**STT 872, 867-868 Fall Preliminary Examination**
**Wednesday, January 6, 2021**
**12:30 - 5:30 pm**

INSTRUCTIONS:

1. This examination is closed book. Every statement you make must be substantiated. You may do this either by quoting a theorem/result and verifying its applicability or by proving things directly. You may use one part of a problem to solve the other part, even if you are unable to solve the part being used. A complete and clearly written solution of a problem will get a more favorable review than a partial solution.

2. You must start solution of each problem on a separate page. Be sure to put the number assigned to you on the top left corner of every page of your solution. Also please number the pages with n/m (top right corner), where n is the current page number and m is the total number of pages, to keep the ordering and to avoid missing any pages during scanning.

3. In ZOOM, the video must be turned on for the whole duration of the exam, while the microphone must be muted for the whole duration of the exam. There should be no other people present in the room during the exam. DO NOT use virtual background. The camera should show a wide angle with you and the desk where your work is visible.

4. If you have questions during the exam (e.g. bathroom break requests) you can send a chat message in ZOOM to the host. Email/cell phone communication with Tami would be a back-up method to ZOOM/ D2L if they fail.

5. The exam will last 5 hours. Additional 30 minutes will be allowed to organize the paper solution (write your assigned number and the page number (n/m) on each page), scan it and upload to D2L. Submit your solution as a PDF file. Before the submission, make sure the PDF is clearly readable and it contains all your answers (check on your laptop). Failing to do so may result in substantial loss of points. Keep your paper solution until the examination result is out. If you run into any upload issues, email your solutions to Tami directly.

6. Please refrain from discussing the exam in any way before the results are made available.

**1.** Let $X_1, \cdots, X_n$ be i.i.d. from a uniform distribution on $U(-\theta, \theta)$.

(a) (3 pts) Find a minimal sufficient statistic $T$ for $\theta$.

(b) (3 pts) Let

$$V = \frac{\bar{X}}{X_{(n)} - X_{(1)}},$$

where $X_{(1)}$ is smallest order statistic, $X_{(n)}$ is the largest order statistic and $\bar{X}$ is the mean. Show that $V$ and $T$ are independent.

(c) (3 pts) Find $a_n$ and $b_n$ such that $a_n X_{(n)} + b_n$ converges to a non-degenerate distribution. **Hint:** Usually $a_n = n$ works.

(d) (3 pts) Construct the most powerful test for $H_0 : \theta = 1$ vs $H_1 : \theta = 2$ at the significance level $\alpha \in (0, 1)$.

**2.** Let $X$ follow an inverse binomial distribution with pmf

$$P(X = x) = \binom{m + x - 1}{m - 1} p^m (1 - p)^x, 0 < p < 1, x = 0, 1, \ldots$$

Suppose $X_1, X_2, \cdots, X_n$ are generated from the above inverse binomial distribution. Based on the sample $X_1, \cdots, X_n$,

(a) (3 pts) Find the UMVUE of $p$.

(b) (3 pts) Find MLE of $p$ and its asymptotic distribution.

(c) (3 pts) Find the Cramer-Rao lower bound for unbiased estimators of $p$. Does the UMVUE attain the bound?

**3.** Consider estimation of the means $\theta_1, \cdots, \theta_p$ of $p$ independent Poisson random variables $X_1, \cdots, X_p$ under the compound squared error loss, $L(\theta, d) = \sum_{i=1}^{p} (\theta_i - d_i)^2$.

(a) (4 pts) Following a Bayesian approach, let the unknown parameters be modeled as random variables $\theta_1, \cdots, \theta_p$ that are i.i.d. with the common density $\lambda e^{-\lambda x}$ for $x > 0, \lambda > 0$. Determine the Bayes estimators of $\Theta_1, \cdots, \Theta_p$ and the overall risk.

(b) (3 pts) Determine the marginal density of $X_i$ in the Bayesian model.

(c) (3 pts) Find an empirical Bayes estimator for $\Theta_1, \cdots, \Theta_p$.

**4.** A random variable $X$ has the Pareto distribution $P(c, \tau)$ if its density is given by $c\tau^c / x^{c+1}$, $0 < \tau < x$, $c > 0$.

(a) (3 pts) Find the distribution of $\log X$.

(b) (3 pts) Suppose $X_1, X_2, \cdots, X_n$ are observations from $P(c, \tau)$. Let $Y_i = \log X_i$. Use the $Y$s to construct the UMP test at the significance level $\alpha \in (0, 1)$ for $H_0 : \tau = \tau_0$ vs $H_1 : \tau < \tau_0$ assuming $c$ is known.

(c) (3 pts) Suppose $X_1, X_2, \cdots, X_n$ are observations from $P(c, \tau)$. Let $Y_i = \log X_i$. Use the $Y$s to construct the UMP test for $H_0 : c = c_0, \tau = \tau_0$ vs $H_1 : c > c_0, \tau < \tau_0$ at the significance level $\alpha \in (0, 1)$. **Hint:** use NP lemma for $c = c_0, \tau = \tau_0$ vs $c = c_1, \tau = \tau_1$.

**5.** This problem focuses on linear fixed effects models. **Hint:** The Woodbury matrix identity is $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$, where $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times k}, C \in \mathbb{R}^{k\times k}, V \in \mathbb{R}^{k\times n}$.

