May 2017 PhD Qualifying Examination
Department of Statistics
University of South Carolina
9:00AM–3:00PM

Instructions: This exam consists of six problems. Answer all six problems. Use separate sheets of paper for each problem, and write on one side of each page (do **not** write on both sides). Write your confidential code letter and the problem number at the top of **each page**.

You are allowed to use the computers and the software in the examination room. However, you are **not** allowed to use the Internet, except to examine help files of the software and to examine data sets that are needed in some of the problems. Provide complete details in your solutions. You have **six hours** to complete this examination. Good luck!

1. Suppose $X_1, X_2, \ldots, X_n$ are iid random variables with common pdf

$$f(x|\theta) = \theta x^{\theta-1},\ 0 < x < 1,\ \theta > 0$$

(a) Prove that $-\sum_{i=1}^{n} \ln(X_i)$ is a complete sufficient statistic for this family.

(b) Derive the probability distribution of $-\sum_{i=1}^{n} \ln(X_i)$.

(c) Find the $UMVUE$ for $\frac{1}{\theta}$. Does it attain the $CRLB$? Give a detailed reason why or why not. Show that the $UMVUE$ also happens to be the MLE of $\frac{1}{\theta}$.

(d) Find the MLE of θ and derive the asymptotic probability distribution of the MLE.

2. The file available at `http://people.stat.sc.edu/hitchcock/EmployeeData2017.txt` contains data on 65 randomly selected employees of a large company. The variable y is the monthly salary (in dollars) of the employee; x_1 represents the sex of the employee (0=male, 1=female); x_2 is a average score (out of 50) over the past six months on a performance rating of the employee; and x_3 is a percentage of the annual personal expense budget that the employee spent over the last year (note that some employees exceeded the personal budget value).

(a) Suppose you only had the data on salary and sex. Perform an appropriate procedure to test whether male and female employees at this company tended to have different monthly salaries, on average. State and verify the assumptions of your procedure.

(b) Now suppose you have access to all the variables in the data. Build an appropriate regression model to help shed light on some questions of interest: Which variables affect or predict monthly salary? What is the apparent relationship between each important predictor and monthly salary, in the context of the entire set of variables in the model? Do these relationships depend on the values of the other predictors? How well can the model explain the variation in monthly salary? State and verify the assumptions of your model.

(c) Answer question (a) again with a formal test, but now considering the information in the whole set of available variables. Discuss whether your conclusions from parts (a) and (c) are the same, and explain why or why not.

(d) Test whether those employees who exceeded their annual personal expense budget tend to have higher monthly salaries than those who did not exceed their annual personal expense budget, accounting for the other variables in the data set. State any limitations of this inference that you notice.

(e) Now suppose that a new goal is to build a regression model to predict a new variable (call it personal profit margin, P), now using monthly salary (the same as y, but call it M here) as a predictor. Suppose a cubic regression model is fit:

$$E(P) = \beta_0 + \beta_1 M + \beta_2 M^2 + \beta_3 M^3,$$

where the M variable is not centered. The file at
`http://people.stat.sc.edu/hitchcock/OtherEmployeeData2017.txt` contains the data for P and M for the 65 employees. Obtain a point estimate for *each* "turning point" of the cubic regression function, where the turning points are defined as the the values of M at which the mean response function goes from increasing to decreasing, or vice versa. Also, under the normal-errors assumption, obtain a set of familywise approximate 90 percent confidence intervals for these turning points of the cubic regression function. Show/explain how you obtained your estimates and intervals.

3. Let $Y_1, \ldots, Y_n$ be random variables and $X_1, \ldots, X_n$ be fixed constants taking values in $[0, 1]$ such that

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with mean 0 and variance σ^2 and $m(\cdot)$ is a function from $[0, 1]$ to the real numbers.

Divide $[0, 1]$ into L intervals of equal length

$$I_\ell = \left[\frac{\ell - 1}{L}, \frac{\ell}{L}\right), \quad \ell = 1, \ldots, L-1, \quad \text{and} \quad I_L = \left[\frac{L-1}{L}, 1\right]$$

and define the indicators

$$\mathbf{1}_\ell(x) = \begin{cases} 1 & \text{if } x \in I_\ell \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $n_\ell := \sum_{i=1}^n \mathbf{1}_\ell(X_i) > 0$ for $\ell = 1, \ldots, L$.

For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^T$, let $m_{\boldsymbol{\beta}}(\cdot) = \sum_{\ell=1}^L \beta_\ell \mathbf{1}_\ell(\cdot)$ be the function which takes the value β_ℓ over the interval I_ℓ and consider the estimator $m_{\hat{\boldsymbol{\beta}}}(\cdot)$ of $m(\cdot)$, where

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - m_{\boldsymbol{\beta}}(X_i)]^2.$$

(a) Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and find the design matrix $\mathbf{Z}$ such that

$$\sum_{i=1}^n [Y_i - m_\beta(X_i)]^2 = \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|_2^2,$$

where $\|a\|_2^2 = \sum_{i=1}^n a_i^2$.

(b) Find $\hat{\beta}_1, \ldots, \hat{\beta}_L$.

(c) Find $\operatorname{Var}\left(m_{\hat{\boldsymbol{\beta}}}(x)\right)$ for x in the interval I_ℓ.

(d) Suppose the true function $m(\cdot)$ has the property that for any $x_1, x_2 \in [0, 1]$,

$$|m(x_1) - m(x_2)| < C|x_1 - x_2|$$

for some $C > 0$.

