PH.D QUALIFYING YEAR AND ADVISORY EXAM MATERIALS

Statistics Centre

April 3, 2008

Chapter 1

Advisory Exams in Statistics

All students entering the Ph.D. (Statistics) program are required to take an Advisory Exam. This should be taken within a week of the start of the first term of classes. It is important for the student to bear in mind that this exam cannot be 'failed'. Rather, the purpose of the exam is to enable us to determine a suitable program of courses for the student's first year, and to take any necessary action to remedy deficiencies in the student's background. After two full terms of course work the Graduate Committee of the department will meet with the relevant instructors, and then make a recommendation as to whether the student should continue in the program.

The Statistics exam is an 'open book' exam, and is in two parts:

STAT I This 3-hour exam is based (loosely) on our more theoretical senior undergraduate courses STAT 366 and STAT 472, and on STAT 312. The STAT 366 and STAT 472 components include the following topics: principles of inference, sufficiency, likelihood, completeness, optimal methods of estimation, testing and interval construction, distribution theory; combinatorial probability, conditioning, laws of large numbers, central limit theory, generating functions, Markov chains, renewal processes, martingales. The STAT 312 component includes: vector spaces, eigenvalues, matrix factorisations, theory of continuity, differentiation and integration of real-valued functions, multidimensional calculus, optimisation methods. Most of the relevant material is contained in the following books:

- Casella and Berger, *Statistical Inference* (1990), ch. 1-10
- Grimmett and Stirzacker, *Probability and Random Processes*, ch. 1, 2, 3.1-3.8, 4, 5, 6.1-6.6, 7.1-7.5, 10, 12.1-12.5
- Khuri, *Advanced Calculus with Applications in Statistics*, ch. 2-8

STAT II This 4-hour exam is based on our more applied courses STAT 479, STAT 361, STAT 368, STAT 378. It includes the following topics: stationary time series, spectral analysis, filtering, Box-Jenkins methodology; basic sampling schemes and methods of estimation; design and analysis of experiments; multiple regression. Most of the relevant material is contained in the following books:

- Schumway and Stoffer, *Time Series Analysis and Its Applications*, ch. 1-4.
- Cochran, *Sampling Techniques*, ch. 1-5, 6 and 7.
- Montgomery, *Design and Analysis of Experiments*, ch. 1-7, 11, 12
- Neter, Wasserman and Kutner, *Applied Linear Statistical Models* , ch. 7-13

The exam questions, like the sample questions given below, are divided into two 'levels':

Level A These questions are meant to be rather elementary in nature. A poor performance on them might call for the student to take the corresponding undergraduate courses before proceeding into our graduate courses.

Level B These questions are somewhat more challenging; they are generally at the level of final exam, or assigned, questions in the corresponding courses. A satisfactory performance here indicates that the student should proceed directly into our 'core' graduate courses.

The core STAT courses are:

STAT 512 Techniques of Mathematics for Statistics

STAT 532 Survival Analysis

STAT 561 Sample Survey Methodology

STAT 562 Discrete Data Analysis

STAT 568 Design and Analysis of Experiments

STAT 571 Applied Measure Theory for Probability

STAT 575 Multivariate Analysis

STAT 578 Regression Analysis

STAT 664 Theory of Statistical Inference

STAT 665 Asymptotic Methods in Statistical Inference

At least 6 of these courses are to be taken by every Ph.D. (Statistics) student with at least 4 of them taken in the Qualifying Year. The 6 must include STAT 664, 665, 571.

1.1 Sample Questions - STAT I, Level A

1. Let $X_1, ..., X_n$ be a random sample from a distribution with p.d.f. $f_X(x|\theta)$. Find a sufficient statistic and the MLE for θ under the following cases:

   (a) $f_X(x|\theta) = e^{-(x-\theta)},\ x > \theta,\ \ -\infty < \theta < \infty,$

   (b) $f_X(x|\theta) = I_{[\theta < x < \theta+1]}$.

2. On a multiple-choice test each question has m multiple-choice answers. A student either knows the answer to a question or guesses. Let p be the probability that a student knows the answer to a question, and $q = 1 - p$ the probability that the student guesses. Assume that a student who guesses will be correct with probability $\frac{1}{m}$. What is the probability that a student who correctly answers a question knew the answer and did not guess?

