Ph. D. Qualifying Exam
Tuesday, January 3, 2012

Put your solution to each problem on a separate sheet of paper.

Problem 1. Applied Statistics (STA5166)
(**Note: Need a Chi-Square table for this problem.**)

A study was conducted to determine whether the age of customers is related to the type of movie he or she watches. A sample is shown in the following table.

Age	Documentary	Comedy	Mystery
12-20	24	19	18
21-40	35	46	58
41 and over	23	60	49

(a) Given that the total sample size $n = 332$ is fixed, what is the distribution of the nine categories? What are the mean value and variance of each cell frequency? Run a chi-square test of independence and draw your conclusion. Use $\alpha = 0.05$.

(b) Run a chi-square test of "Comedy" verse "Mystery". Use $\alpha = 0.05$.

Problem 2. Applied Statistics (STA5166)

In a complete factorial experiment, consider the following random-effects model:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \\ & \qquad i = 1, \ldots, a; \quad j = 1, \ldots, b; \quad k = 1, \ldots, r, \end{aligned}$$

where μ is the overall mean, the random effects α_i, β_j, and $(\alpha\beta)_{ij}$ are assumed to be independent and normally distributed with means of zero and variances σ^2_α, σ^2_β, and $\sigma^2_{\alpha\beta}$, respectively. The random errors $\{\epsilon_{ijk}\}$ are assumed to be independent and normally distributed with mean zero and variance σ^2. Furthermore, $\{\alpha_i\}$, $\{\beta_j\}$, $\{\alpha\beta)_{ij}\}$, and $\{\epsilon_{ijk}\}$ are assumed to be independent of one another.

Define

$$SS(AB) = r \sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$$

and

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (Y_{ijk} - \bar{Y}_{ij\cdot})^2.$$

(a) What are the distributions of $\bar{Y}_{ij\cdot}$, $\bar{Y}_{i\cdot\cdot}$, and $\bar{Y}_{\cdot j\cdot}$? Find the covariance $\mathrm{Cov}(\bar{Y}_{ij\cdot}, \bar{Y}_{uv\cdot})$.

(b) Calculate the mean values of $SS(AB)$ and SSE. Find unbiased estimates for the variances $\sigma^2_{\alpha\beta}$ and σ^2 based on the observations $\{Y_{ijk}, \quad i = 1, \ldots, a; \;\; j = 1, \ldots, b; \;\; k = 1, \ldots, , r\}$.

Problem 3. Applied Statistics (STA5167)

Consider the linear regression model:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $Y = (y_1, \ldots, y_n)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, and X is an $n \times p$ full-rank matrix. $\{\epsilon_i, i = 1, \ldots, n\}$ are assumed to be iid $N(0, \sigma^2)$ variables. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$.

(a) If $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$ has a multivariate normal distribution with mean zero and nonsingular covariance matrix $V = \text{Cov}(\boldsymbol{\xi})$, and if A is a $n \times n$ symmetric matrix with rank r, show that $\boldsymbol{\xi}' A \boldsymbol{\xi}$ has a χ^2 distribution with $df = r$ if AV is idempotent. (Hint: If AV is idempotent and $V = C'C$ where C is nonsigular, then $B = CAC'$ is idempotent with rank r).

(b) Using the result in (a), find the distribution of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(X'X)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

Problem 4. Applied Statistics (STA5167)

Consider the linear regression model:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\xi},$$

where $Y = (y_1, \ldots, y_n)'$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and X is an $n \times p$ full-rank matrix. The process $\{\xi_i\}$ is generated by the model:

$$\xi_i + 0.5\xi_{i-1} = a_i,$$

where $\{a_i, i = 0, \pm 1, \pm 2 \ldots, \}$ are iid $N(0, \sigma^2)$ variables.

(a) Show that $\xi_i = \sum_{j=0}^{\infty} (-0.5)^j a_{i-j}$ is a solution of the equation $\xi_i - 0.5\xi_{i-1} = a_i$.

(b) Based on the expression in (a) for ξ_i, calculate autocorrelations $\gamma_k = \text{Cov}(\xi_i, \xi_{i+k})$ for $k \geq 0$. Show that $\gamma_k = -0.5\gamma_{k-1}$ for any $k \geq 0$.

(c) How do you estimate $\boldsymbol{\beta}$ in this setting? Discuss the properties of your estimate such as mean, covariance matrix, and distribution of $\hat{\boldsymbol{\beta}}$. Compare your estimate with the least squares estimate of $\boldsymbol{\beta}$.

Problem 5. Advanced Probability and Inference (STA6346)

Let $N(t)$ be a homogeneous Poisson process with rate λ. Find $E[N(t)N(s)]$ for $s \neq t$.

Problem 6. Advanced Probability and Inference (STA6346)

Let $X_1, X_2, \ldots$ be the positions of a random walk on the integers starting from 0. That is, X_1 is the position of the random walk after the first step, X_2 is the position after the second step, and so on. Let $P_{i,i+1} = p \in (0, 1)$ be the probability the walk moves from i to $i+1$ for any integer i.

