Ph. D. Qualifying Exam
Thursday, August 23, 2007

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 following linear model for a randomized block design:

$$y_{ti} = \mu + \beta_i + \tau_t + \epsilon_{ti}, t = 1, \ldots, k; i = 1, \ldots, n,$$

where μ is an overall mean, τ_t is the effect of tth treatment, β_i is the effect of ith block, $\{\epsilon_{ti} : t = 1, \ldots, k; i = 1, \ldots, n\}$ are assumed to be iid $N(0, \sigma^2)$.

(a) The least squares estimate of τ_t is $\hat{\tau}_t = \bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot}$. Find the expectation and variance of $\hat{\tau}_t$.

(b) Show the decomposition of variation for the experiment: $S_D = S_B + S_T + S_R$ where

- S_D: Total Variation of the observations,
- S_B: Sum of Squares for Blocks,
- S_T: Sum of Squares for Treatments,
- S_R: Sum of Squares for Experimental Errors.

(c) Find the expectation of S_T and S_R.

Problem 2. Consider an $I \times J$ contingency table that cross-classifies a multinomial sample of n subjects on two categorical responses. Let X denote the row variable and Y the column variable. The cell probabilities are $\{\pi_{ij}, i = 1, \cdots, I; j = 1, \cdots, J\}$ and the expected frequencies are $\{\mu_{ij} = n\pi_{ij}, i = 1, \cdots, I; j = 1, \cdots, J\}$.

(a) Define $\theta_{ij} = \log(\pi_{ij}/\pi_{1,1})$ for $(i, j) \neq (1, 1)$. Show that the multinomial distribution with cell probabilities $\{\pi_{ij}, i = 1, \cdots, I; j = 1, \cdots, J\}$ belongs to a multiple parameter exponential family with $\{\theta_{ij}\}$ as the natural parameters. Find the maximum likelihood estimates for the cell probabilities $\{\pi_{ij}, i = 1, \cdots, I; j = 1, \cdots, J\}$.

(b) Show that the independence of X and Y, i.e., $\pi_{ij} = \pi_{i+}\pi_{+j}$, is equivalent to the expected frequencies $\{\mu_{ij} = n\pi_{ij}, i = 1, \cdots, I; j = 1, \cdots, J\}$ following the loglinear model:

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y.$$

Problem 3. Suppose that $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a sample from a bivariate normal distribution, i.e.,

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), \; i = 1, \cdots, n.$$

(a) Show that the conditional distribution of y_i given x_i is normal and

$$y_i | x_i \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_i - \mu_1), \; \sigma_2^2(1 - \rho^2)\right), \; i = 1, \cdots, n.$$

(b) Define

$$\beta_1 = \rho\frac{\sigma_2}{\sigma_1}, \quad \beta_0 = \mu_2 - \beta_1\mu_1, \quad \sigma^2 = \sigma_2^2(1 - \rho^2). \tag{1}$$

Then y_i given x_i follows the simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, \cdots, n.$$

where $\{\epsilon_i, \; i = 1, \cdots, n\}$ are iid N(0, σ^2). The moment estimates of β_0, β_1, and σ^2, denoted by $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\sigma}^2$, respectively, are obtained by simply substituting the sample means ($\bar{x}$ and $\bar{y}$), sample variances ($\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$), and sample correlation $\hat{\rho}$ into (1). Are the moment estimates the same as the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$?

Probability

Problem 4. Let $\{X_n\}_{n \geq 1}$ be a sequence of discrete random variables such that $P(X_n = 0) = 1 - \frac{1}{n^2}$ and $P(X_n = n^2) = \frac{1}{n^2}$, for each n.

(a) Show that $EX_n \to 1$, as $n \to \infty$.

(b) Show that $X_n \to 0$ a.s. as $n \to \infty$.

(c) State Vitali's Theorem. Does the sequence X_n converge in L_1 to zero?

Problem 5. Show that if X is a Poisson random variable with parameter λ then

$$\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1), \quad \text{as } \lambda \to \infty.$$

Hint. You can use any method of proof. However, you may find it easier to use characteristic functions. Recall that the characteristic function of a Poisson (λ) random variable is $\exp\left(\lambda(e^{it}-1)\right)$ and that of a $N(0,1)$ random variable is $\exp(-t^2/2)$.

Problem 6. Let $p_\theta(x)$ be a *Hellinger* differentiable density in $\mathbb{R}$ with score function $\tau_\theta(x)$.

(a) Show that $\int \tau_\theta(x) p_\theta(x)\, dx = 0$.

(b) Show that the product density $p_\theta(x,y) := p_\theta(x) \times p_\theta(y)$ in $\mathbb{R}^2$ is Hellinger differentiable with score function $\tau_\theta(x,y) = \tau_\theta(x) + \tau_\theta(y)$.

Problem 7. Consider a branching process in which each individual lives exactly one unit of time and then gives birth to either 0, 1, or 2 offspring with probabilities 1/3, 1/6, and 1/2, respectively.

(a) If you start with exactly **5** organisms at time zero, what is the probability this population will eventually become extinct?

(b) What are the mean and variance of the total number of organisms born in the third generation (that is, at time 3)?

