First Year Examination
Department of Statistics, University of Florida
August 18, 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
- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- MOM = method of moments
- ML = maximum likelihood
- pdf = probability density function
- pmf = probability mass function
- α = specified probability of Type I error (in the context of hypothesis testing)
- $\mathbb{R} = (-\infty, \infty)$
- N(μ, σ^2) = normal distribution with mean μ and variance σ^2
- UMVUE = uniformly minimum variance unbiased estimator
- $t_{\nu,\alpha}$ is the number such that $P(T > t_{\nu,\alpha}) = \alpha$ when T is a Student's t random variable with ν degrees of freedom

1. Let X and Y be two random variables with finite second moments and recall that the covariance between X and Y is defined as $\text{Cov}(X,Y) = \text{E}\big[\big(X - \text{E}X\big)\big(Y - \text{E}Y\big)\big]$.

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

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

Now suppose that $P(Y \geq 0) = 1$ and let $\text{E}Y = \mu$. Let $\big(X_0, Y_0\big)$ be an iid copy of (X,Y). In what remains of this problem, you will show that

$$\big[\text{Cov}(X,Y)\big]^2 \leq \mu\,\text{E}\left[\big(X - X_0\big)^2\Big(\sqrt{Y} - \sqrt{Y_0}\Big)^2\right] . \tag{1}$$

(c) Show that

$$\text{Cov}(X,Y) = \frac{1}{2}\,\text{E}\Big[\big(X - X_0\big)\big(Y - Y_0\big)\Big] .$$

(d) State the Cauchy-Schwarz inequality.

(e) Use Cauchy-Schwarz and the fact that $(Y - Y_0) = \Big(\sqrt{Y} - \sqrt{Y_0}\Big)\Big(\sqrt{Y} + \sqrt{Y_0}\Big)$ to establish that

$$\Big\{\text{E}\big[\big(X - X_0\big)\big(Y - Y_0\big)\big]\Big\}^2 \leq \text{E}\left[\Big(\sqrt{Y} + \sqrt{Y_0}\Big)^2\right]\,\text{E}\left[\big(X - X_0\big)^2\Big(\sqrt{Y} - \sqrt{Y_0}\Big)^2\right] .$$

(f) Show that for positive numbers a and b,

$$\Big(\sqrt{a} + \sqrt{b}\Big)^2 \leq 2(a+b) .$$

(g) Put the results in (c), (e) and (f) together to establish (1).

2. This question deals with the concept of completeness and its applications.

(a) Consider a family of pdfs (or pmfs) $\big\{f(t|\theta) : \theta \in \Theta\big\}$. Explain exactly what it means to say that this family is *complete*.

(b) Suppose that $f(t|\theta) = \theta^{-1} I_{(0,\theta)}(t)$ and $\Theta = \{1, 2, 3\}$. Is this family complete? (Remember to **show your work**!)

(c) Let $X_1, \ldots, X_n$ be iid Uniform$(0,\theta)$ where $\theta \in (0,\infty)$ is unknown. Find the best unbiased estimator of θ; that is, the UMVUE of θ.

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

$$f(x|\theta) = \begin{cases} \frac{\gamma x^{\gamma-1}}{\theta^\gamma} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma > 0$ is known and $\theta > 0$ is unknown.

(a) Is θ a location parameter, a scale parameter, or neither of these? Explain.

(b) Find a sufficient statistic for θ.

(c) Find the ML estimator of θ.

(d) Find the MOM estimator of θ.

(e) Let W be a random variable with a Pareto(α, β) distribution; that is, W has pdf given by

$$f(w|\alpha, \beta) = \frac{\beta\alpha^\beta I_{(\alpha,\infty)}(w)}{w^{\beta+1}} ,$$

where $\alpha, \beta > 0$. Find the expected value of W. (Hint: Does it always exist?)

(f) Suppose you are a Bayesian and that your prior distribution for θ is Pareto(α, β). Find the posterior density of θ.

(g) Find the posterior expectation of θ given the data.

4. Let $X_1, \ldots, X_n$ be iid N(μ, σ^2) with both parameters unknown. As usual, let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

(a) Derive the ML estimator of (μ, σ^2).

(b) Define the hypothesis testing terms *size* and *level*.

(c) Consider testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ where $\mu_0 \in \mathbb{R}$ is fixed. Show that the test that rejects H_0 when

$$\bar{X} > \mu_0 + t_{n-1,\alpha}\sqrt{S^2/n}$$

is a size α test.

(d) Show that the test in (c) is the LRT; that is, show that the LRT rejects if and only if

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > c .$$

5. (a) Derive the mgf of a standard normal random variable.

(b) Let Z be a random variable with mgf $M_Z(t)$ and let a and b be constants. Show that the mgf of $aZ+b$ is $e^{tb}M_Z(at)$. Use this result to write down the mgf of a $\mathrm{N}(\mu, \sigma^2)$ random variable.

(c) Suppose that $W_1, \ldots, W_m$ are independent random variables such that $W_i \sim \mathrm{N}(\mu_i, \sigma^2)$. Find the distribution of $\sum_{i=1}^n a_i W_i$ where $a_1, \ldots, a_m$ are known constants.

(d) Suppose that the random variables $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \beta x_i + \varepsilon_i \ ,$$

for $i = 1, \ldots, n$ where $x_1, \ldots, x_n$ are known constants and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathrm{N}(0, \tau^2)$. Find a two-dimensional sufficient statistic for the unknown parameter (β, τ^2).

(e) Find the MLE of β, call it $\hat{\beta} = \hat{\beta}(Y_1, \ldots, Y_n)$. Show that $\hat{\beta}$ is unbiased.