3. Suppose that the joint distribution of a random vector $\mathbf{X}$ is determined by a parameter vector $\boldsymbol{\theta}$. Let $T = T(\mathbf{X})$ be a sufficient statistic for θ. Suppose that $W = W(\mathbf{X})$ is an unbiased estimator for $\tau(\theta)$ with finite variance. Prove that there exists an unbiased estimator for $\tau(\theta)$ that is a function of T and has variance less than or equal to that of W. In other words, prove the Rao-Blackwell Theorem.

4. Male and female customers arrive at a store, independently of each other. The number of males arriving in $[0, t]$ is a Poisson random variable with mean $\lambda_1 t$. The number of females is Poisson with mean $\lambda_2 t$.

   (a) What is the distribution of $N_t = \#$ of customers arriving in $[0, t]$?

   (b) Given that $N_t = n$, what is the conditional distribution of the number of male arrivals in $[0, t]$?

5. Show that if $A \subset \mathbb{R}$ is an open set, then its complement $\bar{A}$ is a closed set.

6. Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ idempotent matrices. Show that $\mathbf{A} - \mathbf{B}$ is idempotent if and only if $\mathbf{AB} = \mathbf{BA} = \mathbf{B}$.

1.2 Sample Questions - STAT I, Level B

1. Let the random variable X have a Poisson distribution with parameter $\theta \in \Omega = (0, \infty)$. Let the loss function be $L(\theta, a) = \frac{(\theta - a)^2}{\theta}$, where $a \in [0, \infty)$.

   (a) Find the risk function of the usual estimator $d(X) = X$ of θ.

   (b) Find the Bayes estimate of θ with respect to the prior distribution $G(\alpha, \beta)$, the gamma distribution with parameters α and β.

(c) Using part (b) show that the estimator $d(X) = X$ is minimax.

2. Let $X_1, \ldots X_n$ be a sample from the $N(\theta, 1)$ distribution. The variance σ^2=1 is known. Fix $0 < \alpha < 1$.

   (a) Does there exist a uniformly most powerful size α test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$? Support your answer with an appropriate argument.

   (b) Does there exist a uniformly most powerful size α test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$? Support your answer with an appropriate argument.

3. Members of a large (assumed infinite) population are either immune to a given disease or are susceptible to it. Let X_n be the number of susceptible members in the population at time period n. Suppose that $X_0 = 0$, and that in the absence of an epidemic $X_{n+1} = X_n + 1$. Thus, in the absence of an epidemic, the number of susceptibles increases in time, possibly owing to individuals losing their immunity, or to the introduction of new susceptible members to the population.

   But in each period there is a constant but unknown probability p of a disease. When the disease occurs all susceptibles are stricken. The disease is non-lethal and confers immunity, so that if T is the first time of disease occurrence, then $X_T = 0$.

   Compute the stationary distribution for the Markov chain $\{X_n : n = 0, 1, 2, \ldots\}$.

4. Prove the following: Let $\{Z_n\}$ denote a branching process with $Z_0 = 1$ and $G_n(s) = E[e^{Z_n}]$. Then

   $$P(Z_n = 0) \rightarrow \eta = P(\text{ultimate extinction}) \text{ as } n \rightarrow \infty,$$

   where η is the smallest non-negative root of the equation $s = G_1(s)$.

5. Suppose that $f$and g are continuous functions on D, and $g(x) \neq 0$ there. Show that f/g is continuous on D.

6. Consider the problem of estimating a parameter vector $\boldsymbol{\theta}$, by Least Squares, in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \textit{random error}$. Here $\mathbf{y} : n \times 1$ and $\mathbf{X} : n \times p$ are constants and $\mathbf{X}$ has rank $p < n$. Suppose that the parameters are required to satisfy q independent linear constraints of the form $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}_{q\times 1}$, where $\mathbf{A}_{q\times p}$ has rank q. Thus the mathematical problem is

   $$\begin{aligned} \text{minimize } S(\boldsymbol{\theta}) &= ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2 \\ \text{over } \boldsymbol{\theta} &\in \mathbb{R}^p, \text{ subject to constraints} \\ \mathbf{A}\boldsymbol{\theta} &= \mathbf{0}_{q\times 1}. \qquad (*) \end{aligned}$$

   Show that any solution to this problem is of the form

   $$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\mathbf{t},$$

   where $\hat{\boldsymbol{\theta}}_0$= $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbf{t}$ is a $q \times 1$ vector chosen so that $\hat{\boldsymbol{\theta}}$ satisfies (*).

1.3 Sample Questions - STAT II, Level A

1. Suppose that two random variables X, Y have joint probability density function

$$g(x, y) = 6e^{-(3x+2y)},\ 0 \leq x, y \leq \infty.$$

   (a) Are X and Y independent? Why or why not?

