First Year Examination
Department of Statistics, University of Florida
May 13, 2005, 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. **Write only on one side of the paper, and start each question on a new page.**
4. There are 10 problems of which you must answer 8.
5. Only your first 8 problems will be graded.
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. You are allowed to use a calculator.

The following abbreviations and terminology are used throughout:

- ANOVA = analysis of variance
- cdf = cumulative distribution function
- SS = sums of squares
- *corrected total sum of squares* = total SS corrected for the mean
- iid = independent and identically distributed
- LRT = likelihood ratio test
- MOM = method of moments
- MSE = mean squared error
- ML = maximum likelihood
- pdf = probability density function
- α = specified probability of Type I error
- $\mathbb{N} = \{1, 2, 3, \dots\}$
- N(μ, σ^2) = normal distribution with mean μ and variance σ^2

1. Suppose the random variables Y_{ij}, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ satisfy the *oneway ANOVA assumptions*; that is,

$$Y_{ij} = \theta_i + \varepsilon_{ij} \ ,$$

where the ε_{ij} are independent with $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ and $\theta_1, \theta_2, \ldots, \theta_k$ and σ^2 are all unknown parameters. Consider testing $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$ versus H_1 : Not H_0.

(a) Write down the usual "F-test." (Carefully define any notation that you introduce.)

(b) Derive the LRT.

(c) Prove or disprove the following statement: "The LRT is equivalent to the F-test."

2. Consider an experiment in which three fair dice are tossed.

(a) Let $\mathcal{S}$ denote the set of all random variables defined on this experiment whose range is a subset of $\{0, 1\}$. (Note that $\{0, 1\}$ is a subset of $\{0, 1\}$.) How many random variables does $\mathcal{S}$ contain?

(b) Each random variable in $\mathcal{S}$ has a probability distribution. Are the distributions all different? If not, how many unique distributions are there?

(c) Define X and Y to be the smallest and largest of the three up faces, respectively. For example, if the result of the toss is $(5, 6, 2)$ then $x = 2$ and $y = 6$. Find the joint mass function of (X, Y).

(d) Are X and Y independent? (A yes/no response is <u>not</u> sufficient.)

(e) Find the conditional mass function of Y given $X = 3$.

(f) Now suppose that after the dice are tossed, we flip a fair coin $X + Y$ times. Let Z denote the number of heads. For each $t \in \{2, 3, \ldots, 12\}$, find the probability that $X + Y = t$ given that $Z = 10$.

3. Let $X_1, \ldots, X_n$ be iid random variables from a continuous population with cdf $F(x)$ and pdf $f(x)$.

(a) Derive the cdf of $X_{(j)}$ for $1 \leq j \leq n$.

Now suppose that $Y_1, \ldots, Y_m$ are iid with common pdf given by

$$f(y|\theta) = \begin{cases} \theta/y^2 & y > \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is an unknown parameter.

(b) Find the ML estimator of θ, call it $\hat{\theta}(Y)$, and calculate its bias and MSE.

(c) Find the best unbiased estimator of θ, call it $\tilde{\theta}(Y)$, and calculate its MSE.

(d) Compare and contrast $\hat{\theta}(Y)$ and $\tilde{\theta}(Y)$.

(e) How does the MOM estimator of θ perform in this case?

(f) Find a function $c : (0,1) \times \mathbb{N} \to (0, \infty)$ such that the interval estimator

$$\Big(c(\alpha, m)\hat{\theta}(Y), \hat{\theta}(Y)\Big)$$

has confidence coefficient $1 - \alpha$ for all $m \in \mathbb{N}$ and all $\alpha \in (0,1)$.

4. Let $X_1, \ldots, X_n$ be iid discrete uniform on $\{1, 2, \ldots, \theta\}$; that is, the common mass function is given by

$$f(x|\theta) = \theta^{-1} I_{\{1,2,\ldots,\theta\}}(x) \ ,$$

where $\theta \in \mathbb{N}$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_0 < \theta_1$ and $\theta_0, \theta_1 \in \mathbb{N}$. The relevant sample space, $\mathcal{X}$, is given by

$$\mathcal{X} = \Big\{(x_1, x_2, \ldots, x_n) : x_i \in \{1, 2, \ldots, \theta_1\} \ \text{ for each } \ i = 1, 2, \ldots, n\Big\} \ .$$

There are θ_1^n elements in $\mathcal{X}$; that is, $\#(\mathcal{X}) = \theta_1^n$. Recall that a test is nothing more than a partition of $\mathcal{X}$ into the rejection region and the acceptance region. Since there are only a finite number of sample points, there are only a finite number of possible tests.

