STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE I and II
April 2014

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

Problem 1. Consider a sample of size n from Uniform$(\theta, \theta+1)$, $\theta \in (-\infty, \infty)$. Find the minimal sufficient statistic (please prove your assertion).

Problem 2. Consider a sample of size n from Uniform$(0, \theta)$. Find a complete sufficient statistic and prove your assertion.

Problem 3. Formulate and prove Basu's Theorem.

Problem 4. Consider a sample of size n from Normal$(\theta, 1)$, $\theta \in [0, \infty)$. Find the maximum likelihood estimate.

Problem 5. Let $f(x|\theta)$ be the logistic pdf,

$$f(x|\theta = \frac{e^{x-\theta}}{(1+e^{x-\theta})^2}, \quad x \in (-\infty, \infty), \quad \theta \in (-\infty, \infty).$$

(a) Does the family have an MLR? (b) Based on one observation, find the UMP size α test for $H_0: \ \theta \leq 0$ versus $H_1: \ \theta > 0$.

Problem 6. Pivoting an exponential density $f(x|\theta) = e^{-(x-\theta)} I(x > \theta)$ find a $1-\alpha$ confidence interval for θ.

Problem 7. Show that if $E_\theta(\delta) = g(\theta)$ and $Var_\theta(\delta)$ attains the information inequality lower bound then

$$\delta(X^n) = g(\theta) + \frac{g'(\theta)}{I(\theta)} \partial \ln(f_\theta(X^n))/\partial\theta.$$

Problem 8. Consider a density $f_\theta(x)$ from an exponential family of distributions with θ being the natural parameter. Show that the likelihood equation has a unique root.

Problem 9. Let X be a Cauchy(0,1) random variable (i.e., it has density $f(x) = [\pi(1+x^2)]I(x \in (-\infty, \infty))$. Set $Y_n = \cos(X/n)$. Show that Y_n converges in probability and determine the limit as $n \to \infty$.

Problem 10. Let $F(x)$ and $G(x)$ be two known cumulative distribution functions on the real line. Consider a mixture distribution $H_\theta(x) = \theta F(x) + (1-\theta)G(x)$, $\theta \in (0,1)$.
(a) For a single observation X from H_θ, find a UMP test of size α for testing $H_0 : \theta \leq \theta_0$ versus $H_a : \theta > \theta_0$.
(b) Based on a sample of size n from H_θ, propose an asymptotically efficient scoring estimate of θ.
Remark: If it is difficult to present complete solutions for (a) and (b), outline main steps.

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE I and II
April 2013

General Instruction: Write your ID number on all answer sheets. Do not put your name. Show all work. Please write neatly so it is easy to read your solution.

Problem 1. (a) Formulate and prove Basu's Theorem. Do not forget to give definitions to all statistics involved.
(b) Consider a sample from expomemntial distribution. Calculate expectation of $(X_1 + X_n)/[X_1 + X_2 + \ldots + X_n]$.

Problem 2. Consider an exponential family with the pdf for a random variable X,

$$f(x|\theta) = h(x)c(\theta)\exp(\sum_{i=1}^{k} w_i(\theta)t_i(x)).$$

Here θ is a vector-valued parameter.
(a) Can you remember why this family is so convenient to deal with in the case of employing the Basu's Theorem?
(b) Prove that

$$\mathrm{Var}_\theta(\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)) = -\frac{\partial^2 \log(c(\theta)}{\partial \theta_j^2} - E\{\sum_{i=1}^{k} \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X))\}$$

Problem 3. (a) Prove the following assertion which is referred to as Stein's lemma and which is a good example of using integration by parts. Let X be Normal(θ, σ^2), and let $g(x)$ be a differentiable function satisfying $E\{|g'(X)|\} < \infty$. Then

$$E\{g(X)(X - \theta)\} = \sigma^2 E\{g'(X)\}.$$

(b) Use this result to calculate $E\{X^3\}$.

