Ph. D. Qualifying Exam
Thursday, August 25, 2005

Please submit solutions to at most **seven** problems. You have four hours. No one is expected to answer all the problems correctly. Partial credit will be given. All problems are worth an equal amount of credit.

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

Applied Statistics

Problem 1. Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i\,, \quad i = 1, \ldots, n,$$

where $\{e_i, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma^2)$ variables. Let $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$ denote the least-squares estimates of the unknown parameters $\{\beta_0, \beta_1, \sigma^2\}$

(a) What are the distributions of the least-squares estimates $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$?

(b) In the model, suppose the predictor values are replaced by $\{cx_i, \ i = 1, \ldots, n\}$ where c is a non-zero constant. How are $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, R^2$ and the t-test of "$H_0 : \beta_1 = 0$" affected by the change?

(c) Suppose the response values are replaced by $\{ay_i, \ i = 1, \ldots, n\}$ where a is a non-zero constant. How are $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, R^2$ and the t-test of "$H_0 : \beta_1 = 0$" affected by the change?

Problem 2. In a generalized linear model, the response variable Y_i is assumed to have a density function with the form:

$$f(y_i; \theta_i, \phi) \;=\; \exp\{[y_i\theta_i - b(\theta_i)]/a_i(\phi) + c(y_i, \phi)\}, \quad i = 1, \ldots, n.$$

Denote $\mu_i = E(Y_i)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, and $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Let $l(\boldsymbol{\mu}; \boldsymbol{y})$ be the log-likelihood function. For a given model with p unknown parameters, let $l(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ denote the maximum of the log-likelihood function.

(a) Suppose that $\{Y_i, i = 1, \ldots, n\}$ are independent Poisson random variables. Show that $l(\boldsymbol{\mu}; \boldsymbol{y})$ reaches its maximum when $\mu_i = y_i$. Letting $a_i(\phi) = \phi/w_i$, write out the scaled deviance $2[l(\boldsymbol{y}; \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}; \boldsymbol{y})]$ in this case.

(b) Suppose that $\{n_i Y_i, i = 1, \ldots, n\}$ are independent with $n_i Y_i \sim$ Binomial(n_i, π_i). Repeat the questions in Part (a).

Problem 3. Consider the following linear model for a one-way layout design:

$$Y_{ij} \quad = \quad \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, k; \;\; j = 1, \ldots, n_i \,,$$

where for each i, the random errors $\{\epsilon_{ij}, j = 1, \ldots, n_i\}$ are assumed to be i.i.d. $N(0, \sigma_i^2)$. The sample variances for the k treatments are denoted by $s_i^2, \;\; i = 1, \ldots, k$.

(a) It is well known that

$$(n_i - 1)s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \sim \sigma_i^2 \chi^2_{n_i - 1}.$$

Based on this result, show that the distribution of $\log(s_i^2)$ can be approximated by $N(\log(\sigma_i^2), 2(n_i - 1)^{-1})$.

(b) In a one-way layout design with $k = 10$ and $n_1 = \cdots = n_{10} = 6$, the s_i^2's and $\log(s_i^2)$'s are given in the following table:

s_i^2	0.56	0.82	0.45	1.02	0.34	0.77	0.52	0.88	0.31	1.20
$\log(s_i^2)$	−0.58	−0.20	−0.80	0.02	−1.08	−0.26	−0.65	−0.13	−1.17	0.18

Based on the result in part (a), perform a test on "$H_0 : \;\; \sigma_1^2 = \cdots = \sigma_{10}^2$" against the alternative hypothesis "$H_a : \quad$ some of the variances are not equal."

Probability

Problem 4.

(a) Suppose that P, Q are two probability measures on the same measurable space $(\Omega, \mathcal{A})$. Suppose that P and Q are both absolutely continuous with respect to the measure μ with densities (Radon-Nikodym derivatives) p and q, respectively. Thus $P(A) = \int_A p \, d\mu$ and $Q(A) = \int_A q \, d\mu$, for $A \in \mathcal{A}$. Show that

$$\sup_{A \in \mathcal{A}} |P(A) - Q(A)| = \frac{1}{2} \int |p - q| d\mu.$$

(b) Suppose that $f_0, f_1, \ldots$ are ≥ 0, defined on a sigma-finite measure space $(\Omega, \mathcal{A}, \mu)$, and satisfy $\int_\Omega f_n \, d\mu = 1$, for all $n = 0, 1, 2, \ldots$. Suppose also that $f_n \to_{\text{a.e.}} f_0$ with respect to μ. Use (a) above to show that

$$\sup_{A \in \mathcal{A}} \left| \int_A f_n \, d\mu - \int_A f_0 \, d\mu \right| \to 0, \quad \text{as} \;\; n \to \infty.$$

Problem 5.

(a) Assume that $X_1, \ldots, X_n, \ldots$ are uncorrelated and that $EX_j^2 \le M < \infty$ for all $j \ge 1$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Show that

$$\bar{X}_n - E\bar{X}_n \to_{\mathrm{L}_2} 0, \quad \text{as} \quad n \to \infty.$$

(b) Assume now that $X_1, \ldots, X_n, \ldots$ are i.i.d. with $E|X_1| \le M < \infty$ and $EX_1 = m$. Use Vitali's theorem to show that

$$\bar{X}_n \to_{\mathrm{L}_1} m, \quad \text{as} \quad n \to \infty.$$

Problem 6. The median of a distribution F on the real line is defined as the smallest value m for which $F(m) \ge 1/2$.