   (b) What is the marginal probability density of X?

2. In an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship between $\log_{10}(dose)$ $(= X_1)$ and $\log_{10}$ $(larva\ weight)$ $(= X_2)$ was sought. The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (i.e., time until death), were used to produce the following computer results and the ANOVA table with $Y = \log_{10}(survivaltime)$:

$$\begin{aligned}\widehat{Y} &= 2.952 - 0.550X_1\\ \widehat{Y} &= 2.187 + 1.370X_2\\ \widehat{Y} &= 2.593 - 0.381X_1 + 0.871X_2\end{aligned}$$

Source	d.f	Sums of squares
Regression (X_1, X_2)	2	0.4637
Residual	12	0.0476

Source	d.f	Sums of squares
Regression (X_1)	1	0.3633
Residual	13	0.1480

Source	d.f	Sums of squares
Regression (X_2)	1	0.3332
Residual	13	0.1780

   (a) Test for the significance of the overall regression involving both independent variables X_1 and X_2.

   (b) Test to see whether using X_1 alone helps significantly in predicting survival time.

   (c) Test to see whether using X_2 alone helps significantly in predicting survival time.

   (d) Compute R^2 for each of the three models.

   (e) Which independent variable do you consider to be the best single predictor of survival time?

(f) Which model involving one or both of the independent variables do you prefer and why?

3. Give a brief but specific example of *(a)* an experimental study and *(b)* an observational study. Use your examples to comment on the differences between the two kinds of studies, with regard to the way in which the studies are carried out and the inferences that can be drawn from the data analysis.

4. In a survey conducted in fall, 1995, by the Survey Research Center at the University of Kentucky, it was found that 268 out of 534 households have no children under age 18 living at home. Obtain an approximate 95% confidence interval for the true proportion of households in the state of Kentucky that have no children under age 18 living at home. Assume that N is effectively large.

5. All of the farms in a county are stratified by farm size and the mean number of hectares of wheat per farm in each stratum, with the following results:

Farm size	**# of farms**	**Mean wheat (hectares)**	**Standard deviation**
$0-20$	368	2.7	2.1
$21-40$	425	8.1	3.6
$41-60$	389	12.1	3.9
$61-80$	316	16.9	5.1
$81-100$	174	20.8	6.1
$100-$	236	28.5	6.5

For a random sample of 100 farms, compute the sample sizes in each stratum under stratified random sampling with

(a) proportional allocation;

(b) Neyman allocation.

Assume that the cost per observation is the same for all strata.

6. Let $\{X_t\}$ be a first-order moving average (MA(1)) process, i.e.

$$X_t = w_t - \theta w_{t-1},$$

where $\{w_t\}$ is a sequence of independent, identically distributed random variables with mean zero and variance σ_w^2. Compute the autocovariance function

$$R(m) = cov[X_t, X_{t+m}],\ m = 0, \pm 1, \pm 2, \ldots.$$

7. An investigator is planning a randomized complete block design to compare four treatments. She plans to run five blocks, with each treatment occurring once in each block. Based on a previous study, she expects the error standard deviation within blocks to be approximately $\sigma = 2$. She hopes to observe a statistically significant difference among the treatment means at level $\alpha = .05$ if the difference between the largest and smallest mean is at least 6.

   (a) Determine the smallest value possible for the noncentrality parameter ϕ in this situation.

   (b) Find the corresponding minimum power attainable. Specify the values of ϕ, α and of the degrees of freedom that you use to calculate the power. You may use the fact that the noncentrality parameter for single factor studies (with r factors) is

$$\phi = \frac{1}{\sigma}\sqrt{\frac{1}{r}\sum_i n_i(\mu_i - \mu_.)^2}.$$

8. Assume that $(a_t)_{-\infty<t<\infty}$ is a white noise process; that is, $E(a_t) = 0$ for all t, and $\mathrm{Cov}\,(a_s, a_t) = \sigma_a^2\delta_{\{s=t\}}$ for all s, t. For $|\phi| < 1$, let $(Z_t)_{t\geq 0}$ be the stationary autoregressive process given by

$$Z_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots + \phi^k a_{t-k} + \cdots$$

   For all $k \geq 0$, calculate $\rho(k) = \mathrm{Cov}\,(Z_t, Z_{t-k})$ in terms of ϕ and σ_a^2.

1.4 Sample Questions - STAT II, Level B

1. Consider the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with *i.i.d.* errors ε. For any possible estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ define the sum of squares of residuals by

$$S(\hat{\boldsymbol{\beta}}) = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2.$$

   Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ be the 'hat' matrix.

