Qualifying Exam 2016 — Statistical Methods

Instructions:

- NOTE: You are not allowed to use internet or email except for downloading the data and emailing the final report (optional). To use these for any other purpose, ask the proctor.
- Load the datasets for questions 2 (OgleLMCSMCWa.csv) and 3 (fertility.txt) from the website http://www.utdallas.edu/~swati.biswas/QE. Let the proctor know if you have any problems with this step.
- You can use any software of your choice. You can use the lab machines or your own laptop.
- Your report should clearly explain the steps, results, conclusions, and justification for the conclusions. Also, include the codes and outputs (ONLY relevant parts; highlighted wherever possible). Do not attach the parts of the output that were not used in answering questions.
- Submit a report, written or typed, hard copy or by email to swati.biswas@utdallas.edu. If you choose to email, then attach only one single PDF file with the whole report. DO NOT email separate files for report, codes, and outputs.
- Write your name on the report. DO NOT write your ID.

1. Monte Carlo simulation for estimating type I error rate of chi-square test of independence. Consider two categorical variables, say X and Y. Suppose X represents race/ethnicity with five categories (White, Black, Hispanic, Asian, and American Indian) and Y represents disease status (Yes/No). Data collected on n individuals on these two variables can be represented in a 5×2 contingency table whose (i,j)th cell contains the observed number of individuals in the ith row and jth column (O_{ij}). Recall that in this situation, a chi-square test statistic given by

$$\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

can be used to test whether X and Y are associated. Here E_{ij} is the expected number of individuals in (i,j)th cell under the null hypothesis of no association between X and Y; E_{ij} is estimated from the observed sample. The asymptotic null distribution of this test statistic is χ^2_4, however, this distribution may not hold in some situations. For example, even though a 5% level (nominal level) test is used for inference, the actual type I error rate may be quite different from 5%. Thus, it is desirable to estimate the actual type I error rate of this test statistic by Monte Carlo simulation.

   Use 1,000 replications, 5% as the (nominal) level for the test, and vary n to be 20, 50, 100, and 500. For simulating X, use the following proportions: White: 62%, Black: 12%, Hispanic: 18%, Asian: 6%, and American Indian: 1%. For simulating Y, use 50% proportion for disease and 50% for non-diseased. Compare the estimated type I error rates with the nominal rate of 5%. For starting seed value, use the last 5 digits of your ID. [35 points]

2. Period-Luminosity-Color relationships in the Large and Small Magellanic Clouds. The key to obtaining distances in our universe was discovered 100 years ago by Henrietta Leavitt. Ms. Leavitt found that a relatiohship exists between the luminosity and periods of a special class

of pulsating variable stars called Cepheids. Calibrating this relatiohship allowed astronomers to determine distances to these stars from their periods of variability. This gave Edwin Hubble the tool he needed to discover that our Milky Way is just one galaxy among many others, and that our universe is expanding. This relationship continues to be studied and refined today because it is used to calibrate other methods for obtaining distances in our universe.

Data contained in the file OgleLMCSMCWa.csv was obtained by The Optical Gravitational Lensing Experiment (OGLE-II) for the Large and Small Magellanic Clouds. This file is in comma-separated-value format and contains the following variables:

#	Variable	Description
1	Star	catalog names for the stars, this column may be used for the row names of the data
2	MWa	an extinction-free measure of absolute magnitude (luminosity) of the star. Note: stellar magnitudes are given in a logarithmic scale, so there is no need to transform them.
3	VI	difference between the magnitude of a star in violet light and the magnitude in visible light, called the color index. It is a measure of the star's surface temperature.
4	logPeriod	base 10 logarithm of the star's pulsation period
5	Type	type of Cepheid, Fundamental Mode (FU) and First Overtone (FO); categorical variable
6	Galaxy	location of the star, LMC (Large Magellanic Cloud) and SMC (Small Magellanic Cloud); categorical variable

Notes:

- Some of these stars were misclassified into the wrong Type, so there may be some apparent outliers. Don't try to identify and remove them.
- Any potential effect of Galaxy on luminosity most likely would be due to differences in relative proportions of Cepheid types between these two galaxies and greater uncertainties associated with the greater distance of the SMC from earth, but not any fundamental physical relationship, so don't use Galaxy in your models.

