# UNL Statistics PhD Qualifying Exam - January 2024

**Print Your Qualifying Exam ID:** _____

## Day 1

1. Let $X_1$, $X_2$, $X_3$, ... be an infinite sequence of independent and identically distributed $Bernoulli(p)$ random variables for $p \in (0, 1)$. For some $\lambda > 0$, let $n$ be a $Poisson(\lambda)$ random variable, which is independent of the Bernoulli sequence $\{X_i\}_{i=1}^{\infty}$. Let

$$S_n = \sum_{i=1}^{n} X_i$$

which is the random variable of interest here.

(a) Derive $E(S_n)$ and $Var(S_n)$.

(b) Derive the moment generating function (MGF) of $S_n$.

(c) Use the MGF obtained above to verify your answers in (a).

(d) Establish using the MGF that

$$\frac{S_n - E(S_n)}{\sqrt{\lambda}} \to N(0, \eta^2) \text{ in distribution, as } \lambda \to \infty$$

and identify $\eta^2$.

2. Let $X_1$, $X_2$, ..., $X_n$ be iid random variable with pdf

$$f(x \mid \theta) = \begin{cases} \frac{5x^4}{\theta^5} & 0 < x < \theta \\ 0 & else \end{cases}$$

Recall that the pdf of the $j^{th}$ order statistics is given by:

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{n-j}$$

(a) Show that $X_{(n)}$ is sufficient for $\theta$.

(b) Is $X_{(n)}$ minimal sufficient? Prove or disprove.

(c) Find the cdf and pdf of $X_{(n)}$.

(d) Find the maximum likelihood estimator of the probability that the largest observation does not exceed 5.

(e) Derive the likelihood ratio test of size $\alpha$ for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Find the finite sample rejection rule for the null hypothesis in terms of $\alpha$. Do not use the large sample approximation.

(f) Derive a $1 - \alpha$ pivotal confidence interval for $\theta$.

(g) Let $\pi(\theta)$ be a prior distribution for $\theta$. Assume that,

$$\int_a^\infty \frac{1}{\theta^k} \pi(\theta) d\theta = \frac{a^{-k+1}}{k-1} \text{ for } k = 1, 2, \ldots$$

Using this assumption, find the joint distribution of $\theta$ and $X_{(n)}$, the marginal distribution of $X_{(n)}$, and the posterior distribution of $\theta$ given $X_{(n)}$.

(h) Find a Bayes estimator of $\theta$.

3. Let $X_1, X_2, \ldots, X_n$ be iid *exponential*$(\theta)$ random variables with density,

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & o.w. \end{cases}$$

(a) Find the UMVUE for $\theta$.

(b) Is the UMVUE a consistent estimator of $\theta$? Prove or disprove.

(c) Find a UMP test of size $\alpha$ for testing $H_o \ : \ \theta \leq \theta_0$ versus $H_1 \ : \ \theta > \theta_0$. Be sure to determine the cut-offs in terms of $\alpha$.

(d) Find the power function of the test in part (c).

4. A system comprises two components, each with independent lifetimes denoted by $X_1$ and $X_2$. These lifetimes follow the exponential distribution with parameters $\lambda_1$ and $\lambda_2$, respectively.

   (a) The system fails if either component fails, meaning the system's lifetime ($Y$) is defined as $Y = \min(X_1, X_2)$. Find the expected lifetime of the system, i.e., $\mathbb{E}(Y)$.

   (b) The system is functional if either component works, meaning the system's lifetime is defined as $Y = \max(X_1, X_2)$. Find the expected lifetime of the system, i.e., $\mathbb{E}(Y)$.

5. As part of an investigation of toxic agents, 48 rates were allocated evenly to 3 poisons $(I, II, III)$ and 4 treatments $(A, B, C, D)$. The response was survival times in tens of hours. Co-factors age (in days since birth) and sex of the rat $(0 = female,\ 1 = male)$ were also recorded for each rat. The following table shows a sample of eight rats (from a total of 48 rats):

| time | poison | treat | age | sex |
|---|---|---|---|---|
| 0.23 | III | A | 34 | 1 |
| 0.43 | I | A | 23 | 1 |
| 0.61 | II | B | 26 | 0 |
| 0.30 | III | B | 28 | 0 |
| 0.45 | I | A | 21 | 0 |
| 0.31 | III | D | 24 | 1 |
| 0.33 | III | D | 25 | 1 |
| 0.76 | I | C | 28 | 1 |

