First Year Examination
Department of Statistics, University of Florida
May 11, 2007, 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 the mean
- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- ML = maximum likelihood
- MLR = monotone likelihood ratio
- MSE = mean squared error
- NP = Neyman-Pearson
- pdf = probability density function
- pmf = probability mass function
- UMP = uniformly most powerful
- $\mathbb{R}^+ = (0, \infty)$
- N(μ, σ^2) = normal distribution with mean μ and variance σ^2

You may use the following facts/formulas without proof:

Beta density: $X \sim \text{Beta}(\alpha, \beta)$ means

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}\, I_{(0,1)}(x)$$

where $\alpha > 0$ and $\beta > 0$.

Gamma Density: $X \sim \text{Gamma}(\alpha, \beta)$ means X has pdf

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}\, I_{(0,\infty)}(x)$$

where $\alpha > 0$ and $\beta > 0$. Also, $\text{E}(X) = \alpha\beta$ and $\text{Var}(X) = \alpha\beta^2$. The mgf is given by $m_X(t) = (1-\beta t)^{-\alpha}$ for $t < 1/\beta$.

Inverse Gamma Density: $X \sim \text{IG}(\alpha, \beta)$ means X has pdf

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, \frac{1}{x^{\alpha+1}}\, e^{-1/x\beta}\, I_{(0,\infty)}(x)$$

where $\alpha > 0$ and $\beta > 0$.

Iterated Expectation Formula: $\text{E}(X) = \text{E}\left[\text{E}(X|Y)\right]$.

Iterated Variance Formula: $\text{Var}(X) = \text{E}\left[\text{Var}(X|Y)\right] + \text{Var}\left[\text{E}(X|Y)\right]$.

Distributional Result: If $X \sim \text{Gamma}(\alpha_x, \beta)$ and $Y \sim \text{Gamma}(\alpha_y, \beta)$ and X and Y are independent, then $X/(X+Y) \sim \text{Beta}(\alpha_x, \alpha_y)$.

1. Suppose n data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are observed.

(a) Write the usual model equation for the simple linear regression of the y_i values on the x_i values, under the assumption of independent, identically distributed, normal errors. State any conditions that the terms must satisfy.

(b) *Derive* scalar formulas (in terms of the data) for the normal equations and the ordinary least squares (OLS) estimates of all mean-related parameters in your model from part (a), assuming that the estimates are unique. State the condition(s) that the data must satisfy for the estimates to be unique. Show *all* steps, beginning with the definition of OLS estimates.

(c) *Derive* expressions for the expected value and variance of the OLS estimate for the *slope* parameter from part (b), assuming the data obey the model of part (a).

2. Three different movie trailers (1, 2, and 3) for a new (unreleased) movie are being evaluated for their ability to draw audiences. Each of 15 separate audiences (of approximately the same size) is shown one of the trailers, such that each trailer is shown to five audiences. Afterwards, each member of each audience is given a free ticket to a special pre-screening of the new movie. (Each ticket is secretly marked so that the audience to which it was distributed can be identified.) You are given the following table that shows the percentage of each audience who actually attend the pre-screening of the new movie:

	1	2	3	4	5
Trailer 1:	30	25	40	20	10
Trailer 2:	45	35	60	20	40
Trailer 3:	45	45	50	35	25

Row 1 gives the percentage attendance from each of the five audiences who viewed Trailer 1, and so forth. You ask a person conducting the experiment to describe how the trailers were assigned to audiences.

(a) "The assignment of trailers to audiences was completely at random, subject to five different audiences being assigned to each trailer. The column numbers above the table are completely arbitrary."

(i) Name the *design* of this experiment.

(ii) Write the usual normal-theory model equation for analyzing the data from this experiment, based on the design you named in part (a). State the conditions on the terms in the equation, including any constraints necessary to make all parameters well-defined.

(iii) Compute an analysis of variance (relevant sources, degrees of freedom, sums of squares, and mean squares).

(iv) Based on your model, perform a test of whether the type of trailer shown has any effect on the mean percentage who attend the pre-screening ($\alpha = 0.05$).

