PhD Qualifying Examination

Department of Statistics
University of South Carolina
May 25, 2018 - 9:00AM–3:00PM

READ FIRST THESE INSTRUCTIONS

1. **Procedure for assigning examinee number and codename:** Each examinee will draw a piece of paper from each of two boxes: Box A and Box B. The piece of paper from Box A will contain a number and the piece of paper from Box B will contain a codename. Write your real name, number, and codename in a sheet of paper provided to you by the proctor. Place this sheet of paper in an envelope provided to you, seal the envelope, and sign the envelope on the seal portion. Give the envelope to the proctor. The proctor should keep these envelopes sealed until all exams are graded and the Examination Committee has decided on the results. The Examination Committee will then open these envelopes to determine the examinees names and their results.

2. DO NOT write your real name in any parts of your answer sheets. Instead, write the number and codename that you drew from Step 1.

3. There are seven (7) sets of problems in this examination, with each set of problems appearing in one page of this examination questionnaire. Choose **exactly** six problems to answer. When you submit your work, **indicate clearly on the sheet of paper titled 'Problems to be Graded' which problems you have decided to answer and should be graded**.

   However, if you answered all seven problems but did not indicate the six problems to be graded, then the Examination Committee will delete the problem where you got the **highest (best)** score. **Each of the six problems that you will answer are equally weighted in the overall score.**

   Write your solutions on the yellow sheets of paper provided in your examination packet. **Write only on one side on each sheet of paper.**

4. The two data sets referred to in two of the problems can be found in the pin drive that is included in your examination packet. These are the files "CallData.txt", "CallData.R", "igrowup.txt", "igrowup.R". You could source the files with the .R extension to read the file into R, or you could use the read.table command to read in the files with the .txt extension.

5. You may use a computer in the examination room, the statistical packages (e.g., R or SAS) in them, and/or a calculator. However, you are **not allowed** to use search engines during the examination. But you may use the HELP manuals of the statistical packages that you use. Please adhere to the HONOR CODE in this instance.

6. You have six hours for this examination. Good luck and may the *Force* be with you.

Examination Committee: Dr. E. Peña (Chair); Dr. P. Chakraborty; Dr. K. Gregory; Dr. Y. Ho.

The Problems

1. A disease epidemic (think of the flu) occurs in a certain country. The number of infected individuals, K, has a Poisson distribution with parameter (mean) λ. Each infected individual has a probability θ of dying from the disease and dying among the infected individuals are independent events. Denote by X the number of deaths arising from this disease.

   (a) What is the conditional distribution of X, given $K = k$.

   (b) Find the (unconditional) mean of X?

   (c) Find the (unconditional) variance of X?

   (d) Find the distribution of X?

   (e) Find the conditional distribution of K, given that $X = x$.

   (f) Suppose it is known that $\theta = \theta_0$. If you only observe the value of X, say $X = x_0$, provide an estimate of λ.

   (g) Provide an estimate of the standard error of the estimate of λ you obtained in (f).

2. The number of calls arriving at an answering service during the lunch-hour on any given day follows a Poisson distribution with mean λ. A manager needs to decide the lunch-hour schedule. This entails deciding whether to keep stand-in operators during the lunch time, in which case she has to assign different lunch-break time slots for different groups of people and also has to schedule how many ports each of the stand-in operators will cover during that time. If she decides not to keep stand-in operators, everybody gets lunch-breaks at the same time and the scheduling will be very easy but at the expense of nobody manning the stations leading to unattended calls.

   One of the in-house statistician proposed the following scheme: if the chance of getting at least one call during the lunch hour is 50% or more, the manager should keep stand-in operators; otherwise, if the chance of getting any calls during the lunch hour is less than 50% then risk missing few calls and opt for the same lunch break for everyone to ease up scheduling burden.

   A 30-day record of the number of calls that arrived during the lunch hour is given in Table 1. This is the file `CallData.txt` in the pin drive.

