2024 PhD Qualification Examination
Department of Statistics
University of South Carolina
Part I: August 15, 2024
9:00AM–1:00PM

Instructions:

Part I consists of four questions. Each question is of equal weight.

Use separate sheets of paper for each problem. Do not write on the back of any page.

Write your assigned code at the top of every page you submit. Do not write your name anywhere on your solutions.

Part II of the exam will start at 9:00am tomorrow.

1. Suppose X is a random variable with a χ^2_k distribution with probability density function

$$f_X(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(\frac{k}{2})}, \quad x > 0,$$

and Y is also independently distributed as χ^2_k. Let $U = X + Y$ and $V = \frac{X}{Y}$.

(a) Find the joint probability density function of U and V. Are U and V independent? Justify your answer.

(b) Determine the marginal probability density function of V.

(c) Assuming $k > 2$, show that the mode of V is $(k-2)/(k+2)$.

(d) Consider n independent random variables $V_1, V_2, ..., V_n$ from the same distribution as V. What is the approximate distribution of

$$\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$$

for large n and $k > 4$?

2. Suppose $X \sim$ Bernoulli(1/2) and $Y \sim$ Bernoulli(p), where $p \in (0,1)$ is an unknown parameter. Suppose X and Y are independent. Due to aggregation, we do not get to observe X or Y. Instead, we only get to observe $Z = X + Y$.

(a) Show the probability mass function of Z is

$$f_Z(z) = \begin{cases} (1-p)/2, & z = 0 \\ 1/2, & z = 1 \\ p/2, & z = 2 \\ 0, & \text{otherwise.} \end{cases}$$

For the remaining parts, suppose $Z_1, Z_2, ..., Z_n$ is an independent and identically distributed (iid) sample from $f_Z(z)$.

(b) Give conditions for which the maximum likelihood estimator (MLE) of p exists and determine what the MLE is when these conditions are satisfied. When the MLE exists, show that it is a consistent estimator of p.

(c) Suppose p is best regarded as random and follows a beta(a, b) prior distribution, where $a > 0$ and $b > 0$ are known. Find $E(p|Z_1, Z_2, ..., Z_n)$, the posterior mean of p.

3. Suppose $X_1, X_2, ..., X_n$ is an independent and identically distributed sample with cumulative distribution function $F(x)$, a continuous function of x. The median m is defined as the value $F^{-1}(0.5)$. Assume $F(x)$ is strictly increasing in an open neighborhood about m so that m is uniquely defined.

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics for $X_1, X_2, ..., X_n$ and let

$$Z = \sum_{i=1}^{n} I(X_i \leq m),$$

that is, Z is a random variable denoting the number of observations in the sample that are smaller than or equal to the median.

(a) What is the exact sampling distribution of Z?

(b) Use a normal approximation to find $P(Z \leq k)$, for $k = 0, 1, ..., n$. You may want to use a continuity correction because k is an integer.

(c) Find $P(X_{(j)} \geq m)$, the probability the jth order statistic exceeds the median m. *Hint:* Try to relate the event $\{X_{(j)} \geq m\}$ to the event $\{Z \leq k\}$, for some k.

(d) Assuming the sample size n is sufficiently large so that the normal approximation in part (b) is reasonable, find a $100(1-\alpha)\%$ confidence interval for m using the order statistics.

4. A group of high-technology companies agreed to share employee salary information in an effort to establish salary ranges for technical positions in research and development. Data obtained for each of $n = 65$ employees included

- current annual `salary` (Y, measured in \$1,000s)
- highest academic `degree` (x_1, a factor variable with 1 = bachelor's degree, 2 = master's degree, 3 = doctoral degree)
- `years` of experience since the last degree earned (x_2)
- `number` of persons the employee supervises (x_3).

Refer to the attachment entitled "R code and output" to answer the questions below. The variable names in the attachment are `salary`, `degree`, `years`, and `number`, respectively.

(a) The attachment shows output from regressing `salary` on all predictors and their two-way interaction terms. Answer the three parts below.

(a1) Comment on the importance of each individual interaction term.