The central question here is to determine how luminosity (MWa) depends on the other variables. Specifically, perform the following:

(a) Obtain an appropriate model to predict MWa based on the other variables. Check assumptions and identify influential points. [12 points]

(b) Use a variable selection method to remove unimportant variables from the model. Summarize and interpret the coefficients of the resulting model. In particular, what effect does Type have on the model? [12 points]

(c) Plot MWa vs logPeriod using different colors or plot symbols for each combination of Type and Galaxy. [6 points]

3. A state food company studied the soil fertility of five different fertility treatments. The standard, a currently used, treatment (Treatment1) and four experimental treatments (Treatments 2, 3, 4, 5) were included in the study. Eight farms were randomly selected, thus

reflecting variations in farms owned by the company throughout the state. At each farm, a random ordering of the treatments to five soil patches was employed (one treatment on each patch). After a suitable period of exposure to weather, a measure of soil fertility was obtained. The data is provided as fertility.txt (the higher the score, the better the result).

(a) Prepare a plot to show the interaction between the treatments and farms. Does this plot suggest any interaction? [5 points]

(b) Build an appropriate model based on the result of part (a). Then obtain the residuals and prepare a normal probability or quantile-quantile plot of the residuals. Summarize your findings about the appropriateness of your model. [10 points]

(c) Obtain the analysis of variance table. Then test whether or not the treatment effect is significant. [10 points]

(d) Treatments 1 and 3 were expected to be similar to each other but to differ from Treatment 4. Use the most efficient multiple comparison procedure with a 95% family-wise confidence coefficient to estimate

$$\begin{aligned} L_1 &= \mu_{1\cdot} - \mu_{3\cdot} \\ L_2 &= \frac{\mu_{1\cdot} + \mu_{3\cdot}}{2} - \mu_{4\cdot} \end{aligned}$$

Summarize your findings. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE I and II
April 2016

General Instructions: Write your ID number on all answer sheets. Do not put your name on any of your answer sheets. Show all work/proofs/references. Please write neatly so it is easy to read your solution.

Problem 1. Consider a sample of size n from the pdf $f(x|\theta) := [\pi(1+(x-\theta)^2)]^{-1}$. Find a minimal sufficient statistic.

Problem 2. Let $X_1, \ldots, X_n$ be a sample from a location family. Prove that $M - \bar{X}$ as an ancillary statistic, where M is the median and $\bar{X}$ is the sample mean.

Problem 3. Let $X_1, \ldots, X_n$ be a sample from $f(x|\theta) = [\log(\theta)/(\theta - 1)]\theta^x I(0 < x < 1) I(\theta > 1)$. Either find a complete sufficient statistic or prove that it does not exist.

Problem 4. Formulate and prove the Cramer-Rao inequality. (Do not forget to write down all assumptions.) Then explain what for it is used.

Problem 5. Formulate and prove the Rao-Blackwell inequality. (Do not forget to write down all assumptions.) Then explain what for it is used and present an example.

Problem 6. Consider a sample of size n from the cdf

$$F(x|\alpha, \beta) = (x/\beta)^\alpha I(0 \le x \le \beta) + I(x > \beta), \quad \alpha > 0,\ \beta > 0.$$

Find the MLEs of α and β.

Problem 7. Consider a sample of size n from Exponential distribution with mean λ. Suppose that the sample is missed and only the minimal observation $X_{(1)} = \min(X_1, \ldots, X_n)$ is available. Is it possible or impossible to suggest an unbiased estimate of λ?

Problem 8. Consider a sample of size n from the pdf

$$f(x|\theta,\nu) = \frac{\theta\nu^\theta}{x^{\theta+1}} I(\nu \le x < \infty),\ \theta > 0,\ \nu > 0.$$

Consider the LRT (likelihood ratio test) for $H_0 :\ \theta = 1$ and ν is unknown, versus $H_a :\ \theta \neq 1$ and ν is unknown. Indicate the critical region. Hint: Begin with finding the MLEs.

Problem 9. Consider a sample of size n from Exponential distribution with mean λ. Find a UMA $1-\alpha$ confidence interval based on inverting an UMP size α test $H_0 : \lambda = \lambda_0$ versus $H_a :\ \lambda < \lambda_0$. Hint: Recall that the sum of iid exponential random variables, appropriately rescaled, has a nice classical distribution.

