Masters Comprehensive Examination
Department of Statistics, University of Florida
August 21, 2003, 2:00-6:00pm

Instructions:

1. You have four hours to answer questions in this examination.
2. You must show your work to receive credit.
3. There are 10 problems of which you must answer 8.
4. Only your first 8 problems will be graded.
5. The last page contains two tables that you may need.
6. While the 10 questions are equally weighted, some problems are more difficult than others.
7. The parts within a given question are not necessarily equally weighted.
8. Write only on one side of the paper, and start each question on a new page.
9. You are allowed to use a calculator.

The following abbreviations are used throughout:

- ANOVA = analysis of variance
- CPA = certified public accountant
- CRD = completely randomized design
- iid = independent and identically distributed
- mgf = moment generating function
- ML = maximum likelihood
- MLR = monotone likelihood ratio
- MOM = method of moments
- pdf = probability density function
- pmf = probability mass function
- UMP = uniformly most powerful
- UMVUE = uniformly minimum variance unbiased estimator
- $\varepsilon_i \sim NID(0, \sigma^2)$ means that the ε_is are iid $\mathrm{N}(0, \sigma^2)$.

1. Let $z_1, \ldots, z_n$ be n real numbers such that $z_1 < z_2 < \cdots < z_n$ where $n \geq 2$. Let Z denote the discrete random variable that takes the values $z_1, \ldots, z_n$ with probabilities $p_1, \ldots, p_n$ where, of course, $p_i \in [0, 1]$ for each $i \in \{1, \ldots, n\}$ and $\sum_{i=1}^n p_i = 1$. Clearly, the distribution of Z depends on the value of $p = (p_1, \ldots, p_n)$. In this problem, we will answer the question: "What value of p leads to the most variable Z?"

(a) Let X be a random variable such that $\mathrm{E}X^2 < \infty$. Show that, for any real number c,

$$\mathrm{Var}(X) \leq \mathrm{E}[(X - c)^2] \ .$$

(b) Give a geometric argument showing that

$$\left| z_i - \frac{(z_1 + z_n)}{2} \right| \leq \frac{z_n - z_1}{2} \ ,$$

for all $i \in \{1, 2, \ldots, n\}$.

(c) Use (a) and (b) to show that

$$\mathrm{Var}(Z) \leq \frac{(z_n - z_1)^2}{4} \ .$$

(d) Find the value of p that results in the most variable Z by showing that it is possible to attain the upper bound in (c).

2. Two sentries (guards) are sent to patrol a road that is θ kilometers long. The two sentries are sent to stand at the points X_1 and X_2, chosen randomly (uniformly) and independently, along the road.

(a) Find the joint distribution of $Y_1 = X_1 - X_2$ and $Y_2 = X_2$.

(b) Find the marginal distribution of $Y_1 = X_1 - X_2$.

(c) Compute the probability that the sentries will be less than $\theta/2$ kilometers apart when they reach their assigned posts.

3. Let $X_1, \ldots, X_n$ be iid with common pdf given by

$$f(x|\alpha, \beta) = \frac{\beta\, \alpha^\beta}{x^{\beta+1}}\, I(x \geq \alpha)$$

where $\alpha > 0$ and $\beta > 2$. The fact that β is restricted to be larger than 2 is very important! Let $X = (X_1, \ldots, X_n)$. Also, let $\overline{X}$ denote the sample mean and set $\overline{\overline{X}} = \frac{1}{n} \sum_{i=1}^n X_i^2$.

(a) Find the ML estimators of α and β, call them $\hat{\alpha} = \hat{\alpha}(X)$ and $\hat{\beta} = \hat{\beta}(X)$.

(b) Show that for any $k \in (-\infty, 2]$,

$$\mathrm{E}X^k = \frac{\beta\, \alpha^k}{\beta - k} \ .$$

(c) Now consider the MOM estimators of α and β, call them $\tilde{\alpha} = \tilde{\alpha}(X)$ and $\tilde{\beta} = \tilde{\beta}(X)$. Show that the MOM estimator of β is a solution of

$$\big(\overline{X}^2 - \overline{\overline{X}}\big)\beta^2 - 2\big(\overline{X}^2 - \overline{\overline{X}}\big)\beta + \overline{X}^2 = 0 \ .$$

(d) Find $\tilde{\alpha}(X)$ and $\tilde{\beta}(X)$.

(e) One serious problem with MOM estimators in general is that (unlike ML estimators) they are not guaranteed to fall inside the parameter space. Can $(\tilde{\alpha}, \tilde{\beta})$ fall outside of the parameter space?

4. Assume that $X_1, X_2, \ldots, X_n$ are iid Poisson(λ).

(a) Use the fact that the Poisson is a member of the exponential family to show that $\sum_{i=1}^n X_i$ is a complete, sufficient statistic.

(b) Derive the mgf of X_1 and use it to find the distribution of $\sum_{i=1}^n X_i$.

(c) Find an unbiased estimator of $e^{-\lambda}$.

(d) Find the UMVUE of $e^{-\lambda}$.

5. Suppose that X_1, X_2, X_3 are iid from the following pmf

$$P_\theta(X = x) = \frac{(x+1)}{\theta^2}\left(\frac{\theta}{\theta+1}\right)^{x+2} I_{\mathbb{Z}^+}(x) ,$$

where $\mathbb{Z}^+ := \{0, 1, 2, \ldots\}$ and $\theta > 0$.

