asymptotic result is not particularly useful when $\theta = 1/2$.

(d) Based on the asymptotic result you obtained in part (c), determine the form of a symmetric (two-sided) 95% confidence interval for $\theta(1-\theta)$. The result of part (c) suggests that when $\theta = 1/2$, this confidence interval will not possess the desired property of approximate 95% coverage, even in very large samples. Illustrate this via simulation (e.g., using R) under the very large sample size of $n =$ 100,000. Please use 5,000 simulation replicates for this problem. **Please include your R code as an appendix**.

(e) Again focusing on the special case of $\theta = 1/2$, determine the sequence a_n and the constant b such that

$$a_n\Big(\widehat{\theta}_n(1-\widehat{\theta}_n) - b\Big) \xrightarrow{d} \chi_1^2.$$

Hint: First, express $f(x) = x(1-x)$ in the form $f(x) = c_0 + c_1(x-1/2) + c_2(x-1/2)^2$ for some real-valued constants c_0, c_1, and c_2 using the Taylor formula; you should find that $c_1 = 0$. The value of c_0 should inform you greatly about the value of b. To then figure out the value of a_n, use the result of part (b) for $\theta = 1/2$.

---

5. $\boxed{\textbf{25 pts}}$ Suppose $X \sim \mathrm{Uniform}(0, \theta)$ for some unknown $\theta \in (0, 1)$ to be estimated. For this problem, you may freely use without proof the fact that this is a *complete* family. In this problem, *all* estimators for θ are based on the single observation, X.

(a) Determine a first-order method-of-moments estimator, $\widehat{\theta}_{\mathrm{MME}}$, for θ.

(b) Naming any theorems you invoke and justifying why they apply, argue that $\widehat{\theta}_{\mathrm{MME}}$ is a minimum-variance unbiased estimator for θ.

(c) Determine the maximum likelihood estimator, $\widehat{\theta}_{\mathrm{MLE}}$, for θ, and show that it is biased.

Now approach this estimation problem as a Bayesian, placing a prior distribution on θ given by

$$\pi(\theta) = \begin{cases} 2\theta & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases} .$$

(d) Determine the posterior distribution, specifically identifying it by name.

(e) Let $\widetilde{\theta} = \mathrm{E}[\theta|X]$ denote the *posterior mean.* Determine the values of the following quantities:

- The posterior variance, $\mathrm{Var}[\theta|X]$.
- The variance of the posterior mean, $\mathrm{Var}[\widetilde{\theta}]$.

Use plain language to describe and interpret what each of these values signifies.

(f) Compare and contrast the three estimators proposed in this problem ($\widehat{\theta}_{\mathrm{MME}}$, $\widehat{\theta}_{\mathrm{MLE}}$, and $\widetilde{\theta}$) in terms of mean squared error. From this vantage point, briefly discuss the relative advantages and disadvantages of each estimator. You may of course use computing resources (e.g., Wolfram Alpha, R) to assist you in this problem.

---

6. $\boxed{\textbf{25 pts}}$ This problem has to do with aggregating information from multiple estimators of an unknown real-valued parameter θ. Suppose you have K independent, unbiased estimators $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$. In particular:

- $\mathrm{E}[\widehat{\theta}_k] = \theta$ for $1 \leq k \leq K$.
- $\mathrm{Var}(\widehat{\theta}_k) = \sigma_k^2$ for $1 \leq k \leq K$.
- $\mathrm{Cov}(\widehat{\theta}_k, \widehat{\theta}_{k'}) = 0$ for $1 \leq k \neq k' \leq K$.

(a) Determine the bias and the variance of the estimator $\overline{\theta} = 0.3 \times \widehat{\theta}_1 + 0.7 \times \widehat{\theta}_2$.

(b) Let $w_1, \ldots, w_K$ denote general real-valued constants. Provide expressions for the bias and the variance of estimators taking the form:

$$\widetilde{\theta} \quad = \quad \sum_{k=1}^{K} w_k \widehat{\theta}_k.$$

(c) Explicitly characterize all combinations of $w_1, \ldots, w_K$ for which $\widetilde{\theta}$ will be unbiased for θ.

(d) Without appealing to any general theorems, explicitly determine the values of $w_1, \ldots, w_K$ in terms of $\sigma_1^2, \ldots, \sigma_K^2$ for which $\widetilde{\theta}$ will have minimal variance among all estimators with $w_1, \ldots, w_K$ chosen as in part (c)—namely, chosen in a way such that $\widetilde{\theta}$ is unbiased for θ. You may find the Lagrange multiplier method to be helpful for this problem.

(e) Consider the special case in which $K = 2$ and $\widehat{\theta}_1, \widehat{\theta}_2$ both follow a normal distribution with a common variance (in addition to being independent and unbiased for θ). Show that $\widehat{\theta}_{(1)} = \min(\widehat{\theta}_1, \widehat{\theta}_2)$ is biased.

Hint: You may utilize the following fact without proof: if $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $X_{(1)} = \min(X_1, X_2)$ has moment generating function given by:

$$M_{X_{(1)}}(t) \quad = \quad 2 \times \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right) \times \Phi\left(-\frac{t\sigma}{\sqrt{2}}\right).$$