Problem 10. Conside the problem of Bayes hypothesis testing of k simple hypotheses. Suggest the Bayes test and prove your assertion.

April 2016 Qualifying Exam in Probability Theory

- This is a closed-book test.
- There are 3 questions.
- Answer each question as fully as possible, showing and justifying your steps.
- Refer clearly to any known results that you are using, and **state such results precisely**.
- Indicate how the assumptions in the formulation of the question are being used in the solution.
- Write your solutions on the blank sheets of paper that are provided. Identify each sheet with your name. Begin each question on a new sheet.
- On each sheet, identify which question and part is being answered.
- *When finished, arrange your sheets in order, number each sheet, and be sure that your name is on each sheet.*

---

1. Let $X_1, X_2, \ldots$ be a sequence of identically distributed random variables defined on a probability space $(\Omega, \mathcal{A}, P)$, with $E|X_1| < \infty$ and put $EX_1 = \mu$. For each $i \geq 1$, let $Y_i = X_i I(|X_i| \leq i)$. Put $S_n = X_1 + \cdots X_n$ and $T_n = Y_1 + \cdots Y_n$.

   (a) Recall that

   $$\sum_{i=1}^{\infty} P(|Z| \geq i) \leq E|Z| \tag{1}$$

   for any random variable Z. Using (1) without proof, obtain

   $$P(X_i \neq Y_i \text{ i.o.}) = 0. \tag{2}$$

   (b) Using (2), obtain

   $$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu \iff \frac{T_n}{n} \xrightarrow{\text{a.s.}} \mu. \tag{3}$$

   (c) Define tail event and verify that the event

   $$A = \left\{ \omega : \frac{S_n(\omega)}{n} \to \mu, \ n \to \infty \right\}$$

   is a tail event.

2. Let $\{\xi_n\}$ and $\{\eta_n\}$ be sequences of random variables satisfying $\xi_n \xrightarrow{p} \alpha$ and $\eta_n \xrightarrow{p} \beta$ as $n \to \infty$.

(a) Prove that $\xi_n + \eta_n \xrightarrow{p} \alpha + \beta$ or give a counter-example.

(b) For a sample of independent random variables $\{X_1, \ldots, X_n\}$ from a distribution F having mean μ and finite variance σ^2, consider the sample second central moment,

$$m_2 = n^{-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2,$$

where $\overline{X}_n$ denotes the sample mean $n^{-1} \sum_{i=1}^{n} X_i$. Prove that

$$m_2 \xrightarrow{p} \sigma^2.$$

(c) Does m_2 converge to σ^2 in distribution? in mean square? almost surely? completely?

3. Consider two measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$, where $\Omega_1 = \Omega_2 = \mathbb{R}$ = the real line, and $\mathcal{A}_1 = \sigma(\mathcal{E}_1)$ and $\mathcal{A}_2 = \sigma(\mathcal{E}_2)$, with $\mathcal{E}_1$ the class of sets $\{[0,1],[2,3]\}$ and $\mathcal{E}_2$ the class $\{\emptyset\}$.

   (a) Determine $\mathcal{A}_1$ and $\mathcal{A}_2$ explicitly.

   (b) Define in general the *product measurable space* $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2)$ and determine it explicitly for the above special case.

   (c) Define in general the section $A^{(1)}(\omega_1)$, for A in $\Omega_1 \times \Omega_2$ and $\omega_1 \in \Omega_1$, and determine $A^{(1)}(\omega_1)$ explicitly for the above special case with some particular choice of A and particular choice of ω_1.

   (d) Prove or give counter-example: *the section of a union is the union of the sections.*

   (e) Prove or give counter-example: *the section of an intersection is the intersection of the sections.*

Ph.D. Qualifying Exam: Spring 2016
Linear models

- Number of questions = 3. Answer all of them. Total points = 60.
- Simplify your answers as much as possible and carefully justify all steps to get full credit.
- There is no need to prove any standard result. Just state the result and use it.
- You can use a calculator.
- **All vectors are column vectors.**

---

1. Consider the set up of linear regression of a covariate X on response Y. The data, denoted as (x_i, y_i), $i = 1, \ldots, n$, consist of observations of (X, Y) from n independent subjects. Define the summary statistics,

