UNL Statistics PhD Qualifying Exam - May 2024

Print Your Qualifying Exam ID: ____________________

- **Date: May 23th (8:00am – 1:00pm, Thursday) and May 24th, 2024**
- **Location: Room 49 Hardin Hall (Subject to change)**
- **Please put your phones on vibrate/silent mode and don't use during the exam.**
- The exam is a written exam over the MS core course that assesses preparedness for the PhD program. Students are allowed to take the exam if they have a GPA of at least 3.5 in their MS and PhD core courses taken.
- You will be assigned an ID number.
- Closed book and note. You will not have access to computers or the Internet. You may not use your phones for any reason. Bring a calculator.
- Your answers will be hand-written. We will provide you with paper. Please ensure that your answers are written as clearly as possible.
  - The assigned ID number must be written on each page of the answerscript in place of your name.
  - Start each problem on a new sheet of paper (you do not need to start each part of each problem with a new sheet).
    - (Ex.) Problem 1, parts 1-5 should be numbered: 1.1, 1.2, 1.3 . . . etc.
    - Example of problems with lettered parts and subsections: 6.a, 6.b...6.e(a), 6.e(b), 6.e(c) . . . etc.
  - When you are finished, number **all** pages of work in the bottom right corner 1 – x pages.
  - All answers must be written only on one side of the paper.

Day 1

1. (100 points) Let $X_1, X_2, \cdots, X_n$ be an independent and identically distributed (IID) sample from Uniform$(0, 1)$. Let $X_{(1)}$ be the smallest order statistic.

   1. (20 points) Find the distribution of $X_{(1)}$ and compute $E\left(X_{(1)}\right)$.
   2. (20 points) Show that $X_{(1)} \to 0$ in probability, further show that $nX_{(1)}$ converges in distribution to an Exponential random variable.
   3. (20 points) Let $Y_i = -\log\left(X_i\right), i = 1, \ldots, n$. Show that $Y_1, \ldots, Y_n$ are IID Exponential(1) distributed.
   4. (20 points) Find the limiting distribution of $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$, i.e., identify the distribution with parameters that $\bar{Y}_n$ converges to when $n \to \infty$, and explain why the convergence holds.
   5. (20 points) Find the limiting distribution of $(\bar{Y}_n)^2$ and specify the parameters of the distribution. Explain why it follows such a distribution.

2. (100 points) Suppose $X_1, X_2, \ldots, X_n$ are IID random variables with the probability density function (pdf)

$$f(x \mid \theta) = e^{-(x-\theta)}, \qquad x \geq \theta.$$

Let $X_{(1)} = \min\limits_{1\leq j\leq n} X_j$ and $X_{(n)} = \max\limits_{1\leq j\leq n} X_j$. You may use the fact that $X_{(1)}$ is a complete statistic for the following questions.

1. (20 points) Provide with justification a minimum sufficient statistic for θ.
2. (20 points) Show that $X_{(n)} - X_{(1)}$ is an ancillary statistic and show that $X_{(1)}$ and $X_{(n)} - X_{(1)}$ are independent
3. (20 points) Does $f(x \mid \theta)$ belong to the exponential family? Justify your answer.
4. (20 points) Derive the method-of-moments estimator (MOM) of $\theta, \hat{\theta}_{MOM}$.
5. (20 points) Find the cumulative distribution function of $X_{(1)}$.

3. (100 points) Continued on Problem 2.

   1. (20 points) Find the maximum likelihood estimator (MLE) of $\theta, \hat{\theta}_{\text{MLE}}$.
   2. (20 points) Show that the MLE of θ is consistent.
   3. (20 points) Find the mean squared errors (MSEs) of the two estimators, $\hat{\theta}_{MOM}$ and $\hat{\theta}_{MLE}$. Which estimator is better in terms of MSE?
   4. (20 points) Find an unbiased estimator that is better than both $\hat{\theta}_{MOM}$ and $\hat{\theta}_{MLE}$ in terms of MSE.
   5. (20 points) Find the UMVUE of θ and explain why it's the UMVUE in detail.