(a) Find $E(X_n)$.

(b) For what values of p is $X_1, X_2, \ldots$ a martingale? Prove your answer. You may assume that $E|X_i|$ is finite.

(c) Create a new process Y_n from $X_1, X_2, \ldots, X_n$ such that $Y_1, Y_2, \ldots, Y_n$ is a martingale for ANY fixed $p \in (0, 1)$.

Problem 7. Advanced Probability and Inference (STA6346)

Suppose μ, ν and λ are three σ-finite measures on a measureable space $(\Omega, \mathcal{F})$.

(a) Find $\frac{d\mu}{d\mu}$.

(b) Suppose all three measures are absolutely continuous to each other. Show $\frac{d\mu}{d\nu} = \left[\frac{d\nu}{d\mu}\right]^{-1}$.

Problem 8. Computational Methods in Statistics I (STA5106)

Let H be an $n \times n$ householder matrix given by

$$H = I_n - 2\frac{vv^T}{v^Tv},$$

for any non-zero n-length column vector $v(\neq 0)$. Show that H is a symmetric, orthogonal, and reflection matrix. That is, H satisfies i) $H = H^T$, ii) $HH^T = I_n$, iii) $\det(H)$ = -1.

Problem 9. Computational Methods in Statistics I (STA5106)

Derive an EM algorithm to find the maximum likelihood estimate of θ where θ is a parameter in the multinomial distribution:

$$(x_1, x_2, x_3, x_4) \sim M(n; 0.2\theta, 0.2(2+\theta), 0.6(1-2\theta), 0.8\theta)$$

Choose a variable for the missing data and derive the EM algorithm for iteratively estimating θ.

Problem 10. Distribution Theory and Inference (STA5326)

A small nation has 2,000,000 inhabitants. The people live in 1,00**1** communities. 1,000 of these communities are small villages, each with a population of 1,000 people. The last community is a single large city with a population of 1,000,000. In each village 90% of the residents are farmers. In the city only 5% of the residents are farmers (who must walk to the countryside to till their fields).

Consider two different methods of sampling a person from this nation. (In the following, when we say that something is done "at random", we mean that all the possibilities are equally likely.)

Method #1: Choose a person at random from census records which list all 2,000,000 people.

Method #2: Choose a community at random, and then choose a person at random from this community.

Answer the following.

(a) Suppose a person is sampled using Method #1. Given that this person is **not** a farmer, what is the probability he/she lives in the city?

(b) Answer the question in (a) if the sampling is done using Method #2.

(c) Suppose we sample people one by one by repeated independent use of Method #1, and continue until we have sampled exactly 2 farmers. Let X be the total number of people sampled in this process. Find $P(X > k)$ for $k \geq 2$.

(d) Answer the question in (c) if the sampling is done by repeated independent use of Method #2.

[Note: In parts (c) and (d) it is possible to choose the same person more than once. In part (d) it is possible to choose the same community more than once.]

Problem 11. Distribution Theory and Inference (STA5326)

Suppose that (X, Y) has the joint density

$$f(x,y) = \begin{cases} \dfrac{\sqrt{3}}{\pi} \exp\left\{-(x^2 - xy + y^2)\right\} & x > 0, \\ 0 & x \le 0 \end{cases}$$

Find the following by direct calculation.

(a) The marginal density of X.

(b) The conditional density of Y given $X = x$.

(c) The density of $Z = \dfrac{Y}{X}$.

Problem 12. Statistical Inference (STA5327)

Suppose we observe data $X_1, \ldots, X_n$ which is a random sample from the density

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

(a) Find the MOM (method of moments) estimator of θ. [Your answer may be written in terms of Γ^{-1}, the inverse of the gamma function defined by $\Gamma(a) \equiv \int_0^\infty x^{a-1} e^{-x}\, dx$.]

(b) Find an ancillary statistic. [Hint: Consider the density of $\log(X_i)$.]

(c) Give a detailed statement of the most powerful test of level α for

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta = 2\,.$$

[Do not try to find the critical value explicitly, but state the condition it must satisfy.]

Problem 13. Statistical Inference (STA5327)

Suppose we observe $X_1, \ldots, X_n$ i.i.d. from the density

$$f(x|\alpha) = \frac{c(\alpha)x^\alpha}{1+x^2}\,, \quad 0 < x < \infty\,,\ -1 < \alpha < 1$$
$$\text{where} \quad c(\alpha) = \frac{2}{\pi}\cos\left(\frac{\pi\alpha}{2}\right).$$

Answer the following. Justify your answers.

(a) Find a complete sufficient statistic for α.

(b) Find the MLE for α.

(c) Is there a function of α, say $g(\alpha)$, for which there exists an unbiased estimator whose variance attains the Cramér-Rao Lower Bound? If so, find it. If not, show why not.

(d) Let T denote the complete sufficient statistic found in part (a). Is there a function of α, say $h(\alpha)$, for which there are two different unbiased estimators which are functions of T? Answer 'Yes' or 'No' and prove your answer.