(a2) Test if the interaction terms as a group are useful.

(a3) Discuss why the results you found in (a1) and (a2) are reasonable and provide support for your discussion.

(b) The attachment shows output from regressing `salary` on all predictors and the two-way interaction terms involving `degree` and `number`. Answer the three parts below.

(b1) Express the corresponding theoretical regression model (make sure to define any dummy variables). Justify why this model is the best model among all models considering `degree`, `years`, and `number` and their two-way interactions.

(b2) Write out the overall fitted model and the corresponding fitted models for each degree level.

(b3) Construct a 95% confidence interval for the mean salary <u>difference</u> between employees who have a master degree with 5 years of experience and supervise 5 employees and those employees who have a doctoral degree with 5 years of experience and also supervise 5 employees. Interpret your result.

(c) Consider performing post-fitting diagnostics for the model in part (b). Answer the three parts below (continues on the next page).

(c1) Conduct residual analysis to check the regression model assumptions. Discuss your findings.

(c2) Are there any outliers and/or influential data points? Justify your answer clearly.

(c3) Based on your findings in (c1) and (c2), do you trust the inference you made in (b3)? Justify your answer clearly.

(d) Describe in detail two alternative approaches to improve the analysis.

2024 PhD Qualification Examination
Department of Statistics
University of South Carolina
Part II: August 16, 2024
9:00AM–1:00PM

Instructions:

Part II consists of four questions. Each question is of equal weight.

Use separate sheets of paper for each problem. Do not write on the back of any page.

Write your assigned code at the top of every page you submit. Do not write your name anywhere on your solutions.

1. Suppose $X_1, X_2, ..., X_n$ are iid from a uniform distribution $U(\theta, \theta + 1)$, where $\theta \in \mathbb{R}$. Suppose $n \geq 2$.

(a) Find the joint distribution of $X_{(1)}$ and $X_{(n)}$.

(b) For testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, we propose the test

$$T(X_{(1)}, X_{(n)}) = \begin{cases} 0, & \text{if } X_{(1)} < 1 - \alpha^{1/n} \text{ and } X_{(n)} < 1 \\ 1, & \text{otherwise,} \end{cases}$$

for some $\alpha \in (0, 1)$. Derive the power function of $T(X_{(1)}, X_{(n)})$ and show that it is a size α test.

(c) Show that T is a uniformly most powerful (UMP) test. *It is important to note that the relevant distribution family does not have monotone likelihood ratio and thus the Karlin-Rubin Theorem would not apply.*

2. Consider the statistical model

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + b_t + \epsilon_t, \quad \text{for } t = 1, 2, ..., n,$$

where

- $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, $t = 1, 2, ..., n$,
- $b_t = \gamma_t + \theta\gamma_{t-1}$, $t = 1, 2, ..., n$, where $\gamma_0, \gamma_1, ..., \gamma_n$ are iid $N(0, \tau^2)$, and
- the ϵ_t's and the γ_t's are mutually independent.

Covariates $\mathbf{x}_1 = (x_{1,1}, x_{1,2}, ..., x_{1,n})'$ and $\mathbf{x}_2 = (x_{2,1}, x_{2,2}, ..., x_{2,n})'$ are best regarded as fixed and measured without error.

(a) Express the covariance matrix for $\mathbf{b} = (b_1, b_2, ..., b_n)'$ in terms of θ and τ^2.

(b) Express the covariance matrix for $\mathbf{y} = (y_1, y_2, ..., y_n)'$ in terms of θ, τ^2, and σ^2. Find the joint distribution of $\mathbf{y}$.

(c) Derive the maximum likelihood estimator of $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ assuming that θ, τ^2, and σ^2 are known.

(d) Assuming that θ, τ^2, and σ^2 are known, propose an appropriate test to determine if the second covariate is needed in the model. Describe how you would carry out the test, providing as many details as possible.

3. Recall the probability density function for the Laplace distribution with location parameter μ and scale parameter σ is

$$f_X(x|\mu,\sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty.$$

Suppose $X_1, X_2, ..., X_n$ are independent and identically distributed according to a Laplace distribution with $\mu \in \mathbb{R}$ and $\sigma = 1$, where μ unknown.