(f) Find the distribution of $\hat{\beta}$.

(g) Find the distribution of the alternative estimator of β given by

$$\tilde{\beta} = \tilde{\beta}(Y_1, \ldots, Y_n) = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}$$

(h) Which of these estimators is better?

6. To investigate whether premiums differ significantly between three particular auto insurance companies, four auto insurance buyers get quotes from each of the three companies. The quoted premiums ($) for each buyer/company combination are shown in the following table:

Buyer	Company 1	Company 2	Company 3
1	250	300	350
2	325	350	375
3	375	400	425
4	350	350	350

Regard the buyers as blocks and the three companies as the treatments to be investigated. (We are only interested in these three particular companies.)

(a) Compute the sum of squares for treatment (companies) and give its degrees of freedom.

(b) Compute the sum of squares for blocks (buyers) and give its degrees of freedom.

(c) Compute the sum of squares for error (interaction) and give its degrees of freedom.

(d) Test whether the three companies differ significantly in the premiums they quote ($\alpha = 0.05$).

(e) Form a two-sided 95% confidence interval for the average difference in quoted premium between company 1 and company 2.

7. Two statistical analysts are given the same data for a single response variable Y, and a single independent variable (X). 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	20	22	18
2	34	36	35
3	36	44	40
4	48	45	42

(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 tests of the hypotheses from part (e), each at $\alpha = 0.05$. (There is no need to adjust for simultaneous testing because John and Jane are working separately.)

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

8. 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)

(a) Derive the mean vector and variance-covariance matrix of the least squares estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} .$$

(b) If $\boldsymbol{k}$ is a constant vector of the same size as $\widehat{\boldsymbol{\beta}}$, what is the distribution of $\boldsymbol{k}'\widehat{\boldsymbol{\beta}}$?

(c) Define $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ to be the vector of residuals. Find a way to write $\boldsymbol{e}'\boldsymbol{e}$ that uses $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ but does not otherwise use $\boldsymbol{X}$.

(d) Find $E(\boldsymbol{y}'\boldsymbol{y})$ in terms of $\boldsymbol{X}$, $\boldsymbol{\beta}$, and σ^2. (Substitute $\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ for $\boldsymbol{y}$ and simplify.)

9. The average yield of a batch of chemicals produced in a chemical production process varies smoothly as a function of the amount of a reactant. Fitting a quadratic polynomial to 20 observations for batch yield (y) as a function of reactant amount (r) over a range of approximately 0.2 to 0.6 produces the fitted model equation

$$\begin{array}{ccccc} \widehat{y} & = & 75.76 & +\ 101.09\,r & -\ 106.09\,r^2 \\ & & (3.23) & (16.30) & (19.45) \end{array}$$

with the usual estimated *standard errors* of the regression parameter estimates listed underneath in parentheses. The sum of the squared residuals is 15.22.

(a) Estimate the amount of reactant that will produce the greatest yield.

(b) Estimate the mean and variance of the yield when the amount of reactant is 0.4.

(c) Test whether the quadratic term r^2 can be dropped from the model ($\alpha = 0.05$). Give the null and alternative hypotheses, test statistic, and critical value.

(d) Fitting a cubic model (with an additional term r^3) results in a sum of squared residuals of 13.59. Test whether the cubic term is statistically necessary ($\alpha = 0.05$).

10. A balanced 2-factor design with one random factor A and one fixed factor B has the model equation

$$y_{ijk} \;=\; \mu \;+\; a_i \;+\; \beta_j \;+\; (ab)_{ij} \;+\; e_{ijk}, \qquad i = 1, \ldots, a, \quad j = 1, \ldots, b, \quad k = 1, \ldots, r$$

where the a_i's are the factor A effects, the β_j's are the factor B effects, the $(ab)_{ij}$'s are AB interaction effects, and the e_{ijk}'s are error terms. You may assume (as usual)

$$E(a_i) \;=\; E((ab)_{ij}) \;=\; E(e_{ijk}) \;=\; 0, \quad \sum_{j=1}^{b} \beta_j \;=\; 0,$$
$$\text{Var}(a_i) \;=\; \sigma_a^2, \quad \text{Var}((ab)_{ij}) \;=\; \sigma_{ab}^2, \quad \text{Var}(e_{ijk}) \;=\; \sigma^2$$

and all of the a_i's, $(ab)_{ij}$'s, and e_{ijk}'s are independent. Suppose an experiment based on this design results in the following analysis of variance table:

Source	Deg. of Freedom	Sum of Squares	Expected Mean Square
Factor A	1	156	$\sigma^2 \;+\; 2\sigma_{ab}^2 \;+\; 8\sigma_a^2$
Factor B	3	360	$\sigma^2 \;+\; 2\sigma_{ab}^2 \;+\; 4\theta_b^2$
Interaction AB	3	45	$\sigma^2 \;+\; 2\sigma_{ab}^2$
Error	8	60	σ^2
			$\theta_b^2 \;=\; \sum_{j=1}^{b} \beta_j^2/(b-1)$

(a) Derive the mean and variance of y_{ijk}.

(b) Determine a = number of levels of factor A, b = number of levels of factor B, and r = number of replications at each combination of levels.

(c) Test for the presence of the interaction term ($\alpha = 0.05$). State the null and alternative hypotheses and give the test statistic and critical value.

(d) Test for factor A effect ($\alpha = 0.05$). State the null and alternative hypotheses and give the test statistic and critical value.

(e) Test for factor B effect ($\alpha = 0.05$). State the null and alternative hypotheses and give the test statistic and critical value.

(f) Give an unbiased ANOVA estimate of σ_{ab}^2.