Ph. D. Qualifying Exam
Monday, January 5, 2009

Please submit solutions to at most **eight** problems. You have five 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. (5166) Consider the following linear model involving two factors A and B in a split-plot experiment:

$$Y_{ijk} = \mu + \alpha_i + e_{ij} + \beta_j + \epsilon_{ijk}, \qquad i = 1, \ldots, a; \quad j = 1, \ldots, b; \quad k = 1, \ldots, r,$$

where μ, $\{\alpha_i, i = 1, \ldots, a\}$, and $\{\beta_j, j = 1, \ldots, b\}$ are unknown constants. The random errors $\{e_{ij}\}$ are iid $N(0, \sigma_e^2)$, $\{\epsilon_{ijk}\}$ are iid $N(0, \sigma_\epsilon^2)$, and the two groups $\{e_{ij}\}$ and $\{\epsilon_{ijk}\}$ are independent. Define

$$Y = (Y_{111}, Y_{112}, \ldots, Y_{11r}, Y_{121}, Y_{122}, \ldots, Y_{12r}, \ldots, Y_{ab1}, Y_{ab2}, \ldots, Y_{abr})'.$$

(a) Find the expectation $\boldsymbol{\mu} = E(Y)$ and the variance-covariance matrix $V = \mathrm{Cov}(Y)$. What is the distribution of Y?

(b) Write an ANOVA table for this experiment, including the Sources of Variation, Degrees of Freedom (df), Sum of Squares, Mean Squares. Derive the expectations for the mean squares.

(c) Construct a test statistic for testing $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$. Give your justification for the test statistic.

Problem 2. (5167) Consider the linear regression model:

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

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 moving-average model:

$$\xi_i = \epsilon_i + \theta_1 \epsilon_{i-1} + \theta_2 \epsilon_{i-2},$$

where $\{\epsilon_i, i = -1, 0, 1, \ldots, n\}$ are iid $N(0, \sigma^2)$ variables. Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$. Define $\hat{Y} = X\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\xi}} = Y - \hat{Y}$.

(a) What are the means and covariance matrices of $\hat{Y}$ and $\hat{\boldsymbol{\xi}}$? What are the distributions of $\hat{Y}$ and $\hat{\boldsymbol{\xi}}$?

(b) Are $\hat{Y}$ and $\hat{\boldsymbol{\xi}}$ independent? Show your reasons.

(c) Can you find an estimate for $\boldsymbol{\beta}$, say $\tilde{\boldsymbol{\beta}}$, such that $\tilde{Y} = X\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\xi}} = Y - \tilde{Y}$ are independent?

Problem 3. (5168) In a generalized linear model (GLM), the response variable Y is assumed to have a density function with the exponential dispersion form:

$$f(y;\theta,\phi) \;=\; \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}.$$

(a) Suppose a random variable $X \sim \text{Binomial}(m,\pi)$ and define $Y = X/m$. Express the binomial mass function in the exponential form in terms of the canonical parameter $\theta = \text{logit}(\pi)$. Derive the deviance $D(y,\pi)$ for the binomial model.

(b) Suppose that $\{X_i \sim \text{Binomial}(m_i,\pi_i), i = 1,\ldots,n\}$ are independent random variables and the probabilities $\boldsymbol{\pi} = (\pi_1,\pi_2,\ldots,\pi_n)'$ follow a linear logistic model. Let $H_0 \subset H_1$ be nested hypotheses (fewer covariates under H_0). Show that the corresponding model deviances satisfy the Pythagorean relationship

$$D(\boldsymbol{y},\hat{\boldsymbol{\pi}}^{(0)}) = D(\boldsymbol{y},\hat{\boldsymbol{\pi}}^{(1)}) + D(\hat{\boldsymbol{\pi}}^{(1)},\hat{\boldsymbol{\pi}}^{(0)}),$$

where $\boldsymbol{y} = (y_1,\ldots,y_n)'$ is the vector of observed proportions, and $\hat{\boldsymbol{\pi}}^{(0)}$ and $\hat{\boldsymbol{\pi}}^{(1)}$ are the vectors of fitted proportions under H_0 and H_1, respectively.

Problem 4. (5507) Suppose X and Y are random samples of size m and n, respectively, from a continuous distribution. Assume that the X's and Y's are independent. Let W be the Wilcoxon rank-sum statistic for the sample Y, i.e., W is the sum of the ranks of the sample Y with respect to the combined sample consisting of X and Y.

(a) Find the distribution of W when $m = 3$ and $n = 2$.

(b) For general integers $n \geq 1$ and $m \geq 1$, find the range of W and evaluate $P(W = n(n+1)/2)$.

Problem 5. (5707) Suppose $\sigma^2 > 0$. Consider the matrix

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

(a) For what values of ρ_1, ρ_2 is Σ a covariance matrix?

(b) Suppose that Σ is a covariance matrix and that ρ_1, ρ_2 are positive. Find the proportion of the total variance explained by the first principal component.