Problem 4. (a) Prove Rao-Blackwell Theorem which asserts that if $\delta(\mathbf{X})$ is an unbiased estimate of θ and T is a sufficient statistic for θ then there exists an estimator which is a function of T and its MSE is not larger than

the estimator's MSE uniformly over all θ.
(b) Let we have a sample of size n from $Uniform([-\theta, \theta])$ distribution. Find, if one exists, the UMVU estimator.

Problem 5. Let we have a sample of size n from the pdf $f(x|\theta) = \theta x^{\theta-1} I(0 < x < 1)$, $\theta \in (0, \infty)$. Find the MLE estimator for the estimand $g(\theta) = \cos(\theta)$.

Problem 6. Consider a family of distributions $\{P_\theta,\ \theta \in \{\theta_1, \theta_2, \ldots, \theta_k\}\}$.
(1) Formulate a proposition that will allow you to solve hypothesis testing problems under Bayesian, Most Powerful, and Minimax approaches. Prove the Most Powerful part assuming that the Bayesian part is valid.
(ii) Use this proposition to propose a minimax solution for the model where $X = \theta + Z$ with Z being an exponential and $\theta \in \{\theta_1, \theta_2, \theta_3\}$ where $\theta_1 < \theta_2 < \theta_3$.

Problem 7. Consider a sample of size n from exponential distribution with mean λ. Then $T = \sum_{l=1}^{n} X_l$ has Gamma distribution with the pdf

$$f(x|\lambda) = \frac{1}{\Gamma(n)\lambda^n} x^{n-1} e^{-x/\lambda}.$$

Suggest $(1-\alpha)$ -confidence interval for λ using the methodology of pivoting.

Problem 8. Let we have a sample of size n from a distribution with the cdf

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

Find the MLE's of the two parameters.

Problem 9. Let X be one observation from the distribution with density $f(x|\theta) = \pi^{-1}(1 + (x-\theta)^2)^{-1} I(-\infty < x < \infty)$. Consider a critical function $\phi(x) = I(1 < x < 3)$.
(a) Show that it is the MP critical function (test) of its size for testing $H_0 :\ \theta = 0$ versus $H_1 :\ \theta = 1$, or disprove this assertion.
(b) Calculate Type I and Type II error probabilities for the test.

Problem 10. Suppose that Y_i, $i = 1, 2, \ldots, k$ are independent Poisson(μ_i) where $\mu_i = \beta N_i$. Here β is unknown and $N_1, \ldots, N_k$ are fixed known constants.

(a) Find the maximum likelihood estimate (MLE) for β.

(b) An alternative estimate is $\beta^* = \sum_{i=1}^{k} (Y_i/N_i)$. Compare variances of this estimate and the MLE, and discuss the outcome.

Last Name: _____ First Name: _____

2014 Qualifying Exam in Probability Theory

There are 3 problems. Answer as many questions as you can, showing 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 with

$$P(X_n = n^2 - 1) = n^{-2}, \quad P(X_n = -1) = 1 - n^{-2}, \quad n \geq 1.$$

(a) Show that $\sum_{n=1}^{\infty} \frac{X_n+1}{n}$ converge almost surely.

(b) Let $S_n = X_1 + \cdots + X_n$. Show that $S_n/n \to -1$ a.s. *(10 points)*

2. Let $\{A_n, n = 1, 2, \ldots\}$ be a sequence of events in a probability space $(\Omega, \mathcal{A}, P)$. Denote by $\{A_n \text{ i. o.}\}$ the set of $\omega \in \Omega$ such that $A_n(\omega)$ occurs for infinitely many n. Denote the complement of an event B by B^c.

(i) Show that $\{A_n \text{ i. o.}\} \subset \cup_{n=m}^{\infty} A_n$, for any m.

(ii) Show that $\{A_n \text{ i. o.}\}^c = \cup_{m=1}^{\infty} \cap_{n=m}^{\infty} A_n^c$, for any m.