Theoretical Statistics

Problem 8. Suppose $X_1, \ldots, X_n$ are iid with pdf

$$f(x \mid \alpha, \beta) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha} \quad \text{for } 0 \le x \le \beta$$

where α and β are positive.

(a) Find a two-dimensional sufficient statistic for (α, β).

(b) Find the MLEs of α and β.

(c) Find an ancillary statistic.

Problem 9. Let $X_1, \dots, X_n$ be iid $N(\mu, \sigma^2)$.

(a) Find the best unbiased estimator of μ^2.

(b) Find the best unbiased estimator of σ^4.

Hint: You may find it useful to know that the χ_k^2 distribution has mean k and variance $2k$.

Problem 10. Consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + e_i \,,$$

where the predictor variables take on the following values

i	1	2	3	4	5	6	7
x_{i1}	1	1	-1	-1	0	0	0
x_{i2}	1	-1	1	-1	0	0	0

and the errors e_i are iid $N(0, \sigma^2)$.

(a) Show that β_1 and $\beta_{11} + \beta_{22}$ are estimable.

(b) Find (nonmatrix) algebraic expressions for the best linear unbiased estimators (BLUEs) of β_1 and $\beta_{11} + \beta_{22}$.

(c) Find an algebraic expression for the MSE for this model.

Biostatistics

Problem 11. For a single covariate x, the proportional hazards model specifies that:

$$\lambda(t|x) = \lambda_0(t) \exp(\beta x)$$

(a) Consider the following observations:

time	x	fail
2	2	1
4	1	0
6	3	1
10	5	0

Write out the partial likelihood associated with the data.

(b) The above contained no ties. Suppose the data were the following:

time	x	fail
4	2	1
4	1	1
6	3	0
10	5	1

Sketch at least two ways we could take the tie into account in forming the partial likelihood contribution at time 4.

Problem 12. The following table presents the results of four logistic models relating the specified covariates to death from coronary heart disease (CHD). Age is age of the participant in years and sbp is the participant's systolic blood pressure in mmHg.

Covariate	**Null Model**		**Univariate**		**Univariate**		**Bivariate**	
	β	$se(\beta)$	β	$se(\beta)$	β	$se(\beta)$	β	$se(\beta)$
age			0.038	0.006			0.029	0.007
sbp					0.019	0.003	0.016	0.003
α	-1.237	0.053	-3.180	0.328	-3.796	0.350	-4.895	0.433
log likelihood	-1071.1		-1052.3		-1042.9		-1032.9	

(a) How do you decide which of these four models best describes the data?

(b) For the covariates age and sbp, provide an interpretation of the estimated coefficients in the bivariate model.

(c) For the covariates age and sbp in the bivariate model, describe two ways to test whether they are significantly different from zero.

(d) In the **Null Model**, what does:

$$\frac{\exp(\alpha)}{1+\exp(\alpha)}$$

estimate?

(e) In the data from which these coefficients were estimated, the average sbp for those 50 years of age or less is 128.8 mmHg, among those greater than 50 years of age, the average sbp is 137.1 mmHg. Similarly, among those with sbp ≤ 130 mmHg the average age is 48.4 years, among those with sbp > 130 mmHg the average age is 51.3 years.

Provide a possible explanation for the change in the coefficients in the bivariate model as compared to the univariate models.

Problem 13.

An investigator wants to do a clinical trial. The primary outcome will involve testing the difference between two normal means.

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

She specifies a minimum difference that she thinks is clinically important: $\Delta = \mu_1 - \mu_2$

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).

(a) 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?

(b) She later comes back with yet another value:

$$\Delta_2 = 2 * \Delta$$

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

(c) Because of cost and practical considerations she can randomize n participants.

n_1 is the number of patients in the treatment group.

n_0 is the number of patients in the control group.

Then, $n = n_1 + n_0$ is a fixed constant.

She asks you what values of n_1 and n_2 provide the maximum power.

Assuming fixed costs, i.e., the costs for all patients is the same regardless of treatment group, derive the values of n_1 and n_0 that maximize the power of the study.

Problem 14. For a given probability density function $f(x)$, we are interested in estimating its right tail:

$$\theta = \int_a^\infty f(x)dx, \quad a \gg \int_{-\infty}^\infty x f(x)dx \ .$$

(a) Show how you can use importance sampling to estimate θ, when the samples are obtained from the tilted density:

$$f_t(x) = \frac{e^{tx} f(x)}{M(t)}, \quad M(t) = \int_{-\infty}^\infty e^{tx} f(x)dx, \quad t > 0 \ .$$

(b) If $f(x)$ is the standard normal density, what form does the tilted density take?

(c) Write an algorithm to estimate θ for the standard normal density, with $a = 3$. What is the optimal choice of t for this problem?

Problem 15. Let X_t be a discrete-time, continuous-state Markov chain with the transition kernel $\Pi(x, y)$. For a probability density function f:

(a) State the detailed balance condition involving Π and f.

(b) Show that if f satisfies the detailed balance condition, then f is a stationary probability density associated with the Markov chain.

(c) Assuming that it is computationally feasible to perform transitions under Π, suggest a Monte Carlo Markov Chain (MCMC) algorithm to estimate the mean of f. You can also assume that the variance under f is finite.