Probability

Problem 6. (5446) Let $X_1, X_2, \ldots$ be independent random variables with common continuous distribution function $F(x)$. Let

$$R_n = \sum_{j=1}^{n} 1_{\{X_j \geq X_n\}}$$

be the (relative) rank of X_n among $X_1, \ldots, X_n$. Here 1_A is the indicator function of the set A. For example, $R_n = 1$ means that X_n is the largest of $X_1, \ldots, X_n$ (we say X_n is a record) and $R_n = 2$ means that X_n is the second largest of $X_1, \ldots, X_n$.

(a) Show that $P(X_i = X_j) = 0$ for $i \neq j$.

(b) Conclude that $P\left(\bigcup_{i \neq j} \{X_i = X_j\}\right) = 0$.

(c) Using the fact that all possible $n!$ orderings of $X_1, \ldots, X_n$ have the same probability, show that

$$P(R_n = k) = \frac{1}{n}, \quad \text{for } k = 1, \ldots, n.$$

(d) Show that the random variables $R_1, \ldots, R_n$ are independent.

Problem 7. (5447) (This exercise uses the notation and situation of the previous exercise, but does **not** require you to have solved that exercise.) Define

$$Y_k = 1_{\{R_k = 1\}} = \begin{cases} 1 & \text{if } R_k = 1 \text{ } (X_k \text{ is a record}) \\ 0 & \text{otherwise} \end{cases}$$

According to the results of the previous exercise $Y_1, Y_2, Y_3, \ldots$ are independent with $P(Y_k = 1) = \dfrac{1}{k}$. Using these facts, do the following:

(a) Show that

$$\sum_{n=2}^{\infty} \mathrm{Var}\left(\frac{Y_n}{\log(n)}\right) < \infty.$$

(b) One can show (and you may use without proof) that

$$\frac{1}{\log(n)} \sum_{k=1}^{n} E(Y_k) \to 1 \quad \text{as } n \to \infty.$$

Prove that

$$\frac{1}{\log(n)} \sum_{k=1}^{n} Y_k \to 1$$

almost surely, as $n \to \infty$. (Here $\sum_{k \le n} Y_k$ is the number of records in the first n observations.) Carefully state which results you use.

Problem 8. (5334) Let the population proportions p_1, p_2, p_3 of three genotypes of a single gene with two alleles be given by the Hardy-Weinberg formula

$$p_1 = (1-\theta)^2\,, \quad p_2 = 2\theta(1-\theta)\,, \quad p_3 = \theta^2\,, \quad 0 < \theta < 1.$$

(a) Find the MLE of θ based on a large sample of size n in which the observed frequencies of the three genotypes are n_1, n_2, n_3.

(b) Construct an asymptotically optimal confidence interval for θ having asymptotic confidence coefficient $1 - \alpha$.

Problem 9. (5326) Suppose that tosses of a coin are independent with probability p of heads. Two people each toss this coin n times.

(a) What is the probability they will toss the same number of heads? Give an **exact** formula. (The formula may be left as a summation.)

(b) Give a normal approximation for the probability of tossing the same number of heads. (Assume here that n is large.) Evaluate this normal approximation explicitly for $n = 100{,}000$ and $p = 0.25$.

Problem 10. (5807) At a certain location, cars pass by according to a Poisson process with rate α cars per minute, and bicycles pass by according to a Poisson process with rate β bicycles per minute. The cars and bicycles are independent. Both cars and bicycles are considered to be "vehicles". Answer the following. (**No** proofs are required.)

(a) Standing at this location, what is the probability that exactly k vehicles will pass by in the next minute?

(b) What is the probability that the first vehicle to pass you will be a bicycle?

(c) What is the distribution of the length of time you must wait for the first vehicle to pass? (State the density.)

(d) Let $W(t)$ be the total number of **wheels** which have passed you by time t. (Assume that all cars have 4 wheels, and all bicycles have 2 wheels.) What kind of process is $W(\cdot)$? (Be as detailed as possible). What are the mean and variance of $W(t)$?

Theoretical Statistics

Problem 11. (5327) Let $X_1, \ldots, X_n$ be iid with the geometric distribution

$$P_\theta(X = x) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots, \quad 0 < \theta < 1.$$

(a) Show that $T = \sum X_i$ is a complete and sufficient statistic for θ.

(b) Show that the level α likelihood ratio test (LRT) of

$$H_0 : \theta = 0.5 \quad \text{versus} \quad H_1 : \theta \neq 0.5$$

rejects when $T \leq c_1$ or $T \geq c_2$ for some choice of constants c_1 and c_2.

Problem 12. (5208) Answer the following. Parts (a) and (b) are unrelated.

(a) Assume $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 25 \end{pmatrix}$.

1. What is the distribution of $\boldsymbol{W} = \begin{pmatrix} \frac{Y_1-3}{2} \\ \frac{Y_2-2}{3} \\ \frac{Y_3-1}{5} \end{pmatrix}$?
2. What is the distribution of $\boldsymbol{W}'\boldsymbol{W}$?