Table 1: Number of calls received during a 30-day period in an answering service office.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Number of calls	2	0	1	1	0	0	5	0	1	1	2	2	0	2	4	1

Day	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Number of calls	2	0	1	0	1	1	2	0	0	0	1	0	1	3

   (a) Propose appropriate statistical hypotheses to test for this problem. Justify your choice for the null and alternative hypotheses.

   (b) Conduct a suitable statistical test at the 5% level of significance. Justify your choice of the test. You should mention whether your test has good properties such as having the best power, and explain why.

   (c) Plot the power function of your proposed test. You may use a software (e.g., R) to produce this plot.

   (d) Does your test have size exactly equal to .05? If not, could you still improve (in terms of power) your test and how would you improve it?

3. Let $X_1, \ldots, X_n \in \mathbb{R}$ be a random sample from a uni-modal distribution with probability density function f and let $\theta = \text{mode}\{f\} = \arg\max_x f(x)$ be the mode of f. For some $h > 0$, let

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

be an estimator of $f(x)$ for all $x \in \mathbb{R}$, where K is a kernel function which is a density on $\mathbb{R}$, and let $\hat{\theta}_{n,h} = \text{mode}\{\hat{f}_{n,h}\}$ be the mode of $\hat{f}_{n,h}$. Suppose the quantity

$$\sqrt{nh^3}(\hat{\theta}_{n,h} - \theta)$$

has cdf $G_{n,h}$.

(a) Assuming $G_{n,h}$ is completely known, give expressions for the upper and lower bounds of a $(1-\alpha)100\%$ confidence interval for θ based on $\hat{\theta}_{n,h}$.

(b) Assuming $G_{n,h}$ is completely known, write down a rejection rule for testing H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$ such that the test has size equal to α.

(c) Suppose f is known. Describe a computational method to approximate the cdf $G_{n,h}$ of $\sqrt{nh^3}(\hat{\theta}_{n,h} - \theta)$.

(d) Having observed the random sample $X_1, \ldots, X_n$, suppose you wish to draw random samples of size n from the distribution which has cdf

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x),$$

where $\mathbb{1}(\cdot)$ is the indicator function. Describe how this can be done.

(e) Having observed the random sample $X_1, \ldots, X_n$ and computed the estimator $\hat{f}_{n,h}$ of f, suppose you wish to draw random samples of size n from the distribution which has probability density function $\hat{f}_{n,h}$. Describe how this can be done. Assume that there is a way to draw realizations from the density K.

(f) Describe in detail a way to estimate the cdf $G_{n,h}$ by drawing random samples from the distribution with density $\hat{f}_{n,h}$.

4. The World Health Organization (WHO) Child Growth Standard data (the `igrowup.txt` file in pin drive) was obtained in order to study the relationships among weight, age and gender in young children. The dataset contains observations from 498 children aged between 0 and around 60 months. Missing values are indicated by NA in `igrowup.txt`.

   The variables in the data are: **id**: child's ID; **Gender**: 1 for males, 2 for females; **agemons**: child's age in months; **WEIGHT**: child's weight in kilogram (kg); and **HEIGHT**: child's height in centimeter (cm).

   (a) Plot weight against **agemons**. Then perform a simple linear regression of **weight** (outcome) on **agemons** (predicator variable). Add the least squares line on the scatter plot.

   (b) Write down the model you performed in (a). Write down the least squares estimator of the regression coefficient vector ($\hat{\beta}$) and the covariance ($Cov(\hat{\beta})$) in matrix form. Make sure to define your notation clearly. Write down the key assumptions of the model.

   (c) Suppose a pediatrician told you that children's growth slows down after 12 month of age. Write down a second model that incorporates this information. Perform a regression analysis using this model and interpret the regression coefficients. Add the fitted values from this second model in the scatter plot obtained in (a).

   (d) Is the change of growth rate after 12 months statistically significant? Perform a hypothesis test to address this question and report corresponding 95% confidence interval. Report your conclusion based on the hypothesis test.

   (e) Make a residual plot using the model specified in (c). Write down the key assumptions of the model and comment on how reasonable each assumption is in light of the residual plot.

