First Year Examination
Department of Statistics, University of Florida
August 18, 2006, 8:00 am - 12:00 noon

**Instructions:**

1. Write your **number** on every page that you plan to submit.
2. **Do not** write your name anywhere on any of the pages that you plan to submit.
3. You have four hours to answer questions in this examination.
4. You must show your work to receive credit.
5. **Write only on one side of the paper, and start each question on a new page.**
6. There are 10 problems of which you must answer 8.
7. Only your first 8 problems will be graded.
8. While the 10 questions are equally weighted, some problems are more difficult than others.
9. The parts within a given question are not necessarily equally weighted.
10. 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
- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- MSE = mean squared error
- ML = maximum likelihood
- pdf = probability density function
- $\alpha$ = specified probability of Type I error
- $\mathrm{N}(\mu, \sigma^2)$ = normal distribution with mean $\mu$ and variance $\sigma^2$

You may use the following facts/formulas without proof:

**Fact about mgfs:** If $X$ has mgf $M_X(t)$ and $a$ and $b$ are constants, then the mgf of $aX + b$ is $e^{bt}M_X(at)$.

**Linear Combinations of Independent Normals:** Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$. If $a_1, \ldots, a_n$ are constants, then the random variable $\sum_{i=1}^n a_i X_i$ has a normal distribution.

**Gamma Density:** $X \sim \mathrm{Gamma}(\alpha, \beta)$ 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 $\beta > 0$.

**Students $t$ Density:** $X \sim t_\nu$ means $X$ has pdf

$$f(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}$$

where $\nu > 0$.

**Poisson moments:** If $X \sim \mathrm{Poisson}(\lambda)$, then $\mathrm{E}X = \mathrm{Var}X = \lambda$.

**1.** Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ is the $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times (p+1)$ matrix with full column rank and $\boldsymbol{\beta}$ is the vector of regression parameters. Suppose, contrary to the usual least squares assumptions, that the error vector $\boldsymbol{\epsilon}$ has mean vector zero and variance-covariance matrix $\sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known matrix not equal to the identity matrix.

(a) Find the mean vector and variance-covariance matrix of the ordinary least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

(b) Suppose $\boldsymbol{V} = \boldsymbol{TT}'$, where $\boldsymbol{T}$ is invertible. Find a square matrix $\boldsymbol{A}$ such that if $\boldsymbol{Y}^* = \boldsymbol{AY}$, $\boldsymbol{X}^* = \boldsymbol{AX}$ and $\boldsymbol{\epsilon}^* = \boldsymbol{A\epsilon}$ then the transformed model $\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ satisfies the ordinary least squares assumptions.

(c) Write the ordinary least squares estimator of $\boldsymbol{\beta}$ for the *transformed* model of the previous part in terms of $\boldsymbol{X}$, $\boldsymbol{V}$, and $\boldsymbol{Y}$. Also find its mean vector and variance-covariance matrix.

(d) Let $\boldsymbol{e}^*$ be the vector of ordinary least squares residuals for the *transformed* model (based on the estimator of the previous part). Find the matrix $\boldsymbol{P}^*$ such that the variance-covariance matrix of $\boldsymbol{e}^*$ is $\sigma^2\boldsymbol{P}^*$. Show that $\boldsymbol{P}^*$ is a projection matrix.

(e) Find an unbiased estimator of $\sigma^2$.

**2.** 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 $\epsilon$ are independent and identically distributed with mean zero and the same variance.

(a) Estimate the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

(b) 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 $\widehat{Y}$.

(c) Test the hypothesis that $\beta_1$ and $\beta_2$ 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 $\alpha = 0.05$.

(d) Compute 95% simultaneous two-sided confidence intervals for $\beta_0$, $\beta_1$, and $\beta_2$, using the Bonferroni method.

**3.** 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:

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

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

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

(a) For each model, write out the null hypothesis and the alternative hypothesis for the test of whether or not there are any factor effects.

(b) In terms of the data values $y_{ij}$, write out the sum of squares for factor effect, $\mathrm{SS(Factor)}$, and the sum of squares for error, $\mathrm{SS(Error)}$. Also give expressions for their corresponding degrees of freedom.

(c) Write an expression for the $F$-statistic (in terms of $\mathrm{SS(Factor)}$ and $\mathrm{SS(Error)}$) for testing the hypotheses in part (a). What is its distribution under each null hypothesis of part (a)?

(d) For each model, find the *correlation* between two different responses that have the same treatment level.

**4.** A balanced two-factor experiment yields responses $y_{ijk}$ for replication $k$ at level $i$ of Factor A and level $j$ of Factor B, for $i = 1, 2,\ j = 1, 2,\ k = 1, 2, 3, 4$. The data are analyzed using the model equation

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk},$$

under the conditions

$$\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 0, \quad \alpha\beta_{11} + \alpha\beta_{12} = \alpha\beta_{21} + \alpha\beta_{22} = \alpha\beta_{11} + \alpha\beta_{21} = \alpha\beta_{12} + \alpha\beta_{22} = 0$$

and the terms $\epsilon_{ijk}$ are iid $\mathrm{N}(0, \sigma^2)$.

Suppose the least squares estimates of the model parameters are

$$\widehat{\mu} = 10 \quad \widehat{\alpha}_1 = 6 \quad \widehat{\alpha}_2 = ? \quad \widehat{\beta}_1 = 2 \quad \widehat{\beta}_2 = ? \quad \widehat{\alpha\beta}_{11} = 1 \quad \widehat{\alpha\beta}_{12} = ? \quad \widehat{\alpha\beta}_{21} = ? \quad \widehat{\alpha\beta}_{22} = ?$$

and the sum of squares for error is 100.

