Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
August 29, 2023
12:30-4:30pm, Room 331 SMI

- There are a total of FOUR (4) problems in this exam. Please do all FOUR (4) problems.
- Each problem must be done in a separate exam book.
- Please turn in FOUR (4) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. **Read all parts of the question carefully before starting.**

   Suppose $X_1, \ldots, X_n$ is a random sample from the Geometric(θ) distribution, which has probability mass function (pmf)

$$p(x;\theta) = \theta(1-\theta)^{x-1}; \quad x = 1, 2, \ldots; \quad 0 < \theta < 1.$$

   This question concerns estimation of the parameter $\eta = \theta(1-\theta)$ based on $X_1, \ldots, X_n$. Solve the following problems, showing all your work:

   (a) Find the uniformly minimum variance unbiased estimator (UMVUE) of η.

   (b) Show that the Cramér-Rao lower bound (CRLB) for the variance of unbiased estimators of η is given by $\frac{\theta^2(1-\theta)(1-2\theta)^2}{n}$. Argue that no unbiased estimator of η achieves the bound for all θ.

   (c) Find $\hat{\eta}_{MLE}$, the maximum likelihood estimator (MLE) of η.

   (d) Derive the non-degenerate asymptotic distribution of $\hat{\eta}_{MLE}$ for the case $\theta = 1/2$. Use the result to calculate the asymptotic variance of $\hat{\eta}_{MLE}$ when $\theta = 1/2$. (**HINT:** use the second order delta method.)

   (e) Suppose θ has a Beta(α, β) prior distribution. Find the Bayes rule for η under squared error loss.

   You may find the following notes helpful:

   1. Suppose $Y \sim$ Geometric(θ). Then $E(Y) = \frac{1}{\theta}$ and $Var(Y) = \frac{1-\theta}{\theta^2}$.

   2. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed Geometric(θ) random variables. Let $T = \sum_{i=1}^{n} Y_i$. Then:

$$Pr(T = t) = \binom{t-1}{n-1}\theta^n(1-\theta)^{t-n}, \quad t = n, n+1, \ldots;$$

      i.e., T follows a *negative binomial* distribution.

   3. The Beta(α, β) distribution has probability density function

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

   4. Let χ^2_ν represent a chi-square random variable on $\nu > 0$ degrees of freedom. Then $E(\chi^2_\nu) = \nu$ and $Var(\chi^2_\nu) = 2\nu$.

2. Let $X_1, \ldots, X_n$ be i.i.d. random variables each with common probability density (or frequency) function $f(x;\theta) = h(x)\exp[\theta x - \psi(\theta)]$. Suppose $\theta \in \Theta \subseteq \mathbb{R}$, where the parameter space Θ is an open interval. Answer the following questions and justify each answer.

   (a) Determine a complete sufficient statistic for θ.

   (b) Compute $\mathrm{E}_\theta[X_1]$ and $\mathrm{Var}_\theta[X_1]$.

   (c) Define $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Obtain the non-degenerate limiting distribution of $\overline{X}_n$, suitably normalized.

   (d) Obtain two different estimators for θ based on the method of moments.

   (e) Obtain the maximum likelihood estimator $\widehat{\theta}_n$ for θ.

   (f) Obtain the non-degenerate limiting distribution of $\widehat{\theta}_n$, suitably normalized.

   (g) Determine the uniformly most powerful test for $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$.

3. Researchers observe a sample of variables $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}$ for $i = 1, \ldots, 100$, and obtain the following summary statistics:

$$\left.\begin{array}{llll} \sum_{i=1}^{50} x_i = 50, & \sum_{i=51}^{100} x_i = 0, & \sum_{i=1}^{50} y_i = 26, & \sum_{i=51}^{100} y_i = 6, \\ \sum_{i=1}^{50} x_i^2 = 200, & \sum_{i=51}^{100} x_i^2 = 50, & \sum_{i=1}^{50} y_i^2 = 60, & \sum_{i=51}^{100} y_i^2 = 63, \\ \sum_{i=1}^{50} x_i^3 = 1100, & \sum_{i=51}^{100} x_i^3 = 900, & \sum_{i=1}^{50} y_i^3 = 310, & \sum_{i=51}^{100} y_i^3 = 209, \\ \sum_{i=1}^{50} x_i y_i = -29, & \sum_{i=51}^{100} x_i y_i = -21. & & \end{array}\right\} \quad (1)$$

(a) Researchers fit the data by the working linear regression model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{for } i = 1, \ldots, 50, \\ y_i &= 4\beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{for } i = 51, \ldots, 100, \end{aligned} \quad (2)$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, 100$ with $\overset{i.i.d.}{\sim}$ representing independent and identically distributed. Give the maximum likelihood estimator of $\hat{\beta}_0$ and $\hat{\beta}_1$ under the model (2), and calculate their values based on summary statistics in (1).