   (f) What would be the impact of violation to the model assumptions described in (e)? Propose a way to correct the issue you identified. Perform the analysis and provide correct inference to answer the question in (d).

   (g) Propose an appropriate model and perform an analysis to answer the following questions:

      (i) Are the birth weights the same for boys and girls?

      (ii) Are the growth rates the same for boys and girls?

      (iii) Use appropriate model selection procedure to select your best model and interpret the regression coefficients from your best model.

5. Consider a pair of random variables (X, Y) and let $\mu_X = \mathbb{E}X$, $\sigma_X^2 = \operatorname{Var} X$, $\mu_Y = \mathbb{E}Y$, $\sigma_Y^2 = \operatorname{Var} Y$, and $\sigma_{XY} = \operatorname{Cov}(X, Y)$.

   (a) Define α_0 and β_0 as

   $$(\alpha_0, \beta_0) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \mathbb{E}\left[Y - (\alpha + \beta X)\right]^2.$$

   Find expressions for α_0 and β_0.

   (b) Let U be a random variable which is independent of X and Y and for which $\mathbb{E}U = 0$ and $\mathbb{E}U^2 = \sigma_U^2$. Suppose one is to observe, instead of the pair (X, Y), the pair (W, Y), where $W = X + U$. We may think of X as being observed with some measurement error U.

   Define α_0^W and β_0^W as

   $$(\alpha_0^W, \beta_0^W) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \mathbb{E}\left[Y - (\alpha + \beta W)\right]^2.$$

   Let $\lambda = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$, and find expressions for α_0^W and β_0^W which involve only α_0, β_0, λ, and μ_X.

   (c) Suppose $Y = \alpha_0 + \beta_0 X + \varepsilon$, where $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma_\varepsilon^2$. Then we may write

   $$Y = \alpha_0^W + \beta_0^W W + \eta,$$

   where $\mathbb{E}\eta = 0$ and $\mathbb{E}\eta^2 = \sigma_\eta^2$. Find an expression for σ_η^2 in terms of σ_ε^2, λ, σ_X^2, and β_0.

   (d) Consider a data set with n independent realizations $(W_1, Y_1), \ldots, (W_n, Y_n)$ of (W, Y) and another data set with n independent realizations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), where $W = X + U$. Let

   $$(\hat{\alpha}_n^W, \hat{\beta}_n^W) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - (\alpha + \beta W_i)\right]^2$$

   and

   $$(\hat{\alpha}_n, \hat{\beta}_n) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - (\alpha + \beta X_i)\right]^2.$$

   i. Give an expression for $\mathbb{E}\hat{\beta}_n^W$ and comment on whether $\hat{\beta}_n^W$ is a good estimator of β_0.

   ii. Suppose $\hat{\beta}_n^W$ is used to test the hypotheses H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$. How do you expect the size and power of the test to be affected by the use of $\hat{\beta}_n^W$ instead of $\hat{\beta}_n$?

6. Assume a model for the random variables $\{Y_{ij}, i = 1, 2; j = 1, 2, \ldots, n_i\}$ given by

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad Var(\epsilon_{ij}) = \sigma_i^2, \quad i = 1, 2; j = 1, 2, ...n_i,$$

where $\sigma_i^2, i = 1, 2$, maybe different; μ_is are constants, and the ϵ_{ij}s are independent. The usual t-statistic used in forming a confidence interval for $\mu_1 - \mu_2$ is

$$T = \frac{(\overline{Y_1} - \overline{Y_2}) - (\mu_1 - \mu_2)}{S(n_1^{-1} + n_2^{-1})^{1/2}}$$

where

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}, \quad n = n_1 + n_2,$$

with $s_i^2, i = 1, 2$, the sample variance of $\{Y_{ij}, j = 1, 2, \ldots, n_i\}$. If $\sigma_1^2 = \sigma_2^2$ and ϵ_{ij}s are normally distributed, then $T \sim t_{n-2} \approx N(0, 1)$ for large n. The confidence interval (CI) at confidence level $1 - \alpha$ for $\mu_1 - \mu_2$ in this case is the

$$(\overline{Y_1} - \overline{Y_2}) \pm (t_{n-2;\alpha/2})\, S\, (n_1^{-1} + n_2^{-1})^{1/2}$$

where $t_{n-2;\alpha/2}$ is the $100(1 - \alpha/2)\%$ quantile of the t-distribution with $n - 2$ degrees of freedom, or

$$(\overline{Y_1} - \overline{Y_2}) \pm (z_{\alpha/2})\, S\, (n_1^{-1} + n_2^{-1})^{1/2}$$

for large n, where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the $100(1 - \alpha/2)\%$ quantile of the standard normal distribution.

