Qualifying Exam 2015 — 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 and 3 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 codes or outputs.
- Write your name on the report. DO NOT write your ID.

1. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be two independent samples of sizes n_1 and n_2 from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ populations, respectively. Let $(\bar{X}_1,\ S_1^2)$ and $(\bar{X}_2,\ S_2^2)$ denote their sample means and variances. Consider two commonly used 95% t-confidence intervals (CI) for $\mu_1 - \mu_2$. One is

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.975, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-1}$ is the pooled sample variance. It is used when $\sigma_1^2 = \sigma_2^2$ can be assumed. The other is

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.975, \nu} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where ν is given by Satterthwaite's approximation as

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

It is used when $\sigma_1^2 = \sigma_2^2$ cannot be assumed. We would like to understand the behavior of these CIs, in particular, how their coverage probabilities are affected by (n_1, n_2) and (σ_1^2, σ_2^2), so that we may deduce which CI should be used when and what happens if the wrong CI is used. Conduct a Monte Carlo study with 1,000 replications to investigate this. Use the following settings and compute both CIs for each sample.

(a) $n_1 = n_2 = 20$, $\sigma_1^2 = \sigma_2^2 = 4$.

(b) $n_1 = n_2 = 20$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$.

(c) $n_1 = 20, n_2 = 10$, $\sigma_1^2 = \sigma_2^2 = 4$.

(d) $n_1 = 20, n_2 = 10$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$.

(e) $n_1 = 10, n_2 = 20$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$.

Set $\mu_1 = \mu_2 = 10$. For starting seed value, use the last 5 digits of your ID. Clearly state your overall conclusions. [40 points]

2. "Uric Acid and Cardiovascular Risk Factors": The cardio dataset contains data on 998 individuals on the following variables (Source: Heritier et al., Robust Methods in Biostatistics, 2009; the original dataset contains more variables):

#	Variable	Description
1	uric	Uric acid level
2	dia	Diastolic blood pressure
3	hdl	High-density lipoprotein cholesterol
4	choles	Total cholesterol
5	trig	Triglycerides level in body fat
6	alco	Alcohol intake (ml per day)

(a) Fit a full model for predicting uric acid levels using all other explanatory variables. Test if the variables hdl and choles can be (jointly) dropped together from the full model. State your conclusion. [6 points]

(b) Find the best model(s) using adjusted R^2 criterion and stepwise selection method. [5 points]

(c) For the best model chosen above, detect any outliers and/or influential points using appropriate diagnostic tools. Based on these, comment on the appropriateness of the model. Suggest a better alternative model (Just state it, no need to fit it). [9 points]

3. "Personality Traits": This data set contains measures of personality traits of 600 students in 3 schools. The 7 variables are listed as follows:

#	Variable	Description
1	id	Identification number of student
2	school	School ID
3	extro	Extroversion score: outgoing/energetic vs. solitary/reserved
4	extrolevel	Extroversion level: high (2) or low (1)
5	open	Openness to new experiences: inventive/curious vs. consistent/cautious
6	agree	Agreeableness: friendly/compassionate vs. analytical/detached
7	social	Social engagement: active vs. inactive participation in community or society

For this part, our response variable is extroversion scores (**extro**).

(a) First, we examine the effects of schools (school) on the extroversion (extro). [10 points]

i. Test whether or not the average extroversion scores are equal in these three schools. Use $\alpha = 0.05$. Verify the assumptions used with the model. If the assumptions are not satisfied, is it safe to use the modeling/test results? Justify your answer.

ii. Obtain confidence intervals for all pairwise comparisons among the three schools at the family-wise confidence level of 90%. Interpret your result and state your conclusions.

(b) Next, we study the effects of openness (open) and agreeableness (agree) in addition to the school effect. Here openness is to be classified into two categories (< 40, ≥ 40) and agreeableness is to be classified into two groups (< 35, ≥ 35). [10 points]

i. Conduct an appropriate ANOVA.

ii. Is the interaction between openness and agreeableness significant?

iii. Conduct tests for the main effects (using $\alpha = .05$).