(a) Show that $W = X_1 + X_2 + X_3$ is a sufficient statistic for θ.

(b) Derive the Law of Total Probability. More specifically, suppose that S is a sample space, A is a subset of S and $\{B_0, B_1, B_2, \ldots\}$ is a partition of S and show that

$$P(A) = \sum_{i=0}^{\infty} P(A|B_i)\, P(B_i) .$$

(c) Derive the pmf of W. (Hint: Start by deriving the pmf of $X_1 + X_2$ - don't bother trying to simplify the sum.)

(d) Show that the family of mass functions of W has MLR.

(e) Find a UMP test (based on X_1, X_2, X_3) of $H_0 : \theta \leq 1$ against $H_1 : \theta > 1$ with level $\frac{15}{16}$.

6. Consider the CRD with 3 random treatments, 4 replicates per treatment and the following model structure:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} ,$$

where $\alpha_i \sim NID(0, \sigma_a^2)$, $\varepsilon_{ij} \sim NID(0, \sigma_e^2)$ and Cov$(\alpha_i, \varepsilon_{ij}) = 0$ for all i, j.

(a) Give the Treatment (Between) and Error (Within) mean squares and derive the expected mean square for treatments.

(b) Suppose the sample means (standard deviations) for the three treatments are: 45 (7), 60 (9), 75 (8), respectively.

1. Obtain the ANOVA.
2. Test $H_0 : \sigma_a^2 = 0$ vs $H_A : \sigma_a^2 > 0$ at $\alpha = 0.05$ significance level.
3. Give an unbiased estimate of σ_a^2.

7. A regression model is fit relating a response Y to a set of three predictor variables, X_1, X_2, X_3, based on $n = 100$ subjects. They consider two models:

$$\text{Model 1: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad \varepsilon \sim NID(0, \sigma_1^2)$$

and

$$\text{Model 2: } Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \varepsilon \sim NID(0, \sigma_2^2) \ .$$

The following output was obtained for Model 1:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.8482 & -0.0219 & -0.1163 & -0.0070 \\ -0.0219 & 0.0025 & 0.0006 & -0.0002 \\ -0.1163 & 0.0006 & 0.0200 & 0.0007 \\ -0.0070 & -0.0002 & 0.0007 & 0.0010 \end{bmatrix}, \qquad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 728 \\ 6067 \\ 3947 \\ 3573 \end{bmatrix}$$

and $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y} = 395.70$. The corresponding quantities for Model 2 were:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.1631 & -0.0192 \\ -0.0192 & 0.0024 \end{bmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 728 \\ 6067 \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y} = 400.77$$

(a) Give the least squares estimate of the regression coefficient for X_1 for each model as well as 95% confidence intervals for the true parameters.

(b) Test whether Y is associated with X_2 and/or X_3 after controlling for X_1 at the $\alpha = 0.05$ significance level.

8. An accounting firm has four CPAs (blocks). They are considering three computer programs (treatments) for filing individual tax returns. The firm obtains a client's tax information and has each CPA use each computer program to file the return. The response, Y, is the time to complete the return (in minutes). The CPAs use the programs in random order and the times are given in the following table.

CPA	Prog 1	Prog 2	Prog 3
1	35	40	45
2	25	35	45
3	45	50	55
4	15	35	55

(a) Give the treatment (program) sum of squares and its degrees of freedom.

(b) Give the block (CPA) sum of squares and its degrees of freedom.

(c) Give the error (block by treatment interaction) sum of squares and its degrees of freedom.

(d) Test whether the three programs differ significantly with respect to mean completion times ($\alpha = 0.05$ significance level.)

(e) Give the minimum significant difference when comparing pairs of programs, with an experimentwise error rate of $\alpha = 0.05$ significance level, based on Bonferroni's method.

9. A regression model is to be fit, relating mean sales to weekly newspaper advertising in the range of $X = 0 - 10$ units of advertising. They know from past experience that the mean sales when $X = 0$ is $\mu_0 = 10$. Further, they believe the relationship will be nonlinear, and can be approximated by a quadratic model in this range of X values. Derive the least squares estimating (normal) equations **(in scalar form)** for their model.

10. Two statistical analysts are given the same data for a single response variable Y, and a single independent variable. The independent variable has been labeled with levels 1,2,3. John treats the independent variable as interval scale, and assumes that the relationship between the dependent and independent variables is linear (he has no reason to believe the mean response is 0 when the independent variable is 0). Jane treats the independent variable as nominal scale (no distinct ordering among the levels), making no assumption about the relationship between the dependent and independent variables. Both John and Jane believe that error terms are independent and normally distributed with constant variance.

(a) Write out John's statistical model.

(b) Write out Jane's statistical model.

(c) The following data were obtained. Give least squares estimates of all model parameters for John and Jane.

Trt (X)	Responses (Y)		
1	46	50	54
2	32	38	35
3	3	5	7

(d) Give John's and Jane's Analyses of Variance.

(e) State the null and alternative hypotheses for John and Jane to determine whether there is an association between treatment (X) and response (Y).

(f) Conduct your tests in part (e), each at $\alpha = 0.05$. Note, there is no need to adjust for simultaneous tests, as John and Jane are working independently.

(g) Use the F-test for lack of fit to determine whether John's model is appropriate (H_0) or Jane's is (H_A), with $\alpha = 0.05$.