First Year Examination
Department of Statistics, University of Florida
August 20, 2009, 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. Questions 1 through 5 are the "theory" questions and questions 6 through 10 are the "applied" questions. You must do exactly four of the theory questions and exactly four of the applied questions.
5. While the 10 questions are equally weighted, some questions 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.

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
- ML = maximum likelihood
- OLS = ordinary least squares
- pdf = probability density function
- pmf = probability mass function
- $\mathbb{N} = \{1, 2, 3, \dots\}$

You may use the following facts/formulas without proof:

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 $\beta > 0$.

Normal density: $X \sim \text{N}(\mu, \sigma^2)$ means X has pdf

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\}$$

where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

1. Consider an experiment involving four independent, identical trials. Each trial entails randomly choosing a day of the week; that is, randomly choosing a member of the set

$$\{\text{Sunday}, \text{Monday}, \text{Tuesday}, \ldots, \text{Saturday}\} \ .$$

Let X denote the number of different days in the resulting sequence. For example, if we get "Saturday" all four times, then $X = 1$. As another example, if the first two draws result in "Monday," the third results in "Sunday," and the fourth yields "Friday," then $X = 3$.

(a) Find $\Pr(X = 2)$.

(b) Find $\mathrm{E}(X)$.

(c) Let Z denote the number of occurrences of Saturday and Sunday in the sequence. Find the conditional pmf of X *given* that $Z = 0$.

(Express your answers as ratios of integers or sums of ratios of integers - do not use any decimals!)

2. Let $X_1, X_2, \ldots, X_n$ be iid with common pdf given by

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

where $\beta > 0$ and $\theta \in \mathbb{R}$. Write the order statistics as $X_{(1:n)}, X_{(2:n)}, \ldots, X_{(n:n)}$. (We're using this notation instead of the more standard notation, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, in order to make the dependence on n more explicit.)

(a) Derive the pdf of $X_{(1:n)}$.

(b) Define $Y_n = X_{(1:n)}$ for $n \in \mathbb{N}$. Prove or disprove the following statement: The sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ converges in probability.

(c) Derive the cdf of $X_{(n:n)}$.

(d) Define $Z_n = X_{(n:n)}$ for $n \in \mathbb{N}$. Find a non-random sequence $\{a_n\}_{n=0}^{\infty}$ such that $Z_n - a_n$ converges in distribution. (Hint: Is the limit a valid cdf?)

3. In this question, we will develop a method of simulating from the Gamma$(3.5, 3.5)$ distribution using only iid Uniform$(0, 1)$ random variables.

(a) Let $U \sim$ Uniform$(0, 1)$ and let θ be a fixed, positive constant. Derive the distribution of $Z = -\theta \log U$.

(b) Suppose that $X_1, \ldots, X_n$ are iid Gamma$(1, \beta)$, where $\beta > 0$. Derive the mgf of X_1 and use it to show that the distribution of $S = X_1 + \cdots + X_n$ is Gamma(n, β).

(c) Suppose we wish to simulate from the Gamma$(3.5, 3.5)$ distribution. Consider using an accept-rejection algorithm with a Gamma(n, β) candidate, where $n \in \mathbb{N}$ and $\beta > 0$. Which values of n and β will lead to valid accept-rejection algorithms?

(d) Chose specific valid values for n and β and then write down the accept-reject algorithm. Remember, you have access to an unlimited supply of iid Uniform$(0, 1)$ random variables and nothing more.

(e) On average, how many iterations of your algorithm will be required to get a single draw from the Gamma$(3.5, 3.5)$ distribution? (You don't have to prove anything here, and your answer may involve the gamma and exponential functions.)

4. Suppose that $n \geq 2$ and that $X_1, \ldots, X_n$ are iid Bernoulli(p). (Throughout this problem, you may use the fact that a function of a complete, sufficient statistic is best unbiased for its expectation.)

(a) Find the ML estimator of p. Is it unbiased?