(a) Find $\widehat{\alpha}_2$, $\widehat{\beta}_2$, $\widehat{\alpha\beta}_{12}$, $\widehat{\alpha\beta}_{21}$, and $\widehat{\alpha\beta}_{22}$.

(b) Write out an ANOVA table that includes the sum of squares and degrees of freedom for all sources (Factor A, Factor B, Interaction, and Error).

(c) Test whether there is any interaction between factors A and B. Test whether there is a main effect of Factor A. Test whether there is a main effect of Factor B. Perform all tests individually at level $\alpha = 0.05$.

(d) Is $\mu + \alpha_1$ a contrast? (Answer yes or no.) Give an unbiased estimate of $\mu + \alpha_1$ and an unbiased estimate of the variance of your estimate.

(e) Is $\alpha_1 - \alpha_2$ a contrast? (Answer yes or no.) Give an unbiased estimate of $\alpha_1 - \alpha_2$ and an unbiased estimate of the variance of your estimate.

**5.** An experiment is conducted in a completely randomized design with three treatment groups. In addition to the measured response $Y$, there is also a measured covariate $X$. The data are as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| Response $Y$ | 10 | 14 | 2 | 7 | 5 | 4 |
| Treatment Group | 1 | 1 | 2 | 2 | 3 | 3 |
| Covariate $X$ | -1 | 2 | -2 | 1 | 0 | 0 |

(a) Estimate the coefficients in the simple linear regression of $Y$ on $X$, *completely ignoring* the treatment group. Also find the residual (error) sum of squares.

(b) Perform an $F$-test for treatment effects using a one-way ANOVA model that *completely ignores* the covariate $X$. (Use $\alpha = 0.05$.)

(c) If both the treatment effects and a linear term in $X$ are included in the model for the response $Y$, the residual (error) sum of squares is $0.75$. Perform an analysis-of-covariance $F$-test for treatment effects (i.e. a test for treatment effects after adjusting for a linear effect in the covariate $X$.) (Use $\alpha = 0.05$.)

**6.** Toss $n$ fair dice. Pick up only those *not* showing a 6 and toss them. Continue doing this until all dice show a 6. Let $X$ denote the total number of tosses made in this experiment.

(a) How many tosses do you expect to make if $n = 1$?

(b) Compute the cdf of $X$ for general $n$. (Hint: Think about what happens with each individual die.)

(c) Suppose that $Y$ is a discrete random variable whose support is contained in $\{0, 1, 2, \dots\}$. Show that

$$\sum_{y=0}^{\infty} \big[1 - F_Y(y)\big] = \mathrm{E}Y \ ,$$

where $F_Y(\cdot)$ is the cdf of $Y$. (Hint: A double sum will help.)

(d) Use the result from (c) to find the expected value of $X$ when $n = 2$.

**7.** Suppose that $Y|Z = z \sim \text{N}(\mu, 1/z)$ and that $Z \sim \text{Gamma}(\alpha/2, \beta)$ where $\alpha, \beta > 0$.

(a) Find the marginal pdf of $Y$.

(b) Show that, in general, for any two random variables $S$ and $T$

$$\text{Var}(S) = \text{E}\left[\text{Var}(S|T)\right] + \text{Var}\left[\text{E}(S|T)\right] \ ,$$

provided the expectations exist.

(c) Use the result in part (b) to find the marginal variance of $Y$.

(d) Suppose that $T \sim t_\alpha$. Find a function $g$ such that if $W = g(\mu, \alpha, \beta, T)$, then $W$ and $Y$ have the same distribution. Note that $\mu$, $\alpha$ and $\beta$ are constant.

**8.** Let $X_1, \ldots, X_n$ be iid random variables from a distribution with a finite second moment. Let $\text{E}X_1 = \mu$, $\text{Var}X_1 = \sigma^2$, $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Show that

(a) $\text{E}(\bar{X}) = \mu$

(b) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

(c) $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right]$

(d) $\text{E}(S^2) = \sigma^2$

For the remainder of this problem, assume $X_1, \ldots, X_n$ are iid Poisson$(\lambda)$ and consider estimating $\lambda$ with

$$\delta_w(X_1, \ldots, X_n) = w\bar{X} + (1-w)S^2 \ ,$$

where $w \in [0, 1]$.

(e) Show that $\delta_w$ is an unbiased estimator of $\lambda$.

(f) Find the Cramér-Rao lower bound for the variance of an unbiased estimator of $\lambda$.

(g) How many values of $w \in [0, 1]$ yield estimators that attain the Cramér-Rao lower bound? (Warning: A correct answer with nothing to back it up is worth 0 points.)

(h) Find

$$\text{E}\left[\delta_w(X_1, \ldots, X_n)|\bar{X}\right] \ .$$

**9.** Suppose that $X \in \mathcal{X}$ is a random vector with pdf $f(x|\theta)$, $\theta \in \Theta$. Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_0^c$, where $\Theta_0^c$ denotes the complement of $\Theta_0^c$; that is, $\Theta_0^c = \Theta \setminus \Theta_0$.

(a) Fix $c \in (0, 1)$. A LRT has rejection region

$$R_1 = \left\{ x \in \mathcal{X} : \lambda(x) < c \right\} ,$$

where $\lambda(x)$ is the LRT statistic. Write down the definition of $\lambda(x)$.

Now suppose that $\Theta_0 =$