4. (100 points) Recall that, for $\alpha > 0$ and $\beta > 0$, a Beta(α, β) random variable Z has the pdf, mean and variance:

$$f_Z(z) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1} & 0 < z < 1 \\ 0 & \text{otherwise,} \end{cases} \qquad E(Z) = \frac{\alpha}{\alpha+\beta},\ Var(Z) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The joint density function of the random variables X and Y is

$$f_{X,Y}(x, y) = 20 \binom{n}{y} x^{y+3}(1-x)^{n-y+1}, \qquad \text{for } y = 0, 1, \ldots, n \text{ and } 0 < x < 1, \quad (1)$$

and $f_{X,Y}(x, y) = 0$ otherwise. Here, n is a fixed integer such that $n > 4$.

1. (20 points) Find $f_Y(y)$ and $f_X(x)$. What's the distribution of X?
2. (20 points) Find the conditional expectation of Y given that $X = x$ and compute the expectation of Y.
3. (20 points) What is the conditional expectation of X given that $Y = 4$?
4. (20 points) Use the Law of total expectation

$$E(X) = E(E(X \mid Y))$$

to prove the Law of total covariance

$$Cov(X_1, X_2) = E(Cov(X_1, X_2 \mid Y)) + Cov(E(X_1 \mid Y), E(X_2 \mid Y)).$$

5. (20 points) Suppose that the pairs (X_1, Y) and (X_2, Y) are each jointly distributed as above pdf and that $X_1 \mid Y$ and $X_2 \mid Y$ are independent. Find $Cov(X_1,\ X_2)$.

5. (100 points) Let $X_1, X_2, \ldots, X_n$ be a random sample from a exponential distribution with the pdf $f(x) = \theta^{-1} \exp(-x/\theta)$, where $\theta > 0$ and $x \in (0, +\infty)$.

    1. (60 points) Derive the uniformly most powerful (UMP) level α test for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are known constants such that $0 < \theta_0 < \theta_1$. To gain full credit, you must explicitly justify every step of your derivation (i.e., cite theorems and any necessary results). You must also specifically identify the rejection region that corresponds to the value of α and the power function.

    2. (40 points) Is the test you developed above also a UMP level α test for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$? θ_0 is a known constant. You must justify your assertion.

6. (100 points) A certain bioinsecticide is used to control insects that live in soil and damage the roots of plants. This bioinsecticide consists of spores of a certain fungus that are suspended in a solution. After they are introduced into the soil, the spores develop into fungal colonies that infect and destroy the insects. The object of one study was to examine how different levels of irrigation affect the viability of spores in the soil.

   Ten different fields were used in this study. These fields were randomly selected from a large set of fields that could have been used in the study. Each field was subdivided into three plots of equal size. Each plot was sprayed with a concentration of the bioinsecticide that deposited about 5×10^8 spores per $10cm^2$ of surface area. One plot in each field was irrigated with $2cm$ of water per day. Another plot in each field was irrigated with $1.0cm$ of water per day. The remaining plot in each field received no irrigation. The three levels of irrigation were randomly assigned to the plots with a different randomization within each field. Soil samples were taken from each plot at $1, 5, 10,$ and 20 days after the plot was sprayed with the bioinsecticide, and the number fungal colonies (Y) was measured for each soil sample.