(iii) Use (i) to show that

$$P(\{A_n \text{ i. o.}\}) \leq \sum_{n=m}^{\infty} P(A_n), \quad \text{for any } m.$$

(iv) Use (ii) to show that, if the events A_n are *mutually independent*, then

$$P(\{A_n \text{ i. o.}\}^c) \leq \sum_{m=1}^{\infty} e^{-\sum_{n=m}^{\infty} P(A_n)}.$$

(v) Use (iii) to give a sufficient condition for $P(\{A_n \text{ i. o.}\}) = 0$.

(vi) Use (iv) to give a sufficient condition for $P(\{A_n \text{ i. o.}\}) = 1$. *(20 points)*

3. Suppose we deposit $X_0 = \$1000$. Every month, the rate of change is random, uniformly distributed between -10% and 10%. It means that X_n, the amount after n months, is

$$X_n = X_0 \cdot \prod_{k=1}^{n} \xi_k,$$

where $\xi_1, \xi_2, \ldots$ are i.i.d. Uniform(0.9, 1.1) random variables.

(a) Show that $(X_n, \mathcal{A}_n)$ is a martingale, where $\mathcal{A}_n$ is a sigma-algebra generated by $\xi_1, \ldots, \xi_n$.

(b) We shall withdraw the entire amount as soon as it reaches \$2000 or after 36 months, whichever occurs first. Find the expected amount at the time of withdrawal. Are we expected to gain or lose with this investment?

(c) Show without a calculator that $E(\log \xi_k) < 0$, and therefore, $\log X_n = \log X_0 + \sum_{k=1}^{n} \log \xi_k$ is a random walk with a negative drift.

(20 points)

Ph.D. Qualifying Exam in Statistical Methods

April 11, 2014

- Exam consists of two projects.
- Load the data sets from the web site **http://www.utdallas.edu/~mbaron/QExam41114/**. Let the proctor know if you have any problems with this step.
- Conduct the necessary data analysis using *the software of your choice* and answer as many questions as you can precisely and accurately. Your solutions should be based on the appropriate statistical methods.
- Conduct regression and ANOVA diagnostics when necessary. If some required assumptions are violated, make an attempt to fix the situation. If this is difficult to do, state so.
- Submit a report, written or typed, hard copy or e-mail. If you choose to e-mail the report, send it to `ammann@utdallas.edu`, `ygl@utdallas.edu`, and `mbaron@utdallas.edu`.
- In the report, describe every step of your analysis: method, reasons, and results. For example:

  Test significance of variable XXX. Use ... ... The F test gives a p-value of 0.0003. Therefore, ...

  Verify assumptions of the test. Use ... ... Variable ... violates assumption ... because ... Therefore, ...

- Attach your computer programs and only relevant parts of the output. Do not attach the parts of output that were not used to answer questions.

Project 1 "Compressive Strength of Concrete"

Concrete is the most important material in civil engineering. Its compressive strength is a function of age and its ingredients. This dataset contains the following variables.

1. Cement (component 1) – quantitative – kg in a m3 mixture – Input Variable
2. Slag (component 2) – quantitative – kg in a m3 mixture – Input Variable
3. FlyAsh (component 3) – quantitative – kg in a m3 mixture – Input Variable
4. Water (component 4) – quantitative – kg in a m3 mixture – Input Variable
5. Superplasticizer (component 5) – quantitative – kg in a m3 mixture – Input Variable
6. CoarseAggregate (component 6) – quantitative – kg in a m3 mixture – Input Variable
7. FineAggregate (component 7) – quantitative – kg in a m3 mixture – Input Variable
8. Age – quantitative – Day (1 365) – Input Variable
9. Concrete compressive strength – quantitative – MPa – Output Variable

Exam Questions

1. First, let us predict **compressive strength**.
   (a) Which variables are significant for this prediction?
   (b) What formula do you recommend for the prediction of compressive strength?
   (c) Which variables cause most of the collinearity, and what can be done about it?
   (d) Test for outliers, keeping the experimentwise error rate within a 5% level.
2. The American Concrete Institute defines high-strength concrete as concrete with compressive strength greater than 50 MPa. Construct a model that predicts whether or not concrete will be high-strength.
   (a) Which variables are significant for this prediction?
   (b) What formula do you recommend to predict whether or not a mix will produce high-strength concrete?
   (c) Use your model to predict whether or not each mix in this data set will be high-strength and compare your predictions to how they actually would be rated.
3. Include the proper regression diagnostics. Apply transformations if necessary.