(b) Consider n independent Bernoulli random variables, $Y_1, \ldots, Y_n$ with $\Pr(Y_i = 1) = \pi_i$ and $\Pr(Y_i = 0) = 1 - \pi_i$. The joint distribution in this case may be written:

$$\prod \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}.$$

1. Show that this is a member of the exponential family.

2. If the link is $g(\pi) = \log(\frac{\pi}{1-\pi}) = \boldsymbol{x'\beta}$, show that this is equivalent to

$$\pi = \frac{\exp(\boldsymbol{x'\beta})}{1 + \exp(\boldsymbol{x'\beta})}\,.$$

Biostatistics

Problem 13. (5172) Answer the following. Parts (a) and (b) are unrelated.

(a) The results of a **cohort** study examining the relationship between gender and coronary heart disease (CHD) are summarized in the following table. In this study, 4,541 individuals were followed 10 years and it was determined which of these individuals actually experienced CHD. For example, 236 out of 2,009 males developed the disease, 1,773 males did not develop the disease, etc.

		CHD		
		Yes	No	Total
Male	Yes	236	1,773	2,009
	No	142	2,390	2,532
	Total	378	4,163	4,541

Provide definitions and estimates for each of the following statistics.

1. The relative risk of CHD for males.
2. the odds ratio of CHD for males.
3. Recall that there are several types (definitions) of attributable risk. Define any one and provide the correct estimator.

(b) The following table provides the results of submitting urine samples with known status as to amphetamine use to a commercial laboratory.

		True Status		
		Present	Absent	Total
Lab Determination	Present	180	20	200
	Absent	400	600	1,000
	Total	580	620	1,200

1. What is the sensitivity of this test?
2. What is the specificity of this test?

3. Suppose at a particular company, .1% (0.001) of employees at the company actually use amphetamines and the company orders a drug test on a random employee. If the test is positive, what is the probability that the employee actually uses drugs.

Problem 14. (5244) Answer the following.

(a) Explain the "phases" of clinical trials, that is, their names and what the main question being addressed is.

(b) Explain what a "run-in" period is and what it is meant to accomplish.

(c) Explain what a "non-inferiority" trial is and how its design differs from a clinical trial that is attempting to demonstrate treatment efficacy.

(d) An investigator wants to conduct a study examining whether the correlation between two laboratory measures differs in the two treatment arms. The primary outcome will involve testing the difference between two normal means.

She is willing to assume the variance is equal for the two groups and is σ^2.

She specifies a minimum difference that she thinks is clinically important: $\Delta = \mu_1 - \mu_0$

She specifies the level of type I error she is willing to accept, α.

Finally, she specifies the level of the type II error she is willing to accept (or alternatively the power she wishes her study to have).

1. After you have done the sample size calculation based on the data she provides, she says that after thinking about it, the minimum difference that she thinks is clinically important is really:

$$\Delta_1 = .5 * \Delta$$

   What effect does this have on the estimated sample size?

2. She later comes backs with yet another value:

$$\Delta_2 = 2 * \Delta$$

   What effect does this have on the original sample size calculation?

3. Let n_1 and n_0 denote the number of patients in the treatment and control groups, respectively, and let $n = n_1 + n_0$. Because of cost and practical considerations she can randomize only a fixed total number n of participants. Derive the values of n_1 and n_0 that maximize the power of the study. (Assume fixed costs, i.e., the costs for all patients is the same regardless of treatment group.)

Problem 15. (5179) Let the survival time $T > 0$ be an integer-valued (discrete) random variable with finite mean residual lifetime $\mu_k = E[T-k \,|\, T > k]$ and discrete hazard rate $h_k = P[T = k \,|\, T \geq k]$ for $k = 0, 1, 2, \ldots$

(a) Show that $\mu_0 - (1 - h_1)\mu_1 = 1$.

Hint: One approach is to write down the terms in the sums for μ_0 and μ_1 in two rows, multiply each term of μ_1 by $(1 - h_1)$, and subtract the rows.

(b) Can you generalize the result of (a) to $\mu_k - (1 - h_{k+1})\mu_{k+1}$?

(c) Use the results of (a) and (b) to show that the Geometric distribution is the only discrete distribution with constant mean residual lifetime.

(d) Show that the exponential distribution is the only continuous distribution for which the mean residual lifetime $r(t)$ is constant for all $t > 0$.

Computational Statistics

Problem 16. (5106) Let $V = (V_1, V_2)$ where V_1, V_2 are iid uniform random variables in the interval $(-1, 1)$. Our goal is to develop a Monte Carlo technique to estimate the quantity

$$\theta = P(V \in C) = E[I]$$

where C is the unit disk in $\mathbb{R}^2$ and I is the indicator variable for V being inside the disk.

(a) Define a new random variable Z such that $Z = E[I|Y]$ for some random variable Y. Show that by choosing an appropriate Y we can get an explicit expression for Z.

(b) Show that by averaging independent samples of Z, we can estimate θ and this estimator has a smaller variance than the one that averages independent samples of I.