(a) Assume, we are interested in the model of the form:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma x_{ijkl} + \delta_k + \epsilon_{ijkl} \tag{1}$$

where

| | |
|---|---|
| $\mu$ | is the average survival time |
| $\alpha_i$ | is the effect of the poison I, II, III for $i = 1, 2, 3$ |
| $\beta_j$ | is the effect of the treatment A, B, C, and D for $j = 1, 2, 3, 4$ |
| $(\alpha\beta)_{ij}$ | is the interaction effect between the poison $i$ and treatment $j$ |
| $x_{ijkl}$ | is the age of a rat (in days since birth) |
| $\gamma$ | is the effect of age on survival |
| $\delta_k$ | is the gender effect, where $k = 1$ is female, $k = 2$ is male |
| $\epsilon_{ijkl}$ | $\sim N(0, \sigma^2)$ independent errors |
| $y_{ijkl}$ | is the survival time of a rat in tens of hours |

Assume zero-sum constraints for the parameters.

i. For the data of the eight rats given in the table above, write out the model of all main effects including an interaction between poison and treatment in the matrix form

$$Y = X\beta + \epsilon.$$

ii. What rank does $X$ have (if all 48 rats are included)?

(b) Which of the following quantities are estimable in model (1), if we do not assume any constraints on the parameters? Explain. Give an interpretation of the quantities (in plain English). Would assuming zero-sum constraints for the parameters change your answers on estimability? Explain.

i. $\mu + \alpha_1$

ii. $\beta_2 - \beta_1$

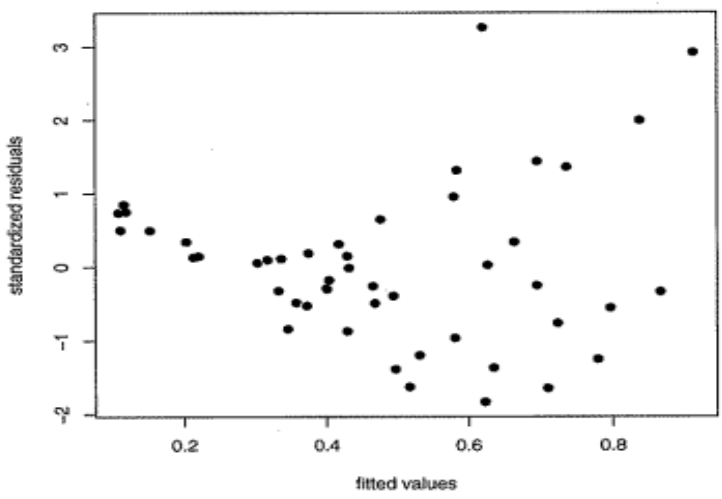


Figure 1: Scatter plot

iii. $\delta_2 - \delta_1$

(c) In a second model, Model 2, the interaction term is excluded. The difference in the residual sum of squares is given as 0.20, which corresponds to a 1.3 fold increase in the residual sum of squares between Model 1 and Model 2. Explain, how you can use these quantities in an $F$ test. Make sure to write out the hypotheses, as well as distributions and parameters of all sum of squares involved. The associated $p$ value is 0.146. Interpret.

(d) Standardised residuals are plotted versus fitted values in the scatter plot in Figure 1 above. Does the plot indicate any problems in the fit? Mention and explain three possible problems you can identify.

(e) Considering model 1, write 3 contrasts statements that could be used in SAS to test research questions. You are not supposed to consider any hypothesis that could be tested with the `lsmeans` statement in SAS.

6. The density of a $Gamma(\mu, \nu)$ distribution is given as (in mean parametrization):

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\mu}{\nu}\right)^{\nu} y^{\nu-1} e^{-\frac{\nu}{\mu}y},$$

where

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

is the Gamma Function (for integer values $n > 1$, $\Gamma(n) = (n-1)!$).

Show that the Gamma distribution is a member of the natural exponential family. Find the natural parameter, the dispersion parameter, and the canonical link.

7. Let $X$ be $N_3(\mu, \Sigma)$ with $\mu = [2, 3, -1]^T$ and

$$\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

(a) Find the distribution of $3X_1 - 2X_2 + X_3$.

(b) Relabel the variables if necessary, and find a $2 \times 1$ vector $a$ such that $X_2$ and $X_2 - a^T \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ are independent.

(c) Find the conditional distribution of $X_3$, given that $X_1 = x_1$ and $X_2 = x_2$.

8. In a study on college students at several universities, a PCA is conducted on the covariance matrix where the variables are (i) hours taken per semester, (ii) grade point average, (iii) hours studying per day, (iv) hours sleep per day and (v) annual cost of attending college ($). What will the first PC likely represent?

9. In a hierarchical cluster analysis, the sum of the diagonal elements of each of the K within-cluster covariance matrices is summed over K clusters, i.e. $sum = \sum_{i=1}^{K} tr(\Sigma_i)$. Should the sum increase or decrease as the number of clusters decreases? Briefly explain why.