Use this property of $m(\cdot)$ to show that $\operatorname{Bias}\left(m_{\hat{\boldsymbol{\beta}}}(x)\right) \leq C/L$ for any x in $[0, 1]$.

(e) Increasing the number of intervals L will have what effect on the variance of $m_{\hat{\beta}}(\cdot)$? How about on the bias of $m_{\hat{\beta}}(\cdot)$?

4. While driving to work one morning, I got stopped in a traffic light at the intersection of Assembly Street and Gervais Street. I noticed that the vehicle in front of me had a broken brake light. I looked at the other vehicles that were also stopped (about 20 of them) and these other vehicles did not have broken brake lights (the police will usually give you a ticket if your brake lights are broken, so the event of a car having broken brake lights is a rare event). I got to wondering: *What is the percentage of all vehicles being driven in Metropolitan Columbia have broken brake lights?* You *are* the statistician that is consulted about this problem.

(a) Describe a statistical sampling plan that could be done within a reasonable amount of time at a reasonable cost to gather relevant data to answer the primary question of inferring about the proportion, denoted by θ, of all vehicles being driven in MetroColumbia that have broken brake lights. Explain why your design is appropriate and describe the type of sample data that you will obtain from this study. In particular will your study have a fixed sample size, or will it have a random sample size?

(b) Based on the data that you will obtain from your study, describe and fully justify your procedure for performing inference (estimation and constructing a confidence interval) about θ. You should describe the appropriate statistical model that you will be postulating and must justify why such a model is reasonable. You should describe the estimator that you will use and justify why such an estimator will be good. For instance, will your estimator have desirable properties and what are these desirable properties? You should also describe how you will obtain a measure of the degree of precision of your estimator.

(c) Based on past information about vehicles in Metropolitan Columbia, a prior distribution about θ is given by a beta distribution with parameters $(\alpha, \beta) = (2, 98)$, so that the prior density function is

$$\pi(\theta) = \frac{1}{B(2,98)}\theta^{2-1}(1-\theta)^{98-1}I\{0 < \theta < 1\}.$$

If you are given this prior information, what will be your Bayes estimator of θ based on squared-error loss function?

(d) Based on your sampling design, could your estimates obtained in (b) and also in (c) be equal to zero? If you get an estimate of zero, will this be a sensible or reasonable estimate?

(e) A certain Professor X, who is *not* so knowledgeable about the intricacies of statistical modeling and inference, insisted that the best sampling design for this study is to observe 500 randomly chosen cars in Metropolitan Columbia. Upon making his observations (of course, with the help of his willing students), he found that none of the 500 cars that were observed have broken brake lights. However, he still claims that a conservative 95% confidence interval for θ based on the observed data is given by $[0, 3/500] = [0, .006]$. Is he justified in his claim? Justify your answer.

(f) Using the sample data obtained by Professor X and the prior distribution of θ in item (c), what would be a 95% Bayesian credible interval for θ?

5. Suppose an agricultural researcher is investigating twelve new corn varieties as well as the currently recommended commercial variety. She is interested in learning whether any of the test (new) varieties have greater yield than the current variety. In the design of the experiment, four complete blocks were used (each treatment appeared once within each block).

(a) Explain the likely reason for the choice of the block design. Also, the agronomist assumed "random block effects" – what does this imply in the context of the experiment?

(b) Carefully write the model for the experiment, defining any symbols and notation used. Include the usual distributional assumptions for an experiment of this type.

(c) Write the ANOVA table, including sources of variation and exact degrees of freedom. Give the formulas for the test statistics for testing about treatment effects and block effects.

(d) Suppose the usual ANOVA F-tests indicate that differences exist among the variety mean yields. Discuss the best approach to formally test the agronomist's main research question.

(e) Under the assumptions of the model, derive the standard deviation of the j-th variety sample mean. Also derive the standard deviation of the i-th block sample mean.

(f) Suppose that after these data are gathered, four more independent yield measurements will taken using variety 1 and block 1. Under the assumptions of our model, find the probability that the sample mean of these four new yields for (variety 1, block 1) will be at least 1.3 times the true expected yield value for (variety 3, block 1). Your answer should be given in terms of the model parameters.

6. Suppose $X_1, X_2, \ldots, X_{30}$ is a random sample from a $Gamma(2, \theta)$ distribution.

(a) Derive the size 0.05, UMP test for $H_0 : \theta \leq 1$ *vs.* $H_1 : \theta > 1$ and derive the power function of the test.

(b) Consider the testing problem $H_0 : \theta = 1$ *vs.* $H_1 : \theta \neq 1$. Argue that an UMP test does not exist in this case. Derive the exact size 0.05 Likelihood Ratio Test (LRT) for this problem.

(c) In the same graph, plot the power function of the test in part (b) along with the power function in part (a) and comment on the plot.

(d) Give a 95% confidence interval for θ using a random sample of size 30, viz. $X_1, X_2, \ldots, X_{30}$.

(e) Suppose the true value of θ is 1. Generate 30 random observations from $Gamma(2, 1)$ distribution and construct the confidence interval derived in (d) from this generated sample observations. Repeat this 10 times to generate 10 different confidence intervals. What is the observed coverage probability (proportion of these 10 constructed intervals that actually covers the true value $\theta = 1$)? Provide the code used for this simulation.
Comment on why the observed coverage probability may not be exactly equal to the confidence coefficient 95% used in part (d).