(b) Later, the person who conducted the experiment says, "Sorry, the description I gave earlier was incorrect. The five columns of the table actually represent five different days. There were three audiences per day, and the three trailers were randomly assigned to these three audiences, independently for each day." Redo every subpart of part (a) based on this new information. (You may reuse results of computations that you have already done in part (a), if appropriate.)

3. Suppose that a data set of triples $(x_{i1}, x_{i2}, y_i), i = 1, \ldots, n$ is analyzed using the following linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad \epsilon_i \sim \text{iid N}(0, \sigma^2) \qquad i = 1, \ldots, n, \tag{1}$$

and that the unique ordinary least squares estimates of β_0, β_1, and β_2 are

$$\hat{\beta}_0 = 10 \qquad \hat{\beta}_1 = -8 \qquad \hat{\beta}_2 = 5$$

Also, letting $\boldsymbol{X}$ be the $n \times 3$ matrix whose i^{th} row is $[\ 1 \ \ x_{i1} \ \ x_{i2}\ ]$, suppose that

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 15 & 10 & -11 \\ 10 & 20 & -10 \\ -11 & -10 & 47 \end{bmatrix} \qquad \text{and} \qquad \sum_{i=1}^{n} y_i^2 = 4755$$

(a) Find the following values:

$$n, \qquad \sum_{i=1}^{n} y_i, \qquad \sum_{i=1}^{n} x_{i1}\, y_i, \qquad \text{and} \qquad \sum_{i=1}^{n} x_{i2}\, y_i$$

(b) Compute the residual (error) sum of squares.

(c) Consider the reduced model

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \qquad i = 1, \ldots, n$$

Compute the ordinary least squares estimates of β_0 and β_1 and the residual (error) sum of squares for this model.

(d) For model (1), compute the *sequential* sums of squares (also called the "Type I" sums of squares, or the sums of squares in the "sequential ANOVA") in the order $\beta_0, \beta_1, \beta_2$ (intercept, x_{i1}, x_{i2}).

(e) Assuming that model (1) is correct, test the following hypotheses (separately) at level $\alpha = 0.05$:

(i) $\text{H}_0 : \ \beta_2 = 0 \quad \text{versus} \quad \text{H}_a : \ \beta_2 \neq 0$

(ii) $\text{H}_0 : \ \beta_1 = \beta_2 = 0 \quad \text{versus} \quad \text{H}_a : \ \beta_1 \neq 0$ and/or $\beta_2 \neq 0$

4. Consider the following two-factor random model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a \quad j = 1, \ldots, b \quad k = 1, \ldots, n$$
$$\alpha_i \sim \mathrm{N}(0, \sigma^2_\alpha), \quad \beta_j \sim \mathrm{N}(0, \sigma^2_\beta), \quad \alpha\beta_{ij} \sim \mathrm{N}(0, \sigma^2_{\alpha\beta}), \quad \epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$$
$$\alpha_i,\ \beta_j,\ \alpha\beta_{ij},\ \epsilon_{ijk} \text{ independent for all } i, j, k$$

where $a \geq 2$, $b \geq 2$, and $n \geq 2$. The effects α_i are the main effects of Factor A, the effects β_j are the main effects of Factor B, and the effects $\alpha\beta_{ij}$ are the AB interaction effects.

(a) Would you call this a *balanced* data model? Explain briefly.

(b) Write out the variance-covariance matrix of the vector

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{122} \\ y_{222} \end{bmatrix}$$

in terms of the parameters.

(c) Give an unbiased estimator of μ that uses *all* of the data values. Find the variance of this estimator (in terms of the parameters).

(d) Write expressions for the *mean squares* for Factor A, for Factor B, for AB interaction, and for error (residual), in terms of the y_{ijk} values. Remember to define any special notation that you use.

(e) Describe how you would perform a level α test for the presence of the main effect term for Factor A, using mean squares from the previous part. Remember to state the null and alternative hypotheses.