However, in this problem, the X_i's are not observed but rather an indicator Y_i for the sign of each X_i has been recorded, that is,

$$Y_i = \begin{cases} 1, & X_i \geq 0 \\ 0, & X_i < 0, \end{cases}$$

for $i = 1, 2, ..., n$.

(a) Find the maximum likelihood estimator of μ based on the observed data $Y_1, Y_2, ..., Y_n$ in the case where $\sum_{i=1}^n Y_i \geq n/2$. You may assume $\mu > 0$.

(b) Is $\sum_{i=1}^n Y_i$ a sufficient statistic for μ? If so, prove it. If not, explain or show why not.

(c) Suppose σ is also unknown. Would it be possible to estimate both μ and σ based on the observed data $Y_1, Y_2, ..., Y_n$? Explain your reasoning.

4. Suppose we measure a response variable Y_i on each of n subjects, $i = 1, 2, ..., n$. For each subject, we have a set of M candidate predictor variables $x_{i1}, x_{i2}, ..., x_{iM}$. We would like to model the expected value of Y as a function of these predictor variables.

Consider the regression model

$$Y_i = f(x_{i1}, x_{i2}, ..., x_{iM}) + \epsilon_i.$$

We know only that the function $f(\cdot)$ is linearly related to a subset of the $M + 1$ candidate predictors (which includes the intercept term), that is,

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_M x_{iM},$$

where some of the β_j's may be equal to zero. We assume $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$, and $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are mutually independent random variables. We also assume σ^2 is **known**.

Our goal is to get as accurate an estimated prediction as possible. Realizing that

- including too many predictors in the model will increase the variance of the estimated prediction, and
- including too few predictors in the model may increase bias due to lack of fit,

we decide to choose the set of predictors that minimizes

$$\text{AMSE} = n^{-1} E\left[\sum_{i=1}^{n} \{\widehat{Y}_i - E(Y_i)\}^2\right],$$

the *average mean squared error* of the estimated prediction.

If we knew $E(Y_i)$, then we could simply calculate AMSE for each of the candidate models and pick the model with the smallest AMSE. However, because $E(Y_i)$ is not known, we don't know AMSE and hence AMSE must be estimated from the data. We consider the statistic

$$C_p^* = \frac{\text{SSE}}{n} + \frac{2p\sigma^2}{n} - \sigma^2$$

for a model with p predictor variables (including the intercept term). The sum of squared residuals SSE is

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2,$$

where $\widehat{Y}_i$ is the ith predicted value from estimating the model with p predictor variables (including the intercept term) using ordinary least squares. The statistic C_p^* is related to the Mallows C_p statistic, which is widely used for model selection in linear regression.

(a) Show that C_p^* is an unbiased estimator of AMSE. *Hint*: Note that $\text{SSE}/(n - p)$ may not be an unbiased estimator of σ^2. It only is when the model is correct, that is, when the correct set of p predictor variables (including the intercept term) is selected.

In the following fictional study, we measured the weight Y on $n = 50$ children. The candidate predictors are

- the intercept term ($x_0 = 1$)
- age (x_1, measured in weeks)
- day of birth (x_2, coded from 1 to 365)
- sex (x_3, coded as 0 for males and 1 for females).

Here were the results:

Predictors in the model	p	C_p^*
x_0	1	3687.7
x_0, x_1	2	73.6
x_0, x_2	2	3216.3
x_0, x_3	2	3657.3
x_0, x_1, x_2	3	73.9
x_0, x_1, x_3	3	72.3
x_0, x_2, x_3	3	3189.9
x_0, x_1, x_2, x_3	4	72.6

(b) Using the table above, which model would you choose as best describing the data? Why? Be brief.

(c) The coefficient of determination

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

is often used as a measure of goodness of fit in multiple regression models. Which of the above models gives the largest value of R^2?

(d) Explain briefly whether you think that C_p^* or R^2 is a better criterion for choosing a regression model.