(c) Last, let us assume that the three schools are a random sample from a large pool of many schools. Assume everything else is the same as in Question (3b). [10 points]

i. Fit an appropriate model concerning all the main effects. Which categorical variables are significant ($\alpha = 0.05$) for modeling the extroversion?

Now we are interested in the extroversion level (**extrolevel**). Please consider factors (school, open, agree, social) that may affect it. Note that we **do NOT** dichotomize any independent variables for this part.

(d) Develop the best model you can. Test whether any interaction exists between open, agree and social. Conduct a goodness of fit test for your final model. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE I and II
April 2015

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 $f(x|\theta) := e^{-(x-\theta)}I(-\infty < \theta < x < \infty)$. Find a minimal sufficient statistic, show that it is complete, and then establish if it is independent of the sample variance.

Problem 2. Formulate and prove Basu's Theorem. Then explain when it is used and present an example.

Problem 3. Formulate and prove the Cramer-Rao inequality. (Do not forget to write down all assumptions.) Then explain what for it is used.

Problem 4. 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 5. Prove that if W is a best unbiased estimator of parameter θ then W is unique.

Problem 6. Consider a sample of size n from the pdf $f(x|\theta) = \theta x^{-4}I(0 < \theta \leq x < \infty)$. Find the MLE of θ.

Problem 7. Suppose that we have two independent random samples $X_1, \ldots, X_n$ from the exponential(θ) and $Y_1, \ldots, Y_n$ from the exponential(μ).
(a) Find the Likelihood Ratio Test of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$.
(b) Show that the test can be based on the statistic
$T = [\sum_{i=1}^{n} X_i]/[\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_i]$.

Problem 8. Suppose that a sample is from the pdf
$f(x|\theta) = h(x)c(\theta)e^{w(\theta)x}I(x > 0)$ where $w(\theta)$ is an increasing function of θ.
(a) Is this family has an MLR?

(b) Explain why this question is important in hypothesis testing and confidence interval estimation.

Problem 9. Consider a sample from the Poisson (λ). Find a UMP $1-\alpha$ confidence interval based on inverting the UMP level α test of $H_0: \lambda = \lambda_0$ versus $H_a: \lambda > \lambda_0$.

Problem 10. Let we observe a sample from Y where $Y = \min(X, d)$. Here d is a known constant and X is Normal(θ, σ^2) where σ^2 is known. Propose an asymptotically efficient scoring estimate of θ. [Remark: At least define the estimator and explain main steps in its construction.]

Last Name: _____ First Name: _____

2015 Qualifying Exam in Probability Theory

There are 3 problems. Answer as many questions as you can, showing and justifying your steps. Refer clearly to any known results that you are using. This is a closed-book test.

1. Let $X_1, X_2, \ldots$ be independent random variables, where X_n has Exponential distribution with mean $(1/n)$ and density

$$f_n(x) = ne^{-nx}, \quad x > 0.$$

Which of the following events occur with probability 1? (Justify your answers.)

(a) $X_n < 1$ for infinitely many n.
(b) $X_n > 1$ for infinitely many n.
(c) $\limsup \{X_n < 1\}$.
(d) $\liminf \{X_n < 1\}$.

2. Let $\mathscr{A}$ be a collection of all Borel sets in $\mathbb{R}$ whose positive and negative parts have the same Lebesgue measure λ. That is,

$$\mathscr{A} = \{A \in \mathscr{B} \ : \ \lambda\{A \cap (-\infty, 0)\} = \lambda\{A \cap (0, \infty)\}\}.$$

(a) Is $\mathscr{A}$ an algebra?

(b) Is $\mathscr{A}$ a monotone class?

(c) What is the smallest sigma-algebra containing $\mathscr{A}$?