Consider two full-rank linear regression models:

$$y_i = x_i'\beta + \epsilon_i, \quad i = 1, 2, \ldots, n,$$
$$z_i = x_i'\gamma + \xi_i, \quad i = 1, 2, \ldots, n,$$

where $y_i, z_i \in \mathbb{R}, x_i \in \mathbb{R}^p$. The two models share the same design matrix, but the error terms are assumed correlated. Specifically, for each $i = 1, 2, \ldots, n$, the vector $(\epsilon_i, \xi_i)$ is independently sampled from $\mathcal{N}(0, \Sigma)$ with the unknown covariance matrix $\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \in \mathbb{R}^{2\times 2}$. It is known that the MLE for $\Sigma$ denoted by $\hat{\Sigma}$ takes the following form

$$\hat{a} = \frac{1}{n}Y'(I - X(X'X)^{-1}X')Y, \quad \hat{b} = \frac{1}{n}Y'(I - X(X'X)^{-1}X')Z,$$
$$\hat{c} = \frac{1}{n}Z'(I - X(X'X)^{-1}X')Z,$$

where $X = (x_1, \ldots, x_n)' \in \mathbb{R}^{n\times p}$, $Y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$, $Z = (z_1, \ldots, z_n)' \in \mathbb{R}^n$. Further denote the OLS under the two models by $\hat{\beta}, \hat{\gamma}$, respectively.

(a) (3 pts) Is $\hat{b}$ unbiased for $b$? Provide your arguments.

(b) (3 pts) An estimator $\hat{\theta}$ is linear if $\hat{\theta} = \sum_{i=1}^n (t_i y_i + s_i z_i)$ for some constants $\{(t_i, s_i)\}_{i=1}^n$. Prove that for any given vectors $\ell, \lambda \in \mathbb{R}^p$, $\ell'\hat{\beta} + \lambda'\hat{\gamma}$ is the BLUE for $\ell'\beta + \lambda'\gamma$ under the model in 5(a).

(c) (2 pts) Prove that $(\hat{\beta}, \hat{\gamma})$ is independent from $\hat{\Sigma}$ under the model in 5(a).

(d) (2 pts) Construct a confidence interval for $\ell'\beta - \ell'\gamma$ with the coverage probability of $1-\alpha$ under the model in 5(a).

**6.** This problem also focuses on linear fixed effects models. **Hint:** The Woodbury matrix identity is $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$, where $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times k}, C \in \mathbb{R}^{k\times k}, V \in \mathbb{R}^{k\times n}$.

Consider the full-rank linear regression model:

$$y_i = x_i'\beta + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $x_i \in \mathbb{R}^p$, and $\epsilon_1, \ldots, \epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

(a) (5 pts) Consider the case where $p = 3, \beta = (\beta_0, \beta_1, \beta_2)', x_i = (1, w_i, z_i)'$. Suppose the data has been standardized, i.e., $\sum_{i=1}^n w_i = \sum_{i=1}^n z_i = 0, \sum_{i=1}^n w_i^2 = \sum_{i=1}^n z_i^2 = 1$. To predict $y$ of a new data point $(y, w, z)$ (independently sampled from the same model), consider two prediction methods (include the intercept in the prediction!): (1) OLS based on the first predictor (2) OLS based on both predictors. Compute the expected predictor error for both methods. Under what conditions does the first method have a smaller predictor error?

**Hint:** Inversion of a $2\times 2$ matrix is $\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

(b) (5 pts) Consider the case when $\frac{1}{n}X'X = I_p$ and $\sigma^2$ is known. Denote the relevant variables by the set $\mathcal{M} = \{1 \le j \le p : \beta_j \neq 0\}$. Characterize the set $\hat{\mathcal{M}}$ selected by the Bayesian information

criterion (BIC). Prove the variable selection consistency of BIC, i.e., $\mathbb{P}(\hat{\mathcal{M}} = \mathcal{M}) \to 1$, as $n \to \infty$ and $p$ is fixed.

**7.** This problem focuses on mixed effects models. **Hint:** The Woodbury matrix identity is $(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$, where $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times k}, C \in \mathbb{R}^{k\times k}, V \in \mathbb{R}^{k\times n}$.

A growth curve data measures the total body bone mineral density for adolescent girls. We denote the measurement for the $i^{th}$ girl at time $t_{ij}$ by $y_{ij}$, and assume the following model:

$$y_{ij} = \alpha_i t_{ij} + \gamma_i + \epsilon_{ij}, \quad i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, n_i, \tag{1}$$

where all the $\epsilon_{ij}$'s are independently sampled from $\mathcal{N}(0, \sigma^2)$. To model the dependence of the bone density gain on calcium intake, we further assume

$$\alpha_i = c_0 + c_1 d_i + \xi_i, \quad i = 1, \ldots, N, \tag{2}$$

where $d_i \in \mathbb{R}$ denotes the daily calcium supplement of the $i^{th}$ girl, and all the $\xi_i$'s are independently sampled from $\mathcal{N}(0, \tau^2)$. The observed data consists of $\{(y_{i1}, \ldots, y_{in_i}, t_{i1}, \ldots, t_{in_i}, d_i)\}_{i=1}^N$, and the unknown parameters are $\{\gamma_i\}_{i=1}^N, c_0, c_1, \tau^2, \sigma^2$.

(a) (5 pts) Show that the model specified in (1) and (2) for the data is a classical linear mixed model.

(b) (5 pts) Consider the special case when $t_{ij} = s_j, n_i = n, i = 1, \ldots, N, j = 1, \ldots, n_i$ (i.e., each measurement is made at the same time for the subjects), and the parameters $\{\gamma_i\}_{i=1}^N$ are known. Derive the closed-form solution for the MLE of $(c_0, c_1)$.