Vanderbilt University Biostatistics Comprehensive Examination

MS Theory Exam/
PhD Theory Exam Series 1

May 22, 2023

Instructions: Please adhere to the following guidelines:

- This exam begins on Monday, May 22 at 9:00am. You will have until 1:00pm 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. [25 pts] Suppose $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ for $i = 1, \ldots, n$.

(a) State the values of $\text{E}[X_i]$ and $\text{Var}[X_i]$.

(b) Identify the distribution of $Y_n = \sum_{i=1}^{n} X_i$ by name. Determine $\text{E}[Y_n]$ and $\text{Var}[Y_n]$ using the values of $\text{E}[X_i]$ and $\text{Var}[X_i]$ you determined in part (a); justify your response.

(c) Determine the large-sample distribution of $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Justify your response.

(d) Determine the large-sample distribution of $\log(\overline{X}_n + 1)$. Justify your response.

(e) Which of the following statements is true? Justify your response.

i. $\text{E}[\log(\overline{X}_n + 1)] \leq \log(\text{E}[\overline{X}_n] + 1)$.

ii. $\text{E}[\log(\overline{X}_n + 1)] \geq \log(\text{E}[\overline{X}_n] + 1)$.

iii. It is not possible to determine which of the above is true without further information.

(f) Prove that the following statement holds if $p = 1/2$:

$$4n\left(\frac{1}{4} - \overline{X}_n(1 - \overline{X}_n)\right) \xrightarrow{d} \chi^2_1.$$

(g) Suppose now that $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_1)$ for $i = 1, \ldots, n$ and $Y_j \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_2)$ for $j = 1, \ldots, m$, with all X_i's and Y_j's independent. Let $\widehat{p}_1$ denote the sample mean for the X_i's and $\widehat{p}_2$ the sample mean for the Y_j's. Determine the large-sample distribution of the sample log-odds ratio:

$$\log\left(\frac{\frac{\widehat{p}_1}{1-\widehat{p}_1}}{\frac{\widehat{p}_2}{1-\widehat{p}_2}}\right) = \log\left(\frac{\widehat{p}_1(1-\widehat{p}_2)}{\widehat{p}_2(1-\widehat{p}_1)}\right).$$

---

2. [25 pts] Suppose that $P \sim \text{Beta}(\alpha, \beta)$, with probability density function given by:

$$f_P(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 < p < 1; \quad \alpha > 0, \beta > 0.$$

Further suppose that, given $P = p$, X follows a NegativeBinomial(r, p) distribution, having probability mass function (PMF) given by:

$$p_{X|P}(x; r, p) = \binom{x+r-1}{x} (1-p)^x p^r, \quad x = 0, 1, 2, \ldots; \quad r = 1, 2, 3, \ldots$$

(a) Show explicitly that $\sum_{i=0}^{\infty} \text{P}(X = i|P = p) = 1$ for $r = 1$.

(b) Use the definition of expectation to determine $\text{E}[X|P]$ for $r = 1$.

(c) Use the fact that $\text{Var}[X|P] = \text{E}[X(X-1)|P] + \text{E}[X|P] - \text{E}[X|P]^2$ to determine $\text{Var}[X|P]$ for $r = 1$.

(d) Suppose that, given $P = p$, Y_1 and Y_2 are independent NegativeBinomial(1, p) random variables. Derive the conditional PMF of $S_2 = Y_1 + Y_2$ (given $P = p$) and verify that it is that of a NegativeBinomial$(2, p)$ distribution.†

(e) Determine the value of $\text{E}[X]$ for any positive integer r; state the values of α and β for which this expectation is defined.

(f) Determine the value of $\text{Var}[X]$ for any positive integer r; state the values of α and β for which this variance is defined.

(g) Justifying your steps, determine the marginal PMF of X, $p_X(x)$.

---

† More generally, it is the case that $S_r = Y_1 + \cdots + Y_r \sim$ NegativeBinomial(r, p) when $Y_1, \ldots, Y_r$ are independent NegativeBinomial$(1, p)$ random variables; you are free to use this fact later in the problem without proof.

3. [25 pts] Suppose you obtain four independently sampled observations, X_1, X_2, X_3, and X_4, shown below:

i	X_i
1	1
2	2
3	5
4	20

You propose a parametric model for X based on the following density function:

$$f_X(x;\theta) \quad = \quad \frac{1}{2}\theta^3 x^{-4}\exp(-\theta/x)\mathbf{1}_{(0,\infty)}(x),$$

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

(a) Determine the maximum likelihood estimate (MLE) of θ based on the four observations (call it $\widehat{\theta}$).

(b) Determine a 95% confidence interval for θ of the form $\widehat{\theta} \pm 1.96 \times \widehat{\mathrm{SE}}(\widehat{\theta})$, where $\widehat{\mathrm{SE}}(\widehat{\theta})$ is based on an asymptotically valid approximation.

(c) Briefly describe your concerns about the validity of the confidence interval you formed in part (b).

(d) Consider placing the following prior distribution on θ:

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

where α and β are known parameters (i.e., set by you, the researcher). Determine the posterior density, $\pi(\theta|X_1, X_2, X_3, X_4)$, and identify it by name.