   Consider the model

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \delta_{ij} + \tau_k + \gamma_{jk} + \epsilon_{ijk}$$

   where
   Y_{ijk} is an observation on the number of fungal colonies in a soil sample
   α_j a corresponds to the j^{th} level of irrigation
   τ_k corresponds to the k^{th} sampling date
   γ_{jk} corresponds to the j^{th} level of irrigation and the k^{th} sampling date
   β_i is a random field (block) effect with $\beta_i \sim NID(0, \sigma_f^2)$
   δ_{ij} is a random effect with $\delta_{ij} \sim NID(0, \sigma_p^2)$
   ϵ_{ijk} is a random effect with $\epsilon_{ijk} \sim NID(0, \sigma_e^2)$
   and β_i, γ_{jk}, and ϵ_{ijk} are mutually independent.

   a. Write out the model in matrix notation. Be specific about the dimensions of each term. What is E[Y]? What is Var[Y]?

   b. An analysis of variance table for this model is presented in the Appendix. Report the appropriate degrees of freedom and use these results to obtain formulas for estimates of the variance components.

   c. For the proposed model, show how you would test the null hypothesis that there is no interaction between the levels of the irrigation factor and time (sample dates). Give a formula for your test statistic and explain how you would use it to make an inference about the potential interaction between levels of irrigation and time.

   d. Suppose the researchers want to compare the results at 20 days for the irrigation levels correspond to 1 cm of water per day and 2 cm of water per day. Averaging across fields, let $\bar{Y}_{.24} = \frac{1}{10}\sum_{i=10}^{10} Y_{i24}$ denote the average of the observations at 20 days using 1 cm of water per day, and let $\bar{Y}_{.34} = \frac{1}{10}\sum_{i=10}^{10} Y_{i34}$ denote the average of the observations at 20 days using 2 cm of water per day. Find the variance of $\bar{Y}_{.24} - \bar{Y}_{.34}$ and show how to construct a 95% confidence interval for the difference in the mean results at 20 days for irrigation levels of 1cm and 2cm of water per day.

e. Which of the following are true statements about the proposed model? Circle any statement that is true. For each of them, provide explanation why they are true/false.

   (a) Each observation has the same variance.

   (b) Any two observations taken in the same field have the same level of correlation.

   (c) Two observations taken from two different fields are uncorrelated.

   (d) The REML estimates of the variance components are equal to the estimates of the variance components that you were asked to derive in part (a).

   (e) $(\alpha_2 + \gamma_{24}) - (\alpha_3 + \gamma_{34})$ is estimable.

Appendix

Sources	DF	Expected MS
Fields		$\sigma_e^2 + 4\sigma_p^2 + 12\sigma_f^2$
Plots within Fields		$\sigma_e^2 + 4\sigma_p^2$
Error		σ_e^2

7. (100 points) State the Gauss-Markov theorem and prove it. Please be consistent with your notation.

8. 1. (25 points) Suppose (x_1, x_2, x_3) follow a multivariate normal distribution with mean (μ_1, μ_2, μ_3) and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & 0 \\ \rho^2 & 0 & 1 \end{bmatrix}$$

show that the conditional distribution of (x_1, x_2) given x_3 has mean vector $[\mu_1 + \rho^2(x_3 - \mu_3), \mu_2]^T$ and covariance matrix:

$$\begin{bmatrix} 1-\rho^4 & \rho \\ \rho & 1 \end{bmatrix}.$$

2. (25 points) If $x \sim N_p(\mu, \Sigma)$ random variables and $Q\Sigma Q^T$ $(q \times q)$ is non-singular, then, given $Qx = q$, show that the conditional distribution of X is normal with $\mu + \Sigma Q^T(Q\Sigma Q^T)^{-1}(q - Q\mu)$ and covariance matrix $\Sigma - \Sigma Q^T(Q\Sigma Q^T)^{-1}Q\Sigma$.

9. Consider two p-variate normal populations $\Pi_1 \;:\; N_p(\mu_1, \Sigma)$ and $\Pi_2 \;:\; N_p(\mu_2, \Sigma)$. Assume that the parameters are known. Suppose x comes from Π_1.

   1. (25 points) Let $h(x)$ be the discriminant function (i.e. the boundary) obtained from linear discriminant analysis with equal prior and equal cost of misclassification. Find the distribution of $h(x)$.
   2. (25 points) What is the probability of misclassifying x to Π_2 using the above linear discriminant analysis rule?