5. The amount of carbohydrates in public school lunches may affect the performance of students on standardized exams. To test this, 30 public school students from the same grade are completely randomized into two groups of 15 students each. On the exam days, students in group 1 are given an ordinary lunch, and students in group 2 are given a special lunch with higher carbohydrate content. Let y_{ij} be the overall score on after-lunch exams of the j^{th} student in group i. The overall score x_{ij} of the j^{th} student in group i on similar exams from an earlier grade is taken as a covariate. Fitting the analysis of covariance (ANCOVA) model

$$y_{ij} \;=\; \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij} \;=\; \mu_i + \beta x_{ij} + \epsilon_{ij}, \qquad i = 1, 2 \quad j = 1, \ldots, 15$$

$$\sum_{i=1}^{2} \alpha_i = 0 \qquad \epsilon_{ij} \sim \text{iid N}(0, \sigma^2)$$

produces unique ordinary least squares estimates $\hat{\mu}_1 = 13.41$, $\hat{\mu}_2 = 55.10$, and $\hat{\beta} = 0.873$, whose usual unbiased variance and covariance estimates are

$$\widehat{\text{Var}}(\hat{\mu}_1) = 928.15 \qquad \widehat{\text{Var}}(\hat{\mu}_2) = 1314.17 \qquad \widehat{\text{Var}}(\hat{\beta}) = 0.00287$$

$$\widehat{\text{Cov}}(\hat{\mu}_1, \hat{\mu}_2) = 1002.40 \qquad \widehat{\text{Cov}}(\hat{\mu}_1, \hat{\beta}) = -1.541 \qquad \widehat{\text{Cov}}(\hat{\mu}_2, \hat{\beta}) = -1.866$$

(a) Find the least squares estimates of μ, α_1, and α_2.

(b) Compute the covariate-adjusted mean overall score estimates for the two groups, adjusting to the average covariate value of 594. Also compute unbiased estimates of the variances of these mean estimates (conditional on the covariate values).

(c) Form a 95% two-sided confidence interval for the difference in (covariate-adjusted) mean overall score between group 1 and group 2. Perform a test of the null hypothesis that the carbohydrate content of the lunch has no effect, based on this confidence interval ($\alpha = 0.05$).

(d) Would the test from the previous part be equivalent to the usual ANCOVA F-test for treatment effects in this two-group experiment? Explain your answer in terms of full and reduced models.

(e) Suppose we were to perform the same experiment in a completely randomized design *without* a covariate. How would this change the way we randomize students to groups?

6. A penny and dime are tossed. Let X denote the total number of heads up. Then the penny is tossed again. Let Y denote the total number of heads up on the dime (from the first toss) plus the penny from the second toss.

(a) Find the joint pmf of X and Y, and hence compute the marginal pmf's of X and Y. Are X and Y independent?

(b) Find the conditional distribution of Y given $X = 1$.

(c) Compute the correlation between X and Y.

7. The following lottery game is played by $N + 1$ people, among them Alex. Each person randomly and independently selects one of b available boxes in which to place a card with their name written on it. The lottery official then randomly selects a box (the winning box) and picks a card from it. The person whose name appears on the card is the winner. (If there are no cards in the winning box, then there is no winner.) Define the events: $A = \{\text{Alex is the winner}\}$, and $B = \{\text{Alex selects the winning box}\}$. The first part of this question involves a general result, and the remaining parts concern the calculation of $P(A)$.

(a) Derive the Law of Total Probability. More specifically, suppose that S is a sample space, C is a subset of S and that $\{E_1, E_2, E_3, \dots\}$ is a partition of S and show that

$$P(C) = \sum_{i=1}^{\infty} P(C|E_i)\, P(E_i)\ .$$

Note that this law remains valid for calculating conditional probabilities. That is, if D is another subset of S, then

$$P(C|D) = \sum_{i=1}^{\infty} P(C|E_i, D)\, P(E_i|D)\ .$$

(You do not have to prove this.)

(b) In the lottery problem described above, find $P(A)$ when $b = 1$.

(c) Assume $b \geq 2$. Show that $P(A) = P(A|B)/b$.

(d) Assume $b \geq 2$. Define F_j, $j = 0, \dots, N$, to be the event that j people (other than Alex) place their card in the winning box. By conditioning, or otherwise, find an expression for $P(A|B)$.

(e) Putting together the results from (c) and (d), or otherwise, write down an expression for the probability that Alex is the winner when more than one box is available.