Project 2 "Working Hours"

This is an extraction from the 1994 Census database. The social study of overworked employees focuses on the working hours per week. This dataset contains the following variables.

1. AGE - quantitative
2. WORKCLASS - categorical
3. EDUCATION - categorical
4. MARITAL_STATUS -categorical
5. OCCUPATION - categorical
6. RACE - categorical
7. GENDER - categorical
8. INCOME - categorical (>50K, <=50K)
9. HOURS_PER_WEEK - quantitative

Exam Questions

1. Does the number of work hours depend on the age? Test the lack of fit of this linear model.
2. Which categorical variables are significant for modeling the number of work hours? Are there significant interactions?
3. Is there a significant dependence between occupation and gender? Between occupation and marital status?
4. Compare the mean working hours of three groups of people - (a) with some college education or professional school, (b) high school graduates, (c) incomplete school education. Construct Tukey confidence intervals for the mean differences between these groups.

Ph.D. Qualifying Exam: Spring 2014
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. Consider a regression model with 3 regression parameters β_1; β_2 and β_3, and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)'$; $\boldsymbol{\varepsilon} = (\varepsilon_1; \varepsilon_2; \varepsilon_3; \varepsilon_4)' \sim N(\mathbf{0}; \sigma^2 I_4)$; I_4 is an identity matrix of order 4; and

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ -1 & 1 & -2 \\ 1 & -1 & 2 \end{pmatrix}.$$

(a) Are all linear functions of the parameter $\boldsymbol{\beta}$, $\mathbf{c}'\boldsymbol{\beta}$ ($\mathbf{c} \neq \mathbf{0}$), estimable? If yes, why? If not, why not, and under what conditions $\mathbf{c}'\boldsymbol{\beta}$ is estimable?

(b) Can you test the hypothesis H_0: $\frac{1}{2}\beta_1 = \beta_2$? If yes, develop the test statistic. Justify your answer.

(c) Can you test the hypothesis H_0: $\beta_1 = \frac{1}{2}(\beta_2 - 4\beta_3)$ and $\beta_3 = -\frac{1}{2}\beta_1$? If yes, develop the test statistic. Justify your answer.

2. (Continuation of Question 1(c)) (a) What is the distribution of the test statistic under the null hypothesis?

(b) What is the distribution of the test statistic under a general scenario (not necessarily under the null hypothesis)?

(c) Assume that we observe $\mathbf{Y} = (2, 2, 2, 2)'$. Construct a 95% confidence interval for $\beta_1 - \frac{1}{2}(\beta_2 - 4\beta_3)$.

3. Suppose $\mathbf{X} = (X_1, X_2)'$ has a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix, $\Sigma = (1 - \rho)I_2 + \rho J_2$. Here, I_2 is an identity matrix of order 2, and J_2 is a 2×2 matrix of 1's.

(a) Derive the range of ρ.

(b) Suppose $Q_1 = (X_1 - X_2)^2$ and $Q_2 = (X_1 + X_2)^2$. Find the distributions of Q_1 and Q_2?

(c) Are Q_1 and Q_2 be distributed independently? Justify your answer.

4. Develop a test for testing whether the following K regression lines are parallel or not. The K regressions lines are:

$$y_{ki} = \alpha_k + \beta_k x_{ki} + \varepsilon_{ki};\ k = 1, 2, ..., K;\ i = 1, 2, ..., n_k,$$

where $\varepsilon_{ki} \sim$ independent $N(0, \sigma^2)$, and σ^2 is known.