Vanderbilt University Biostatistics Comprehensive Examination

MS Theory Exam/
PhD Theory Exam Series 1

May 23, 2022

Instructions: Please adhere to the following guidelines:

- This exam begins on Monday, May 23 at 9:00am. You will have until 5:00pm to complete it.
- There are six equally weighted problems of varying length and difficulty. Note that not all 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. [25 pts] Consider a research study of a particular disease in mice. Researchers evaluate offspring from two animals; each offspring has a known 25% chance of having the disease of interest. Assume for the purposes of this problem that the disease status is independent across all offspring.

Suppose the first 11 offspring from the breeding pair are evaluated, and let X denote the number of offspring with the disease.

(a) Determine the probability mass function, $p_X(x)$.

(b) Use the definition of expected value to compute $\mathrm{E}[X]$.

(c) Use the definition of variance to compute $\mathrm{Var}[X]$.

Suppose the breeding pair continues to breed and produce offspring. Let Y denote the number of offspring examined (not including the original 11) until the next offspring with the disease is identified.

(d) Determine the moment generating function, $M_Y(t)$, for Y.

(e) Use your answer to part (d) to compute $\mathrm{E}[Y]$.

(f) Use your answer to part (d) to compute $\mathrm{Var}[Y]$.

Assume now that the total number of offspring examined, N, is a random variable, with $N \sim \text{Poisson}(\lambda = 20)$. Assume that, given N, the number of offspring in this sub-study with the disease, Z, follows the same family of distributions as in part (a).

(g) Determine the value of $\mathrm{E}[Z]$.

(h) Determine the value of $\mathrm{Var}[Z]$.

The researchers devise a diagnostic test for the disease of interest based on a continuous biomarker. They now randomly sample $n_0 = 100$ offspring without the disease and $n_1 = 100$ offspring with the disease (these sample sizes are fixed and known). Suppose that at a particular threshold for the continuous biomarker, 95 mice test positive: 20 mice without the disease and 75 mice with the disease.

(i) Determine an estimate of each of the following quantities:

- The sensitivity, P(test positive|disease).
- The specificity, P(test negative|no disease).
- The positive predictive value, P(disease|test positive).
- The negative predictive value, P(no disease|test negative).

(j) The researchers use the data to construct an empirical receiver operating characteristic curve based on the diagnostic test. Briefly explain why the empirical "area under the ROC curve" can be no larger than 0.95 (based on the data provided in the description above). You may feel free to draw a picture to justify your argument—no formal proof is required.

---

2. [25 pts] Suppose X and Y are random variables having joint density function given by:

$$f_{X,Y}(x,y) = x + y, \quad 0 < x < 1, \;\; 0 < y < 1.$$

Further, let $U = X + Y$ and let $V = X - Y$.

(a) Show that $f_{X,Y}(x,y)$ is a valid joint probability density function.

(b) Determine the joint cumulative distribution function of (X,Y).

(c) Determine $\mathrm{P}(Y^{1/3} < X)$.

(d) Determine the value of $\mathrm{Cov}[U,V]$.

(e) Determine the joint density function for (U,V).

(f) Are U and V independent? Justify your response.

For parts (g) and (h), suppose you sample n i.i.d. pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ from $f_{X,Y}(x,y)$.

(g) Let $U_n = \sum_{i=1}^{n}(X_i + Y_i)$. Determine a sequence of positive numbers, a_n, and a real number, b, such that the following statement is true:

$$a_n U_n \xrightarrow{a.s.} b.$$

Justify your steps by naming any theorems you invoke.

(h) Let $V_n = \sum_{i=1}^{n}(X_i - Y_i)$. Determine a sequence of positive numbers, c_n, such that the following statement is true:

$$c_n V_n^2 \xrightarrow{d} \chi_1^2.$$

Justify your steps by naming any theorems you invoke.

---

3. [25 pts] The number of colds per year, X, for a randomly selected resident in a northern region is assumed to have the discrete distribution $p_\theta(x) = \theta^{-1}$, $x = 1, 2, \ldots, \theta$, where the parameter θ is an unknown positive integer. It is desired to find a reasonable candidate for the minimum variance unbiased estimator (MVUE) of θ using the information contained in a random sample $X_1, \ldots, X_n$ from $p_\theta(x)$.

(a) Prove that $U = \max\{X_1, \ldots, X_n\}$ is the maximum likelihood estimator of θ and is a sufficient statistic.

(b) Show that $T = 2X_1 - 1$ is an unbiased estimator of θ.

(c) Given that U is a complete sufficient statistic for θ, derive an explicit expression for the MVUE, $\widetilde{\theta}$, of θ.

(d) Check your work in part (c) by confirming that $\widetilde{\theta} = 2X_1 - 1$ when $n = 1$. Then, show directly that $\mathrm{E}[\widetilde{\theta}] = \theta$ for all n.

(e) Do you notice any undesirable properties of $\widetilde{\theta}$ as an estimator of θ? Explain.

---

4. [25 pts] Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables having common density function

$$f_\lambda(x) \quad = \quad \frac{\lambda}{2} e^{-\lambda|x|}$$

for an unknown parameter $\lambda > 0$.

(a) Suppose you seek a method-of-moments estimator for λ. Very briefly, why would a method-of-moments estimator based on $\mu_1 = \mathrm{E}[X]$ not be an appropriate approach to this problem?

(b) Determine the value of $\mu_2 = \mathrm{E}[X^2]$.