(a) Define a set as follows

$$S = \Big\{(x_1, x_2, \ldots, x_n) \in \mathcal{X} : \max_{1 \leq i \leq n} x_i > \theta_0\Big\} \ .$$

Let $P_{\theta_0}(R)$ and $P_{\theta_1}(R)$ denote the size and power of the test with rejection region R. Show that $P_{\theta_0}(R)$ and $P_{\theta_1}(R)$ can both be written as simple functions of n, θ_0, θ_1, R and S.

(b) Prove that if $S \subset R$, then R is a most powerful test of its size.

(c) Suppose that the set $S \cap R^c$ is not empty. Is there a simple way to improve the test R? Explain.

5. Let X and Y be two random variables with finite second moments.

(a) Prove that $\text{Var}(X) = \text{Var}\big[\text{E}(X|Y)\big] + \text{E}\big[\text{Var}(X|Y)\big]$.

(b) Use the function $h(t) = \text{E}\Big\{\big[(X - \text{E}X)t + (Y - \text{E}Y)\big]^2\Big\}$ to prove that $-1 \leq \rho_{XY} \leq 1$ where ρ_{XY} denotes the correlation between X and Y.

(c) Prove or disprove the following statement: If $\text{Cov}(X, Y) = 0$, then X and Y are independent.

6. Consider a randomized complete block design with 10 blocks and a single factor having 4 treatment levels. Let Y_{ij} denote the response measured for the experimental unit in block j that receives treatment i for $i = 1, \ldots, 4$, $j = 1, \ldots, 10$. Suppose there is also a covariate whose value X_{ij} is measured for each experimental unit.

The following four models are fit to the data (using least squares), with the resulting residual (error) sums of squares as specified:

Model 1:	$Y_{ij} = \mu + \gamma_j + \epsilon_{ij}$	$\text{SS(Res)} = 400$
Model 2:	$Y_{ij} = \mu + \tau_i + \gamma_j + \epsilon_{ij}$	$\text{SS(Res)} = 300$
Model 3:	$Y_{ij} = \mu + \tau_i + \gamma_j + \beta X_{ij} + \epsilon_{ij}$	$\text{SS(Res)} = 100$
Model 4:	$Y_{ij} = \mu + \gamma_j + \beta X_{ij} + \epsilon_{ij}$	$\text{SS(Res)} = 250$

The treatment effects are $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_4)$ and the block effects are $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{10})$.

The corrected total sum of squares is 750.

1. Find the *sequential* sums of squares for $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, and β, in that order.
2. Form an ANOVA table for the randomized complete block design *without* the covariate X_{ij}, that is, based on Model 2. The table should include all appropriate sources of variation (including the corrected total), with degrees of freedom, sums of squares, and mean squares where appropriate. Then test whether or not there is a treatment effect based on this model. Use $\alpha = 0.05$.
3. Test whether there is a treatment effect, after accounting both for blocking and for the covariate. Use $\alpha = 0.05$.
4. Suppose the (possibly incorrect) model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ is fit to the data. Compute the residual sum of squares for this model.

7. Consider the linear model in the general matrix formulation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ is the vector of dependent variables, $\boldsymbol{X}$ is a matrix with full column rank, $\boldsymbol{\beta}$ is the vector of regression parameters, and the error vector $\boldsymbol{\epsilon}$ has a multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{I}\sigma^2$. ($\boldsymbol{I}$ = identity matrix)

In the following, carefully define any notation you use that is not introduced above.

1. Write the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in terms of $\boldsymbol{Y}$ and $\boldsymbol{X}$.
2. Form the vector $\boldsymbol{e}$ of ordinary least squares residuals. Derive its (multivariate) distribution.
3. A quadratic form in $\boldsymbol{Y}$ is an expression $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ where $\boldsymbol{A}$ is an appropriate symmetric matrix. Write the *un*corrected total and residual (error) sums of squares as quadratic forms in $\boldsymbol{Y}$.
4. If the variance-covariance matrix of $\boldsymbol{Y}$ is a multiple of the identity matrix, then two quadratic forms $\boldsymbol{Y}'\boldsymbol{A}_1\boldsymbol{Y}$ and $\boldsymbol{Y}'\boldsymbol{A}_2\boldsymbol{Y}$ are called *orthogonal* if $\boldsymbol{A}_1\boldsymbol{A}_2$ is a matrix of zeros. Show that the quadratic form for the residual (error) sum of squares is orthogonal to the quadratic form $\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

