First Year Examination
Department of Statistics, University of Florida
August 20, 2004, 8:00 am - 12:00 noon

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. While the 10 questions are equally weighted, some problems are more difficult than others.
6. The parts within a given question are not necessarily equally weighted.
7. You are allowed to use a calculator.
8. **Write only on one side of the paper, and start each question on a new page.**

The following abbreviations are used throughout:

- ANOVA = analysis of variance
- CPA = certified public accountant
- iid = independent and identically distributed
- mgf = moment generating function
- ML = maximum likelihood
- MLR = monotone likelihood ratio
- MSE = mean squared error
- pdf = probability density function
- pmf = probability mass function
- UMP = uniformly most powerful
- $\varepsilon_i \sim NID(0, \sigma^2)$ means that the ε_is are iid N$(0, \sigma^2)$.

1. Let $f(x)$ be a pdf and let a be a number such that, for all $\varepsilon > 0$, $f(a+\varepsilon) = f(a-\varepsilon)$. Such a pdf is said to be *symmetric* about the point a.

(a) Show that if $X \sim f(x)$, symmetric about a, then the median of X is the number a.

(b) Show that if $X \sim f(x)$, symmetric about a, and $\mathrm{E}X$ exists, then $\mathrm{E}X = a$.

(c) Show that if $X \sim f(x)$, symmetric about a, then $Y = X - a$ is symmetric about the point 0.

(d) Suppose $X \sim f(x)$, symmetric about a, and define $Y = X - a$. Identify another random variable, call it Z, that has the same distribution as Y, but is such that $P(Y = Z) = 0$.

2. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and (finite) variance σ^2. Let $a_1, \ldots, a_n$ be constants.

(a) Show that the estimator $\sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ if and only if $\sum_{i=1}^n a_i = 1$.

(b) Among all unbiased estimators of this form (called *linear unbiased estimators*) find the one with minimum variance *and* calculate the variance.

Now let $W_1, \ldots, W_k$ be unbiased estimators of a parameter θ with $\mathrm{Var}(W_i) = \sigma_i^2$ and $\mathrm{Cov}(W_i, W_j) = 0$ if $i \neq j$. Assume that the σ_i^2s are all known and finite and, again, let $a_1, \ldots, a_k$ be constants.

(c) Show that, of all the unbiased estimators of θ having the form $\sum_{i=1}^k a_i W_i$, the one with the smallest variance is

$$\phi(W_1, \ldots, W_k) = \frac{\sum_{i=1}^k \frac{W_i}{\sigma_i^2}}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} ,$$

and that

$$\mathrm{Var}\left[\phi(W_1, \ldots, W_k)\right] = \frac{1}{\sum_{i=1}^k \frac{1}{\sigma_i^2}} .$$

3. Let $X_1, \ldots, X_n$ be iid Uniform$(0, \theta)$ where $\theta > 0$; that is, the common pdf is

$$f(x|\theta) = \theta^{-1} I(0 \leq x \leq \theta) .$$

(a) Find the ML estimator of θ, call it $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

(b) Find the pdf of $\hat{\theta}$ and show that $\hat{\theta}/\theta$ has a beta distribution.

(c) Show that $n\left(1 - \frac{\hat{\theta}}{\theta}\right)$ converges in distribution and find the limiting distribution.

(d) Find the method of moments estimator of θ, call it $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$.

(e) Compare the two estimators using MSE.

4. Suppose that $X_1, \ldots, X_n$ are iid random variables such that

$$P(X_1 = x) = p(1-p)^x \text{ for } x = 0, 1, 2, \ldots$$

where $p \in (0, 1)$.

(a) Does the the mgf of X_1 exist? If so, what is it?

(b) Suppose that $Y \sim \text{NB}(r, s)$; that is,

$$P(Y = y) = \binom{r+y-1}{y} s^r (1-s)^y \text{ for } y = 0, 1, 2, \ldots$$

where $s \in (0, 1)$ and $r \in \{1, 2, 3, \ldots\}$. Find the mgf of Y.

(c) Find the pmf of the random variable $Z = \sum_{i=1}^n X_i$.