(b) Without appealing to results for exponential families, prove that $\sum_{i=1}^n X_i$ is a complete statistic.

(c) Find the best unbiased estimator of p.

(d) Find the ML estimator of $p(1-p)$. Is it unbiased?

(e) The sample variance, $\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$, is an unbiased estimator of $p(1-p)$. Is it best unbiased?

5. Suppose that $X_1, \ldots, X_n$ are iid N$(0, \sigma^2)$ and that $Y_1, \ldots, Y_m$ are iid N$(0, \tau^2)$. Assume further that $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are independent.

(a) Find the ML estimator of σ^2.

(b) Construct the LRT statistic for testing $H_0 : \sigma^2 = \tau^2$ against $H_1 : \sigma^2 \neq \tau^2$.

(c) Show that the LRT statistic can be written in such a way that it involves the data only through the statistic

$$F = \frac{\frac{1}{m}\sum_{i=1}^m Y_i^2}{\frac{1}{n}\sum_{i=1}^n X_i^2} \, .$$

(d) Find the distribution of F under H_0.

(e) The general LRT theory tells us to reject H_0 when the LRT statistic is small. Give an equivalent rejection rule in terms of F.

(f) Explain exactly what you would have to do to identify the rejection region of the size 0.05 LRT.

6. Data pairs $(X_i, Y_i),\ i = 1, \ldots, n$, are used to fit a simple linear regression model of the Y values on the X values.

(a) Write a scalar model equation for the simple linear regression. Identify the mean-related parameters.

(b) Write an expression for a real function of the mean-related parameters that is minimized only by OLS estimates.

(c) Derive the *normal equations* (in scalar form) in terms of the data and parameters. Show your work.

(d) State the necessary and sufficient condition(s) under which the OLS estimates are unique.

(e) Suppose you impose the restriction that the intercept parameter equals zero. Derive an expression for the OLS estimate of the slope parameter under this restriction (when it uniquely exists).

7. For a series of data $y_t,\ t = 1, \ldots, n$, consider the model

$$y_t = \alpha_0 + \alpha_1 \cos(\pi t/2) + \beta_1 \sin(\pi t/2) + \epsilon_t, \qquad \epsilon_t \sim \text{iid N}(0, \sigma^2), \qquad t = 1, \ldots, n \tag{1}$$

where α_0, α_1, β_1, and σ^2 are unknown parameters, and the arguments of $\cos$ and $\sin$ are in radians (so that $\cos(\pi/2) = 0$, $\sin(\pi/2) = 1$, $\cos(\pi) = -1$, $\sin(\pi) = 0$, and so forth). Suppose $n = 8$.

(a) Is (1) a linear model? Explain.

(b) Write out the full matrix $\boldsymbol{X}$ in the matrix representation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ of this data model, where $\boldsymbol{Y} = (y_1, \ldots, y_8)'$, $\boldsymbol{\beta} = (\alpha_0, \alpha_1, \beta_1)'$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_8)'$.

(c) The data values are

$$y_1 = 18 \qquad y_2 = 13 \qquad y_3 = 10 \qquad y_4 = 17 \qquad y_5 = 18 \qquad y_6 = 11 \qquad y_7 = 6 \qquad y_8 = 27$$

(i) Compute the OLS estimates of α_0, α_1, and β_1 under model (1).

(ii) Compute the corrected total sum of squares and the residual (error) sum of squares. What are their degrees of freedom?

(iii) Form a 95% two-sided confidence interval for $E(y_9)$, the expected value of the (unobserved) datum at $t = 9$.

(iv) Test $H_0:\ \alpha_1 = \beta_1 = 0$ versus the alternative $H_a:\ \alpha_1 \neq 0$ or $\beta_1 \neq 0$. Use $\alpha = 0.05$.

(d) Suppose you tried to fit the higher-degree expansion of the form

$$y_t = \alpha_0 + \alpha_1 \cos(\pi t/2) + \beta_1 \sin(\pi t/2) + \alpha_2 \cos(\pi t) + \beta_2 \sin(\pi t) + \epsilon_t$$

using OLS. What problem would you encounter? Explain specifically.