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2,\ S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2,\ S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Our goal is to fit a line $y = \alpha + \beta x$ to describe the points (x_i, y_i), $i = 1, \ldots, n$. In ordinary linear regression, the covariate is assumed to be measured without error and the method of least squares is used to estimate the coefficients. This method minimizes the sum of squares of vertical distances between the points and the line. Here, however, we assume that the covariate is measured *with* error. In this case, it is reasonable to use the method of *orthogonal least squares*, which minimizes the sum of squares of perpendicular distances between the points and line (see Figure 1). From basic geometry, we can see that the perpendicular from (x_i, y_i) intersects the line $y = \alpha + \beta x$ at the point $(\tilde{x}_i, \tilde{y}_i)$, where

$$\tilde{x}_i = (\beta y_i + x_i - \alpha\beta)/(1 + \beta^2),\ \tilde{y}_i = \alpha + \beta\tilde{x}_i.$$

(a) Show that the criterion to be minimized in the method of orthogonal least squares is [10 points]

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2/(1 + \beta^2).$$

(b) Show that the orthogonal least squares estimates of α and β are [15 points]

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x},\ \hat{\beta} = \frac{-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

2. Let $\mathbf{Y} \sim N_n(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ and let $Q_k = (\mathbf{Y} - \boldsymbol{\theta})'\mathbf{P}_k(\mathbf{Y} - \boldsymbol{\theta})/\sigma^2$, $k = 1, 2$. Assume that $Q_k \sim \chi^2_{r_k}$, $Q_1 - Q_2 \geq 0$, and $r_1 - r_2 > 0$.

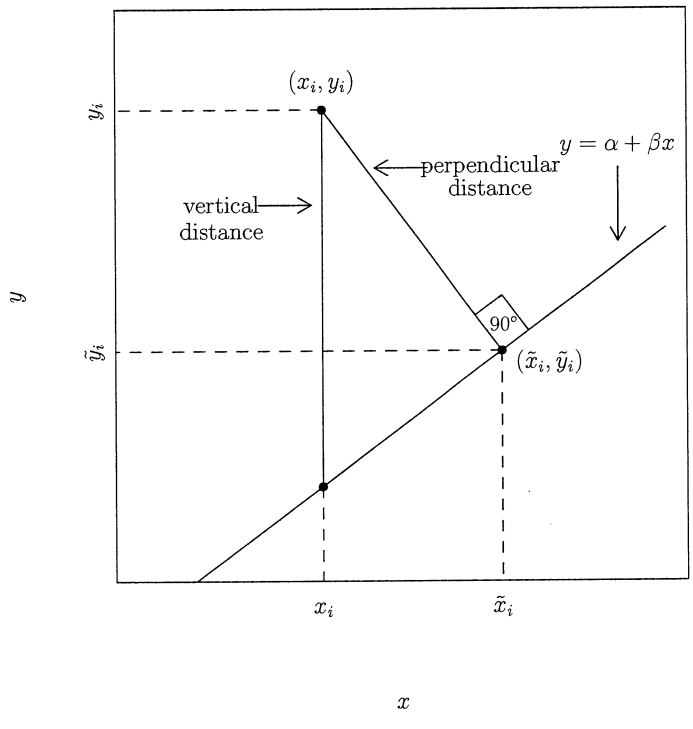

Figure 1: Vertical and perpendicular distances of a point from a line. The former is used in ordinary least squares whereas the latter is used in orthogonal least squares.

(a) Show that $Q_1 - Q_2 \sim \chi^2_{r_1 - r_2}$ [10 points]

(b) Show that $Q_1 - Q_2$ is independent of Q_2. [10 points]

[*Hint*: If $\mathbf{P}_k$, $k = 1, 2$ are symmetric idempotent matrices and $\mathbf{P}_1 - \mathbf{P}_2$ is positive semi-definite, then $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_2$.]

3. Assume that $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_4(\mathbf{0}, \sigma^2\mathbf{I})$ and $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 0$. We would like to test the null hypothesis $H : \theta_1 = \theta_3$ against the alternative $K : \theta_1 \neq \theta_3$.

(a) Show that the relevant F-statistic is [12 points]

$$\frac{2(Y_1 - Y_3)^2}{(Y_1 + Y_2 + Y_3 + Y_4)^2}.$$

(b) What is the null distribution of the test statistic in (a)? [3 points]