(a) Assume $F(t) > 1/2$ for each $t > m$. Show that for every $\varepsilon > 0$ there is a $\delta > 0$ such that $|m(F) - m(G)| \le \varepsilon$ whenever $\sup_t |F(t) - G(t)| \le \delta$.
[*Hint:* Choose $\delta > 0$ such that $F(m-\varepsilon) < \frac{1}{2} - \delta$ and $F(m+\varepsilon) > \frac{1}{2} + \delta$.]

(b) Assume $F(t) > 1/2$ for each $t > m$. Let $X_1, \ldots, X_n$ be an independent sample from F. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^n I_{\{X_i \le x\}}$ be the empirical distribution function. Using part (a), show that the sample median $m(F_n)$ is strongly consistent.

(c) Assume that $f(x) = F'(x)$ exists and is continuous and positive at m. Show that the standardized sample median $\sqrt{n}(m(F_n) - m)$ is asymptotically normal and compute the asymptotic variance.

Problem 7. Consider a Markov chain with 9 states arranged in a 3 by 3 rectangular lattice. If the chain is in a particular state at time n, then at time $n+1$ it moves to one of the neighboring states to the north, south, east, or west, choosing among them with equal probability. (The center state has 4 neighbors, edge states have 3 neighbors, and corner states have only 2 neighbors.) Answer the following. Clearly state any results you use in your answer.

(a) If the chain starts from the southwest corner, what is the expected length of time until it returns to this state?

(b) If the chain starts from the southwest corner, what is the expected number of times it will visit the center state before first visiting the northeast corner? (Your answer may be given as a number or as a matrix expression. All matrices which appear in your answer must be given explicitly.)

Problem 8. Consider the linear model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + e_i$, where e_i are i.i.d. $N(0, \sigma^2)$. Let $\{\hat{Y}_i\}$ denote the fitted values obtained by the least squares method.

(a) Show that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

(b) In particular, let $p = 2, n = 4, X_{i1} = 1, X_{i2} = 0$ for $i = 1, 2$, and $X_{i1} = 0, X_{i2} = 1$ for $i = 3, 4$. Is β_2 estimable? (Justify your answer.)

(c) Compute $E\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)Y_i\right)$

Problem 9. Let X be a random variable whose distribution depends on a parameter θ. Consider the three different families of distributions for X described below. The first two are families of **mass** functions in which $X \in \{0, 1, 2\}$ and $0 < \theta < \infty$. The third is a **finite** family of **density** functions in which $0 < X < 1$ and $\theta \in \{0, 1, 2\}$. In each case determine whether the family of distributions of X is complete. In other words, in each case determine whether X is a complete statistic.

	$P(X=0)$	$P(X=1)$	$P(X=2)$	
Family #1	$c\,\theta$	$c\,2\theta$	$c\,(1+\theta)$	for $\theta > 0$
Family #2	$c\,\theta$	$c\,\theta^2$	$c\,\theta^3$	for $\theta > 0$
Family #3	$f(x \mid \theta) = c\,x^\theta$ for $0 < x < 1$ and $\theta \in \{0, 1, 2\}$.			

In the above table, c is a normalizing constant whose value depends on θ.

Problem 10. Let $X_1, X_2 \ldots, X_n$ be a random sample from a $N(\theta, \sigma^2)$ distribution, and suppose that the prior distribution on θ is $N(\mu, \tau^2)$. Find the posterior distribution of θ and give the mean and variance of this distribution.

[Hint: You may find it useful to use the following fact: If $f(y)$ is a probability density function satisfying

$$f(y) \propto e^{-\frac{1}{2}(ay^2 - 2by)} \quad \text{for } -\infty < y < \infty,$$

then f is the density of a normal distribution with mean b/a and variance $1/a$.]

Problem 11. Assume that we are interested in drawing a random sample of size n to examine a random variable X. We wish to test whether the mean in the population is equal to some specified value, μ_0, versus that it is less than this value, that is:

$H_0 : \mu = \mu_0$ vs $H_a : \mu < \mu_0$

For our test, we want the values of the Type I and Type II errors to be .05 and .10 respectively, that is, $\alpha = 0.05$ and $\beta = 0.10$.

You may assume that X is normally distributed with variance σ^2, and that σ^2 is known.

What is the minimum size of n to assure the correct levels of α and β? That is, derive the formula for computing the sample size, based on the probability statements involved and the assumption of normality.

Problem 12. If Y is a dichotomous random variable, and X is a continuous random variable, then the logistic model specifies that:

$$\Pr(Y = 1) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

(a) Sketch out a derivation of how, given a sample of size n in which we have measured Y and X, we would establish estimates of α and β.

(b) State two different ways we might test the hypothesis: $H_0 : \beta = 0$

Problem 13. Consider the results of a diagnostic test, Y, used to classify disease status ($Y = 0$ signifies that on the basis of the test we call people non-diseased, and $Y = 1$ signifies that on the basis of the test results we call people diseased). We denote disease state by D ($D = 0$ signifies non-diseased and $D = 1$ signifies diseased).

(a) Provide a definition of each of the following in terms of probabilities

1. TPF=True Positive Fraction
2. FPF=False Positive Fraction
3. PPV=Positive Predictive Value
4. NPV=Negative Predictive Value

(b) If we also denote ρ=the prevalence of disease in the population and $\tau = \Pr(Y = 1)$, then we can examine misclassification problems in terms of (TPF, FPF, ρ) or in terms of (PPV, NPV, τ). Are these two parameterizations equivalent? If they are not, provide a counter example. If they are, provide at least a sketch of a proof.