(b) Assume (2) is the true model that generates the data $\{(y_i, x_i) : i = 1, \ldots, 100\}$.

(i) Specify an unbiased estimator of σ^2, and calculate its value using the summary statistics in (1).

(ii) Researchers want to test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. Provide a test statistic, calculate its value using (1), fully specify the distribution of the test statistic under $H_0 : \beta_1 = 0$, and describe the testing procedure.

(c) Assume the true model that generates the data $\{(y_i, x_i) : i = 1, \ldots, 100\}$ is

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \epsilon_i, \quad \text{where } z_i = -r x_i^2 + w_i, \quad (3)$$

where $w_i \overset{i.i.d.}{\sim} N(0, 1)$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$, all w_i's and ϵ_i's are jointly independent, and $\alpha_0, \alpha_1, \alpha_2, \sigma^2$, and r are unknown parameters.

(i) Are $\hat{\beta}_0$ and $\hat{\beta}_1$ in (a) unbiased for estimating α_0 and α_1? If yes, prove it. If not, provide the formula of the bias based on the summary statistics in (1).

(ii) Derive $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1)$ based on summary statistics in (1).

(iii) Given $x_{new} \in \mathbb{R}$, researchers define $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$ as a predictor to use. Suppose (x_{new}, y_{new}) is generated following the true model (3). Specifically, $y_{new} = \alpha_0 + \alpha_1 x_{new} + \alpha_2 z_{new} + \epsilon_{new}$, where $z_{new} = -r x_{new}^2 + w_{new}$, $w_{new} \sim N(0, 1)$, and $\epsilon_{new} \sim N(0, \sigma^2)$. Moreover, z_{new}, w_{new}, z_i's, and w_i's for $i = 1, \ldots, n$ are jointly independent. When $x_{new} = 1$, derive the expected squared prediction error $E(y_{new} - \hat{y})^2$ based on the summary statistics in (1).

You may find the following notes helpful:

(i) All expectations are taken conditioning on the observed x_i's.

(ii) 2×2 matrix inversion: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

4. In a controlled laboratory environment, researchers examined the impact of a certain cognitive-enhancing drug on human memory performance. The drug was assigned to a group of 100 participants at four distinct dosage levels: 0 mg, 2 mg, 4 mg, and 10 mg. Participants were randomly divided into 20 groups, with 5 groups to each dosage level. Participants took memory tests before and after the experiment, and their performances were assessed based on changes in cognition scores. Let y_{ij} denote the average score changes of the jth group that received x_i mg of the drug dosage. Let γ be the expected value of average score changes for groups receiving 4 mg of drug; that is, $\gamma = E(y_{3j})$ for $j = 1, \ldots, 5$. Researchers are interested in estimating γ.

   The researchers employed the following model for data analysis:

$$y_{ij} = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_{ij}, \quad i = 1, \ldots, 4, \quad j = 1, \ldots, 5, \tag{4}$$

   where $\varepsilon_{ij} \sim_{\text{i.i.d.}} N(0, \sigma^2)$ are independent and identically distributed random variables, $\bar{x} = 4$ represents the average dosage of the drug, and $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters.

   (a) Derive an estimator for γ based on model (4). Express your estimator in terms of y_{ij}.

   For the remainder of the problem, assume the ground-truth generative model between averaged cognition score and drug dosage is given by:

$$y_{ij} = \mu_i + \varepsilon^0_{ij}, \quad i = 1, \ldots, 4, \quad j = 1, \ldots, 5, \tag{5}$$

   where the parameter $(\mu_1, \mu_2, \mu_3, \mu_4) = (120, 150, 180, 274)$ represents the drug effect, and $\varepsilon^0_{ij} \sim_{\text{i.i.d.}} N(0, \sigma_0^2)$ are independent and identically distributed random variables with $\sigma_0 = 6$. In the remainder of the problem, you will evaluate the performance of researchers' working model (4) under the ground-truth model (5).

   (b) Find the variance of the researchers' estimator for γ in part (a), considering that the data is generated from model (5).

   (c) Find the mean squared error of the researchers' estimator for γ in part (a), considering that the data is generated from model (5).

   (d) Assume that the researchers now fit model (5) to their data; i.e., the researchers estimate μ_i and σ_0 in model (5), both of which are unknown to them. Find the mean squared error for their estimator of γ from this model. Compare your answer with part (c). Which estimator has a smaller mean squared error?

   (e) Construct an F-test for evaluating the lack of fit of model (4) against the ground-truth model in (5). Specify the null and alternative hypotheses, and determine the distributions of the test statistic. In your answers, please plug in the specific values to **fully** describe the distributions of the test statistics under the null and under the alternative, respectively.