(e) Show explicitly—i.e., without appealing to a general theorem that would suggest as such—that the posterior mean can be expressed as a weighted average of the prior mean and the MLE, $\widehat{\theta}$.

(f) Consider two choices of a prior, each belonging to the family described in part (d): π_1 (based on $\alpha = 1$ and $\beta = 1/10$), and π_2 (based on $\alpha = 20$ and $\beta = 2$). Compute the posterior means under each choice of a prior; call them $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$.

(g) Note that the prior means are equivalent under π_1 and π_2. Briefly explain why the posterior means $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ nevertheless disagree.

(h) Your collaborator argues that since π_1 and π_2 have the same mean, π_1 is less informative about the value of θ because it has higher variance. Briefly comment on the most serious limitations of this argument.

---

4. 25 pts Certain types of measurements (e.g., biological concentrations) tend to exhibit greater variation for high values than for low values. In many settings, it is reasonable to assume that the standard deviation of a measure is proportional to the mean, suggesting the use of models that presume a constant *coefficient of variation*. The coefficient of variation is the ratio of the measure standard deviation to the measure mean:

$$C_v = \frac{\text{SD}[Y]}{\text{E}[Y]}.$$

Under such a model, C_v serves as a measure of precision. The U.S. Food and Drug Administration recommends an $m : n : \tau$ procedure for validating an assay's precision: m different concentration levels are each measured using n independent samples, and if the sample coefficient of variation for each of the m levels is less than some fixed and known threshold, τ, then the assay passes the precision validation. When working with a measurement system where you anticipate C_v to be constant, there are advantages to analyzing log-transformed data instead of the raw values (many statistical tests require error variation across the values of a predictor to be independent of the value of the predictor, and the assumption of a constant C_v suggests that this holds on the log-scale).

Let Y be a random variable denoting a single measure of an assay. Assume that $Y \sim \text{log-normal}(\mu, \sigma^2)$, which is to say that $\log(Y) \sim \mathcal{N}(\mu, \sigma^2)$.

(a) Derive the PDF, $f_Y(y; \mu, \sigma^2)$, of Y.

(b) Determine $\text{E}[Y]$ and $\text{Var}[Y]$. You may recall that $X \sim \mathcal{N}(\mu, \sigma^2)$ has moment-generating function $M_X(t) = \exp(t\mu + \sigma^2 t^2/2)$, a fact you are permitted to use without further proof.

(c) Confirm that $C_v = \sqrt{\exp(\sigma^2) - 1}$, and express σ^2 as a function of C_v.

Let θ_j denote the true C_v for the j^{th} concentration, and assume $\theta_1 = \cdots = \theta_m$ so that we are in the setting of constant C_v. Let $\widehat{\theta}_j$ be the j^{th} sample C_v estimate. If $\widehat{\theta}_j < \tau$ for all $j = 1, \ldots, m$, then the assay passes the $m : n : \tau$ validation procedure, otherwise it fails. We first consider the statistical properties of passing the j^{th} level. Let $\widehat{\theta}_j < \tau$ be a decision rule for a hypothesis test of $H_{0j} : \theta_j \geq \tau$ versus $H_{1j} : \theta_j < \tau$.

(d) Suppose $X_i = \log(Y_i)$ where $Y_1, \ldots, Y_n$ is an i.i.d. sample from log-normal(μ, σ^2).

  i. Determine the MLE of C_v. Label this $\widehat{C}_v$.

  ii. A test of the stated hypotheses uses the rejection region $\{\mathbf{y} : \widehat{C}_v < \tau\}$. Determine the size of this test and express it in a way such that it could be implemented easily in statistical software. Label this α_1.

  iii. Determine the power of this test and express it in a way such that it could be implemented easily in statistical software.

(e) In practice, it is more common to use the sample variance, S^2, of the X's to estimate σ^2 and then use this to estimate C_v. Label this estimator $\widetilde{C}_v$. An alternate test of the stated hypotheses uses the rejection region $\{\mathbf{y} : \widetilde{C}_v < \tau\}$. Describe how you would compute the size of this test in a way that it could be implemented easily in statistical software. Label this α_2.

(f) The tests proposed in parts (d) and (e) apply to the n replicates of one sample concentration level.

  i. Let $\widehat{C}_{vj}$ be the estimator of C_v for the j^{th} level where $j = 1, 2, \ldots, m$. What is the size of the test with rejection region $\left\{\mathbf{y} : \max\left(\widehat{C}_{v1}, \ldots, \widehat{C}_{vm}\right) < \tau\right\}$? Label this $\ddot{\alpha}_1$.

  ii. What is size of the test with rejection region $\left\{\mathbf{y} : \max\left(\widetilde{C}_{v1}, \ldots, \widetilde{C}_{vm}\right) < \tau\right\}$? Label this $\ddot{\alpha}_2$.

(g) One choice for $m : n : \tau$ is $3 : 5 : 0.15$, in which $\alpha_1 = 0.713, \ddot{\alpha}_1 = 0.362, \alpha_2 = 0.594, \ddot{\alpha}_2 = 0.210$. Noting that each replica is costly and sampling levels are invasive, costly, and time consuming, briefly discuss the pros and cons of these tests.