8. A type of electronic component undergoes accelerated life testing by subjection to an extraordinarily high constant temperature during continuous operation until failure. Twenty-four components are tested, four at *each* of six distinct temperatures, and each component's lifetime is recorded. A plot of lifetime (hours) versus temperature suggests variance stabilization by using a base-10 log transform. Using OLS to fit constant-only, simple linear, and quadratic polynomial models for log-lifetime versus temperature (°C) yields the following approximate residual (error) *sums* of squares:

constant-only: 1.20771 simple linear: 0.19933 quadratic: 0.17910

A one-way ANOVA model, fit using OLS with each distinct temperature corresponding to a different factor level, yields a residual *sum* of squares of 0.12806.

(a) What specific model *assumption* was the log transformation chosen to help satisfy?

(b) Assuming the quadratic model is adequate, perform a test for whether the simple linear model is adequate. State H_0 and H_a, and use $\alpha = 0.05$.

(c) Perform a lack-of-fit test for the quadratic model. State H_0 and H_a, and use $\alpha = 0.05$.

(d) The OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of the intercept and slope in the simple linear model, and the usual (unbiased) estimate of the variance-covariance matrix of those estimates, are as follows:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} \approx \begin{bmatrix} 3.92 \\ -0.012 \end{bmatrix} \qquad \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) \approx \begin{bmatrix} 17495 & -148.85 \\ -148.85 & 1.2944 \end{bmatrix} \times 10^{-6}$$

Based on the simple linear model (assuming it is entirely correct),

(i) test $H_0 : \beta_1 = -0.01$ versus $H_a : \beta_1 < -0.01$. Use $\alpha = 0.05$.

(ii) form a symmetric two-sided 95% *prediction* interval for the log-lifetime of a component at the standard operating temperature of 25°C.

(e) What is the highest degree polynomial model that could be uniquely fit to these data using OLS?

9. 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:

$$\begin{array}{lll} \text{Model I:} & y_{ij} = \mu + \tau_i + \epsilon_{ij}, & \sum_{i=1}^{t} \tau_i = 0 \\ \text{Model II:} & y_{ij} = \mu + a_i + \epsilon_{ij}, & a_1, \ldots, a_t \sim \text{iid } \mathrm{N}(0, \sigma_a^2) \end{array}$$

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

(a) For each model, write out H_0 and H_a for the test of whether or not there are any factor effects, in terms of the parameters.

(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) Letting $F_{\nu_1,\nu_2}(\cdot)$ represent the (cumulative) distribution function of the F distribution with ν_1 numerator and ν_2 denominator degrees of freedom, write expressions for the p-values corresponding to the F-tests of part (c).

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

10. Designers of a portable solar oven are investigating the effect of design factors on maximum sustained temperature. There are three factors: window thickness (single vs. double), insulation type (air only vs. fiberglass), and reflective material (foil vs. mylar). Sixteen ovens are constructed, representing each possible factor level combination twice, then tested in a completely randomized fashion. Resulting temperatures (°C) are presented in the following table:

Foil	single		double	
air	90	80	70	80
fiberglass	110	130	100	120

Mylar	single		double	
air	120	100	80	80
fiberglass	130	110	90	110

(a) Name the type of *design* of this experiment. List all of the *treatments* it uses.

(b) Compute the treatment sum of squares and the error (residual) sum of squares. What are their corresponding degrees of freedom?

(c) Test for three-way interaction among the factors. State H_0 and H_a, and use $\alpha = 0.05$.

(d) Test for two-way interaction between window thickness and reflective material. State H_0 and H_a, and use $\alpha = 0.05$.

(e) Suppose this experiment had instead been conducted in a randomized complete block design, but was otherwise the same. How many blocks would there be, and why? What (if anything) would change about the analysis?