(d) Find the ML estimator of $g(p) = p(1-p)$, call it $\widehat{g(p)}$.

(e) Is $\widehat{g(p)}$ the best unbiased estimator of $g(p)$? If not, find the best unbiased estimator of $g(p)$.

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. 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 1,2,3,4. 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	17	20	23
2	29	21	25
3	31	29	30
4	45	43	47

(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 $\alpha = 0.05$.

7. A chemical production process consists of a first reaction with an alcohol and a second reaction with a base. A 3×2 factorial with three alcohols and 2 bases (these are the only alcohol and base types of interest to the researcher) was conducted with three replicate reactions per treatment. The design was completely randomized. The data are given in the following table:

	Base	
Alcohol	1	2
1	93, 90, 87	77, 83, 80
2	80, 78, 82	91, 88, 91
3	81, 89, 85	84, 84, 87

(a) Write a linear model for the experiment, explaining all terms. Compute the Analysis of Variance.

(b) Test whether there is a base by alcohol interaction with respect to yields. Use the $\alpha = 0.05$ significance level. Fully state null and alternative hypotheses with respect to your model parameters, give test statistic, rejection region, and a sketch representing the P-value of the test.

(c) Compare the two bases under each alcohol type using Bonferroni's method with simultaneous 95% confidence intervals (how many comparisons are being made?). Clearly describe your conclusions. Give an interaction plot of results in terms of the sample means.

8. An internet based firm has 2 sources of advertising: newspaper/magazines and television/radio. They vary the amounts of advertising of each type (X_1 and X_2, in thousands of dollars, respectively) across a sample of $n = 8$ similar sized markets, and obtain how many times their website is reached by people from each market (Y, in 1000s of "hits"). They fit the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \varepsilon_i \sim NID(0, \sigma^2)$$

and obtain the following estimated regression equation, residual mean square, and $(X'X)^{-1}$ matrix:

$$\hat{Y}_i = 1.8 + 4.5X_{i1} + 3.5X_{i2} \qquad s^2 = 1.87 \qquad \left(X'X\right)^{-1} = \begin{bmatrix} 0.4 & -0.12 & -0.12 \\ -0.12 & 0.096 & 0.016 \\ -0.12 & 0.016 & 0.096 \end{bmatrix}$$

(a) Give the predicted number of "hits" if they spend \$2000 on newspaper/magazines and \$2000 on television/radio (be careful of units).

(b) Test whether the mean number of hits when $X_1 = 2.0$ and $X_2 = 2.0$ is equal to 20.0. Set this up and conduct it as a test of the form: $H_0 : K'\beta = m$ and test at $\alpha = 0.05$ significance level.

(c) Suppose they fit a response surface and obtained the following least squares estimate of the regression equation containing an intercept, all linear and quadratic terms, and a cross-product term.

$$\hat{Y} = 0.5 + 6.0X_1 + 4.0X_2 - 1.0X_1^2 - 0.5X_2^2 + 1.0X_1X_2 \ .$$

Their budget is limited to \$5000 (recall units), and they must spend all of it. How should they allocate their budget between newspaper/magazine ads (X_1) and television/radio advertising (X_2) to maximize the predicted number of hits based on the estimated regression equation.

9. 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.

10. Consider the matrix form of the linear regression model: $Y = X\beta + \varepsilon$ where $\varepsilon \sim \text{N}(0, \sigma^2 I)$. The least squares estimator of β is, of course, $\hat{\beta} = (X'X)^{-1}X'Y$.

(a) Derive the mean vector and variance-covariance matrix of $\hat{\beta}$.

(b) Write out the vector of predicted values $\hat{Y}$ as a function of $\hat{\beta}$. Derive its mean vector and variance-covariance matrix.

(c) Write out the vector of residuals, e. Derive its mean vector and variance-covariance matrix.

(d) Show that the vector of residuals and the vector of predicted values are orthogonal; that is, that the sum of the products of the fitted values and residuals is 0.

(e) Was the restriction on the residual vector necessary in parts (a)-(d)? **Yes** or **No**