(c) Use your answer to part (b) to propose a method-of-moments estimator for λ (call it $\widetilde{\lambda}_n$), and determine its asymptotic distribution.

(d) Determine the maximum likelihood estimator for λ (call it $\widehat{\lambda}_n$), and determine its asymptotic distribution.

(e) How does the asymptotic efficiency of $\widehat{\lambda}_n$ compare to that of $\widetilde{\lambda}_n$?

---

5. [25 pts] A scientist is studying the teratogenic effects of a certain chemical on rat fetuses. Two fetuses from each pregnant female rat that was exposed to the chemical during gestation were observed. A fetus was determined to be abnormal (i.e., dead or malformed) or normal. Suppose that π is the probability that a fetus is abnormal, $0 < \pi < 1$. Further, for the i^{th} of n litters, each litter being of size two, let the random variable X_{ij} take the value 1 if the j^{th} fetus is abnormal and take the value 0 otherwise, $j = 1, 2$.

Since the two fetuses in each litter have experienced the same gestational conditions, the dichotomous random variables X_{i1} and X_{i2} are expected to be correlated. To allow for such a correlation, the following correlated binomial model is proposed: For $i = 1, 2, ..., n$,

$$\begin{aligned} p_0 &= \mathrm{P}(X_{i1} + X_{i2} = 0) = (1-\pi)^2 + \theta \\ p_1 &= \mathrm{P}(X_{i1} + X_{i2} = 1) = 2(\pi(1-\pi) - \theta) \\ p_2 &= \mathrm{P}(X_{i1} + X_{i2} = 2) = \pi^2 + \theta, \end{aligned}$$

where $-\min\{\pi^2, (1-\pi)^2\} < \theta < \pi(1-\pi)$. Note that by this setup, $\mathrm{Cov}[X_{i1}, X_{i2}] = \theta$. Let Y_k denote the number of litters out of n where k of the two fetuses are abnormal, $k = 0, 1, 2$.

(a) Show that the maximum likelihood estimators $\widehat{\pi}$ of π and $\widehat{\theta}$ of θ are, respectively:

$$\begin{aligned} \widehat{\pi} &= \frac{1}{2} + \frac{Y_2 - Y_0}{2n}, \text{ and} \\ \widehat{\theta} &= \frac{Y_2}{n} - \widehat{\pi}^2 \end{aligned}$$

(b) Derive expressions for $\mathrm{E}[\widehat{\pi}]$ and $\mathrm{Var}[\widehat{\pi}]$.

(c) The scientist wants to test the null hypothesis that normal and abnormal fetuses are equally likely against the general alternative of different likelihoods. State the null hypothesis in terms of π. State the null hypothesis in terms of p_k, $k = 0, 1, 2$.

(d) Determine a likelihood ratio test statistic and describe specifically how you would use it to conduct an α-level hypothesis test.

(e) Another test is based on

$$Z = \frac{\widehat{\pi} - \pi_0}{\sqrt{\widehat{\mathrm{Var}}[\widehat{\pi}]}},$$

and rejects H_0 if $|Z| > z_\alpha$. Argue that this is equivalent to the Wald test.

(f) Under what conditions would you expect the tests of parts (d) and (e) to have type 1 error rates of approximately α?

---

6. [25 pts] Let $x_1, \ldots, x_n$ denote a fixed and known sequence, with $1/c \leq x_1, \ldots, x_n \leq c$ for a known $c \geq 1$ that does not depend upon n. Suppose $Y_1, \ldots, Y_n$ are i.i.d., with $Y_i \sim \text{Poisson}(x_i\theta)$. The goal is to estimate the fixed, unknown parameter $\theta > 0$.

(a) Obtain the maximum likelihood estimator, $\widehat{\theta}_n$, for θ and show that it is unbiased for θ.

In parts (b) and (c), consider Bayesian estimation of θ under the prior $\theta \sim \text{Gamma}(\alpha, \beta)$. To be clear, the prior density function is given by:

$$\pi(\theta) \quad = \quad \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta},$$

with prior mean given by $\text{E}[\theta] = \alpha/\beta$.

(b) Determine (and name) the posterior distribution $\pi(\theta|Y_1, \ldots, Y_n)$.

(c) Determine the posterior mean, $\text{E}[\theta|Y_1, \ldots, Y_n]$. Show that it can be expressed as a convex combination of the prior mean and the maximum likelihood estimator in the following way:

$$\text{E}[\theta|Y_1, \ldots, Y_n] \quad = \quad w_n \times \text{E}[\theta] + (1 - w_n) \times \widehat{\theta}_n,$$

Specifically determine the weights, w_n, as part of your response.

Now, consider the posterior mean, $\text{E}[\theta|Y_1, \ldots, Y_n]$, as an estimator of θ (call it $\widetilde{\theta}_n$) under the frequentist paradigm.

(d) Argue that $\widetilde{\theta}_n$ is not, in general, unbiased for θ.

(e) Argue that $\widetilde{\theta}_n$ is consistent for θ.

(f) Suppose the only conditions imposed on the fixed, known values of $x_1, \ldots, x_n$ is that they are positive and finite. Argue that these conditions are insufficient to guarantee consistency of $\widetilde{\theta}_n$ (i.e., for θ). For this problem, all that is required is to determine a sequence of positive numbers, $x_1, \ldots, x_n$, such that consistency does not hold.