3. Let $\xi_0, \xi_1, \xi_2, \ldots$ be independent random variables with the distribution

$$P(\xi_n = 1) = 3/4, \quad P(\xi_n = -1) = 1/4.$$

A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ is defined recursively as

$$X_0 = 1, \quad \text{and for } n \geq 0, \; X_{n+1} = \xi_n X_n.$$

(a) Show that $\{(2^n X_n), \mathscr{F}_n\}$ is a martingale, where $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$ is a sigma-algebra generated by $X_1, \ldots, X_n$.

(b) Use this result to find the expected value $E(X_n)$ for any n.

(c) Use (b) to find the probability mass function of X_n for any n. (Notice that $P(|X_n| = 1) = 1$.)

(d) Does the sequence $\{X_n\}$ converge almost surely to some random variable? Does it converge in probability? Does it converge in distribution? Justify your answers.

Ph.D. Qualifying Exam: Spring 2015

Linear models

- Number of questions = 4. Answer all of them. Total points = 100.
- 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 it and use it.

---

1. Suppose $\mathbf{X} = (X_1, X_2, X_3)' \sim N_3(0, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix. Define $\mathbf{Y} = (X_1, \sqrt{2}X_2, \sqrt{2}X_3)'$ and

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/4 & -1/4 \\ 0 & -1/4 & 1/4 \end{pmatrix}$$

   (a) Derive the distribution of

$$U = \frac{1}{6}\left(X_1^2 + 4X_2^2 + X_3^2 - 4X_1X_2 + 2X_1X_3 - 4X_2X_3\right)$$

   (b) Derive the distribution of $V = \mathbf{Y}'\mathbf{A}\mathbf{Y}$.

   (c) Are U and V independent? Justify your answer.

2. Suppose $Y_1, Y_2, \ldots, Y_{3n}$ are observed from the model

$$Y_i = \beta_0 + \beta_1 + \beta_2 + \epsilon_i,\ Y_{n+i} = \beta_0 + \beta_1 + \epsilon_{n+i},\ Y_{2n+i} = \beta_0 + \beta_2 + \epsilon_{2n+i},\ i = 1, 2, \ldots, n.$$

where $\epsilon_1, \ldots, \epsilon_{3n}$ are i.i.d. with mean 0 and variance σ^2. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$.

   (a) Is $\boldsymbol{\beta}$ estimable? If yes, justify, and if not, explicitly find a column vector $\boldsymbol{c}$ such that $\boldsymbol{c}'\boldsymbol{\beta}$ estimable.

   (b) Suppose we want to test the null hypothesis $H_0 : \beta_1 = \beta_2$ against the alternative hypothesis $H_1 : \beta_1 \neq \beta_2$ assuming normality for the errors. Derive the test statistic, its distribution under H_0, and provide the critical region for the test.

3. Consider a logistic regression model for a binary response Y (success or failure) and a single explanatory variable X. Suppose the estimates of the regression coefficients are $\widehat{b}_0$ (intercept) and $\widehat{b}_1$ (slope), with respective estimated standard errors s_{b_0} and s_{b_1}.

   (a) Can you use this information to assess whether X is a significant predictor? Justify your answer.

   (b) Suppose we wish to compare the effect on probability of success for subjects with $X = a$ to subjects with $X = b$. Provide a suitable quantitative measure to do this and find its expression in terms of a and b.

4. Suppose a study is conducted to relate the number of children ever born to married women of a certain race with 4 explanatory variables $X_1, \ldots, X_4$.

   (a) Write an appropriate model and list the model assumptions.

   (b) How can you test whether the third explanatory variable X_3 is a significant predictor? Find an appropriate test statistic and its distribution under the null hypothesis.

   (c) Suppose the researcher wants to have some flexibility in the model to allow for either overdispersion or underdispersion, whichever the case may be. As a statistician what will you recommend the researcher?