8. A movie production company pre-releases its films in two test markets, Los Angeles and New York City, before the general release. The total attendances at these test screenings in Los Angeles (X_1, in thousands) and New York City (X_2, in thousands) are used as predictors of eventual total box office revenue (Y, in millions of \$) in the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

Least-squares fitting of this model to data for 15 recent films yields (approximately)

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.80 \\ 6.28 \\ 13.59 \end{bmatrix} \qquad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 0.68 & -0.20 & -0.14 \\ -0.20 & 0.21 & -0.14 \\ -0.14 & -0.14 & 0.27 \end{bmatrix} \qquad \mathrm{SS(Res)} = 840$$

where $\boldsymbol{X}$ is the usual matrix of independent variables conforming to $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, and $\mathrm{SS(Res)}$ is the residual (error) sum of squares. Assume that the model is adequate and that the errors ϵ are independent and identically distributed with a normal distribution having mean zero.

1. Estimate the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.
2. Give an unbiased prediction $\widehat{Y}$ of the total box office revenue (millions of \$) of a new film with test market attendances of 1.5 thousand and 2.2 thousand in Los Angeles and New York City, respectively. Give an unbiased estimate of the variance of $\widehat{Y}$.
3. Test the hypothesis that β_1 and β_2 are equal: Determine $\boldsymbol{K}$ such that the null hypothesis may be written in the form $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{0}$, then perform an appropriate test. Use $\alpha = 0.05$.
4. Compute 95% simultaneous two-sided confidence intervals for β_0, β_1, and β_2, using the Bonferroni method.

9. An experiment is conducted to compare the bond strengths of 3 brands of glue and to determine whether or not cleaning the surfaces prior to glue application affects the strength. Treatments are completely randomized to 30 test surfaces such that each of the six treatment combinations of brand and cleaning status (cleaned or not) is assigned to five of the surfaces. The following table gives the average force required to break the bond, along with the corresponding sample *standard deviation* in parentheses, for each treatment combination:

	Brand 1	Brand 2	Brand 3
Cleaned	17.0 (1.0)	18.8 (0.9)	21.2 (1.1)
Not Cleaned	15.6 (0.6)	16.8 (1.2)	18.6 (0.8)

Assume that these three specific brands of glue are the only brands of interest to the experimenter, and that the error variance does not depend on the treatment combination.

1. Write a (univariate) linear model equation for this experiment. Explain each term and specify any conditions it satisfies.
2. Produce an ANOVA table with all appropriate sources of variation, including the (corrected) total. Include sums of squares, degrees of freedom, and appropriate mean squares.
3. Test whether there is any interaction between brand of glue and the condition of the surface (cleaned or not). Test whether brand of glue has any effect on bond strength. Test whether cleaning has any effect on bond strength. Use $\alpha = 0.05$ in all tests.

10. A balanced one-factor experiment with t factor levels and r replications at each level yields responses y_{ij} for replication j at treatment level i. Consider the following two alternative models for the data:

$$\text{Model I:} \quad y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \sum_{i=1}^{t} \tau_i = 0$$

$$\text{Model II:} \quad y_{ij} = \mu + a_i + \epsilon_{ij}, \quad a_1, \ldots, a_t \sim \text{iid } N(0, \sigma_a^2)$$

where the terms ϵ_{ij} are independent and identically distributed as $N(0, \sigma_e^2)$ (with $\sigma_e^2 > 0$) and are independent of all a_i in Model II.

1. For each model, write out the null hypothesis and the alternative hypothesis for the test of whether or not there are any factor effects.
2. In terms of the data values y_{ij}, write out the sum of squares for factor effect, $\mathrm{SS}(\mathrm{Factor})$, and the sum of squares for error, $\mathrm{SS}(\mathrm{Error})$. Also give expressions for their corresponding degrees of freedom.
3. Write an expression for the F-statistic (in terms of $\mathrm{SS}(\mathrm{Factor})$ and $\mathrm{SS}(\mathrm{Error})$) for testing the hypotheses in part (a). What is its distribution under each null hypothesis of part (a)?
4. For each model, find the *correlation* between two different responses that have the same treatment level.