8. (a) Suppose that $X \sim f(x|\theta)$ and that $\tilde{\theta}(X)$ is an unbiased estimator of θ. Suppose further that $\hat{\theta}(X)$ is another estimator such that $\hat{\theta}(X) = c\,\tilde{\theta}(X)$ where c is a constant. Show that the MSE of $\hat{\theta}(X)$ can be written as

$$c^2 \operatorname{Var}\big(\tilde{\theta}(X)\big) + \theta^2(c-1)^2 \ ,$$

and use this to prove that, if $c > 1$, then one of the two estimators dominates the other in terms of MSE.

(b) Suppose that $Y \sim \text{IG}(\alpha, \beta)$. Find a formula for EY^p where p is any real number. Does your formula hold for all $p \in \mathbb{R}$?

(c) Now suppose that $X|Y = y \sim \text{Gamma}(4, y)$ and that $Y \sim \text{IG}(\alpha, \beta)$ with $\alpha > 1$. Show that the marginal mass function of X is given by

$$f(x; \alpha, \beta) = \frac{x^3 \beta^4 \Gamma(\alpha + 4)}{6\Gamma(\alpha)} \left(\frac{1}{x\beta + 1}\right)^{\alpha+4} I_{\mathbb{R}^+}(x) \ . \tag{2}$$

(d) Find the *marginal* mean of X.

(e) Suppose that X is a single observation from (2) with α fixed and known. Find the ML estimators of β and $1/\beta$. Are these estimators unbiased?

(f) Find an unbiased estimator of $1/\beta$ that dominates the ML estimator of $1/\beta$ in terms of MSE.

9. Consider the pdf given by

$$f(x|\theta) = \frac{3\theta^3}{(x+\theta)^4} I_{\mathbb{R}^+}(x) \ ,$$

where $\theta > 0$.

(a) Let $\overline{F}(x|\theta) = P_\theta(X > x)$. Find $\overline{F}(x|\theta)$ and show that it is a monotone function of θ.

(b) Suppose X is a single observation from $f(x|\theta)$, θ is unknown, and that we wish to test $H_0 : \theta = 1$ versus $H_1 : \theta = 2$. Find the level 0.10 NP test and let R denote its rejection region.

(c) Now suppose that we are interested in testing $H_0' : \theta \leq 1$ against $H_1' : \theta > 1$. Consider using the rejection region R from part (b) to test H_0' versus H_1'. Find the power function of this test and use it to calculate the size of the test. Denote the size by α'.

(d) *Without appealing to MLR*, explain why the test with rejection region R is a UMP level α' test of H_0' versus H_1'.

10. Suppose that $X_1, \ldots, X_n$ are iid Beta$(\mu, 1)$ and that $Y_1, \ldots, Y_m$ are iid Beta$(\theta, 1)$. As usual, let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$. Assume further that X and Y are independent. We will consider testing $H_0 : \mu = \theta$ versus $H_1 : \mu \neq \theta$ using the statistic

$$T = \frac{\sum_{i=1}^n \log X_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i} \, .$$

(a) *Derive* the distribution of $Z = -\mu \log X_1$.

(b) *Identify* the distribution of T under H_0.

(c) Show that the ML estimator of μ is $\hat{\mu}(X)$ where

$$\hat{\mu}(x) = \frac{-n}{\sum_{i=1}^n \log x_i} = \frac{-n}{\log \prod_{i=1}^n x_i} \, .$$

Obviously, the ML estimator of θ is $\hat{\theta}(Y)$ where $\hat{\theta}(y) = -m / \sum_{i=1}^m \log y_i$.

(d) Define

$$\hat{\mu}_0(x, y) = \frac{-(n+m)}{\sum_{i=1}^n \log x_i + \sum_{i=1}^m \log y_i} \, .$$

Simplify the following expression

$$\frac{\left[\left(\prod_{i=1}^n x_i \right) \left(\prod_{i=1}^m y_i \right) \right]^{\hat{\mu}_0(x,y)-1}}{\left(\prod_{i=1}^n x_i \right)^{\hat{\mu}(x)-1} \left(\prod_{i=1}^m y_i \right)^{\hat{\theta}(y)-1}} \, .$$

(e) Show that the LRT statistic for testing H_0 versus H_1 can be written in such a way that it involves the data only through the statistic T.

(f) Give the rejection region of the size 0.10 LRT in terms of T. You may leave your answer in the form of an unsolved equation (or equations).