However, **suppose** that $\sigma_1^2 \neq \sigma_2^2$. Then, heuristically, $S^2 \approx \frac{1}{n}(n_1\sigma_1^2 + n_2\sigma_2^2)$ for large n, and T is approximately normally distributed with mean 0 and variance v^2.

(a) Show that $v^2 = Var(T) \approx \frac{\frac{\sigma_1^2}{\sigma_2^2} + \frac{n_1}{n_2}}{\frac{n_1\sigma_1^2}{n_2\sigma_2^2} + 1}$.

(b) Investigate the error rate of the confidence interval, i.e., calculate $P(\mu_1 - \mu_2 \notin \text{CI} | \mu_1 = \mu_2)$ at $\alpha = 0.05$ using the normal approximation to the test statistic T.

(c) Investigate the error rate using a simulation study. Discuss in detail your simulation set up. Construct two tables of error rates using (i) normal approximation and (ii) a simulation study where the columns and rows are specified by:

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{1}{8}, \frac{1}{2}, 1, 2, 8, \quad \text{and} \quad \frac{n_1}{n_2} = \frac{1}{2}, 1, 2, 8.$$

(d) Based on results from (a), (b), and (c), comment on the effect of unequal variances on the coverage probability in relation to the sample sizes.

7. Let $X_i, i = 1, 2, \ldots, n$, be IID from a $N(\mu, \sigma^2)$ with both μ and σ^2 unknown. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

be the sample mean and sample variance, respectively. $S = +\sqrt{S^2}$ will be the sample standard deviation.

It is generally accepted that $\bar{X}$ and S are good estimators of μ and σ, respectively. However, a budding mathematical statistician, named *WillBeAGreatOne*, claims that he has better estimators of (μ, σ) than $(\bar{X}, S)$. He says that, for some M a positive integer, one should specify fixed real numbers $-\infty < a_1 < a_2 < \ldots < a_M < \infty$ and then define

$$N_k = \sum_{i=1}^{n} \mathbb{1}\{X_i \leq a_k\}, k = 1, 2, \ldots, M.$$

where $\mathbb{1}(\cdot)$ is the indicator function. His estimators of (μ, σ) are the *minimizer*, with respect to (μ, σ), of the quantity

$$Q(\mu, \sigma; a_1, \ldots, a_M) = \sum_{k=1}^{M} \left\{ \frac{[N_k - n\Phi((a_k - \mu)/\sigma)]^2}{[n\Phi((a_k - \mu)/\sigma)]^2} \right\}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

(a) Discuss and provide reasons why $(\bar{X}, S)$ are considered to be good estimators of (μ, σ).

(b) Outline a method to computationally obtain *WillBeAGreatOne*'s estimates if you have the sample observations $x_1, x_2, \ldots, x_n$.

(c) What considerations should *WillBeAGreatOne* use to decide on the constants $a_k, k = 1, 2, \ldots, M$, and M?

(d) Discuss how estimators of (μ, σ) should be evaluated and compared. What characteristics should you compare?

(e) Assuming that *WillBeAGreatOne* is able to find the best a_k's and M when obtaining his estimates, discuss in detail whether his estimators will be better (or worse) than the usual estimators $(\bar{X}, S)$ for (μ, σ). In particular, argue whether one should or should not use *WillBeAGreatOne*'s estimators of (μ, σ).

(f) Describe a computational or a simulation approach to compare and evaluate the estimators $(\bar{X}, S)$ and *WillBeAGreatOne*'s estimators.