   (a) Show that $S(\hat{\boldsymbol{\beta}}) = ||\mathbf{H}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})||^2 + ||(\mathbf{I} - \mathbf{H})\mathbf{y}||^2$.

   (b) Whether you have answered part (a) or not, continue from there to derive the Least Squares Estimate $\hat{\boldsymbol{\beta}}$.

2. Three plants (Factor A) of the same variety were randomly selected in an experiment to investigate the concentration of a particular acid. Two leaves (Factor B) per plant were randomly selected and two separate determinations of the acid concentration were obtained per leaf. The data are:

Plant 1		Plant 2		Plant 3	
Leaf 1	Leaf 2	Leaf 1	Leaf 2	Leaf 1	Leaf 2
11, 15	10, 12	11, 13	12, 16	17, 21	14, 16

(a) Write down an appropriate model for this study, indicating which factors are nested or crossed and describing distributional assumptions or side conditions on effects.

(b) Complete the following ANOVA table:

Source	df	SS	MS	F	E(MS)
A					
B(A)					
Error					

3. In the multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ with *i.i.d.* errors ε, consider the problem of constructing simultaneous confidence intervals on all linear combinations $\mathbf{x}^T\boldsymbol{\beta}$. We want intervals of the form

$$\mathbf{x}^T\hat{\boldsymbol{\beta}} \pm c \cdot \ \text{(estimated standard deviation of } \mathbf{x}^T\hat{\boldsymbol{\beta}}),$$

and so we seek that value of c so that, before sampling, the simultaneous coverage probability is $1-\alpha$:

$$1-\alpha = P\left(|\mathbf{x}^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\boldsymbol{\beta}| \leq c \cdot \text{(estimated std. dev. of } \mathbf{x}^T\hat{\boldsymbol{\beta}}) \text{ for } \underline{\text{all}}\ \mathbf{x}\right).$$

Show that $c^2 = pF^p_{n-p}(1-\alpha)$, where n and p are the number of rows and number of columns of $\mathbf{X}$.

4. From a list of 100 households in a certain small town, 10 households are drawn using SRSWOR. The data for the number of teenagers(x) and the average number of hours spent per week (y) spent by a teenager watching television are as follows:

x	2	1	3	2	4	3	2	1	3	1
y	15	20	16	14	17	12	9	5	15	8

Find a ratio estimate of the average number of hours spent by a teenager watching television in that town. Also, obtain an estimate for the variance of the estimate.

5. In building a model to study automobile fuel consumption, Biggs and Akcelik (1987, *J. Transportation Engineering*) begin by looking at the relationship between idle fuel consumption (y) and engine capacity (x) for a sample of $n = 20$ cars. Suppose a summary of the data is as follows:

$$\sum y = 6.25,\ \sum y^2 = 2.90,\ \sum x = 39.5,\ \sum x^2 = 120.4,\ \sum xy = 18.2.$$

   (a) If the average engine capacity is 2.5 litres for the population of automobiles, estimate the average idle fuel consumption using a regression type estimator, with a bound on the error of estimation.

   (b) Find the relative efficiency of your estimator to the ratio type estimator.

   (a) Identify the following as a particular $ARIMA\ (p, d, q) \times (P, D, Q)_S$ model:

$$X_t = X_{t-1} + X_{t-12} - X_{t-13} + w_t - \theta w_{t-1} - \Theta w_{t-12} + \theta\Theta w_{t-13}.$$

   (b) Write out an expression, similar to that given in part (a), for an $ARIMA\ (0, 1, 1) \times (1, 0, 1)_4$ model.

6. In a drilling operation four factors (A, B, C and D), each at three levels, are thought to be of importance in influencing the volume of crude oil pumped. Using Taguchi's orthogonal array the factors A, B, C and D are assigned to columns 1, 2, 3 and 4 respectively. The response variable showing the number of barrels (in thousands) pumped per day for each of the nine experiments is shown in the table.

Experiment #	1	2	3	4	5	6	7	8	9
Number of barrels (000's)	6.8	15.8	10.5	5.2	17.1	3.4	5.9	12.2	8.5

   Show the experimental design and the response variable for the corresponding experiments. Determine the main effects. Plot the average response curves. What are the optimal settings of the design parameters?

7. Obtain the power spectrum of an $ARMA(p, q)$ series, defined in operator notation by $\phi(B)X_t = \theta(B)w_t$, in terms of the characteristic polynomials.