# First Year Examination
# Department of Statistics, University of Florida

May 8, 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
- CRD = completely randomized design
- iid = independent and identically distributed
- mgf = moment generating function
- ML = maximum likelihood
- pdf = probability density function
- pmf = probability mass function
- $\mathbb{Z}^+ = \{0, 1, 2, 3, \dots\}$
- $\mathbb{N} = \{1, 2, 3, \dots\}$

You may use the following facts/formulas without proof:

**Beta density:** $X \sim$ Beta$(\alpha, \beta)$ means $X$ has pdf

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

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

**Gamma density:** $X \sim$ 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:** $X \sim$ 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 $\sigma^2 > 0$.

**Poisson density:** $X \sim$ Poisson$(\lambda)$ means $X$ has pmf

$$f(x; \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}\, I_{\mathbb{Z}^+}(x)$$

where $\lambda > 0$.

**1.** We have studied the dice game called *craps*. In case you've forgotten the rules, here is the definition of craps from Webster's dictionary: "A gambling game played with two dice; a first throw of seven or eleven wins, and a first throw of two, three, or twelve loses; any other first throw, to win, must be repeated before a seven is thrown." Suppose Lucky Chucky plays a game of craps (with a pair of fair dice) and let the random variable $X$ denote the result of his first roll.

(a) Find the pmf of $X$.

(b) Find the conditional pmf of $X$ *given* that Lucky Chucky wins the game.

(Express all of your probabilities as ratios of integers; i.e., *do not use any decimals*!)

**2.** Suppose $X \sim \text{Gamma}(\alpha, \beta)$, $Y_1, Y_2, \ldots$ are iid $\text{Gamma}(1, 1)$, and $U_1, U_2, \ldots$ are iid $\text{Uniform}(0, 1)$.

(a) Derive the mgf of $X$ and use it to calculate the expected value and variance of $X$.

(b) Fix $\theta \in \mathbb{R}^+$ and $n \in \mathbb{N}$. Derive the pdf of $\theta \sum_{i=1}^{n} Y_i$.

(c) Derive the pdf of $-\log(U_1)$.

(d) Fix $n \in \mathbb{N}$ and assume $n$ is even. Derive the pdf of

$$U_1(1 - U_2)U_3(1 - U_4)U_5(1 - U_6) \cdots U_{n-1}(1 - U_n) .$$

(e) Consider the quadratic function with random coefficients given by $U_1 x^2 + U_2 x + U_3$. Find the probability that it has real roots.

**3.** Suppose that $X_1, \ldots, X_n$ are iid $\text{Poisson}(\lambda)$.

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

(b) Find the ML estimator of $\lambda$, call it $\hat{\lambda}(X)$.

(c) Find the mean and variance of $\hat{\lambda}(X)$. What can you conclude from this?

(d) Find the ML estimator of $h(\lambda) = \lambda^2 e^{-\lambda}$, call it $\hat{h}(X)$.

(e) Prove or disprove the following statement: $\hat{h}(X)$ is an unbiased estimator of $h(\lambda)$.

(f) Find the best unbiased estimator of $h(\lambda)$.

**4.** (The Accept-Reject Algorithm.) Suppose that we want to simulate random variables from the pdf $f : \mathbb{R} \to [0, \infty)$. Define $\mathsf{X} = \{x \in \mathbb{R} : f(x) > 0\}$. Let $g(x)$ be another pdf satisfying $\{x \in \mathbb{R} : g(x) > 0\} = \mathsf{X}$, and suppose that we have a positive number $M$ such that, for all $x \in \mathsf{X}$, $f(x) \leq Mg(x)$. Consider the following two-step algorithm.

---

1. Draw $Y \sim g(\cdot)$ and, independently, draw $U \sim \text{Uniform}(0, 1)$.
2. If $U < \frac{f(Y)}{Mg(Y)}$, then accept $Y$ and stop; otherwise, return to Step 1.

---

(a) Prove that the output of this algorithm is a draw from $f$.

(b) Prove that $M \geq 1$.

(c) What is the distribution of the number of times that Steps 1 and 2 must be repeated before the algorithm terminates?

(d) Now consider a situation where $f(x) = c\,h(x)$ and $h(x)$ is known, but we cannot integrate $h(x)$ in closed form, so the normalizing constant, $c$, is *unknown*. Suppose we can find a number $M^*$ such that, for all $x \in \mathsf{X}$, $h(x) \leq M^* g(x)$. Is there an alternative version of the accept-reject algorithm that can be used to make *exact* draws from $f(x)$ in this case? (Hint: $c$ is unknown, so it cannot appear in an algorithm.)

**5.** Let $X_1, \ldots, X_n$ be iid Uniform$(\theta, \theta + 1)$ where $\theta \in \Theta = [0, \infty)$, and assume that $n \geq 2$. In this question, we consider testing $H_0 : \theta = 0$ vs $H_1 : \theta > 0$.

(a) Consider the function

$$f(s, t) = n(n-1)(t-s)^{n-2} I(0 < s < t < 1) \; .$$

Show that this function is a valid joint pdf by demonstrating that it satisfies all the required properties.

(b) Let $Y_1, \ldots, Y_n$ be iid Uniform$(0, 1)$. The joint pdf in part (a) is actually the joint pdf of two members of the set $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$. Which two? (Explain your reasoning!)

(c) Using the results in parts (a) and (b), derive the joint pdf of $X_{(1)}$ and $X_{(n)}$.

(d) Now consider testing $H_0 : \theta = 0$ vs $H_1 : \theta > 0$ using a test with rejection region given by

$$\Big\{(x_1, \ldots, x_n) \in [0, \infty)^n : x_{(1)} \geq k \;\text{ or }\; x_{(n)} \geq 1\Big\}$$

where $k \in (0, 1)$. Find the value of $k$ that leads to a size $\frac{1}{10}$ test, and fix $k$ at that value for the remainder of this question.

(e) Find a closed form expression for the power function of the test.

(f) Which values of $n \in \{2, 3, 4, \ldots\}$ make the following statement true: The power of the test is greater than $\frac{9}{10}$ for all $\theta > \frac{1}{2}$?

**6.** Data pairs $(X_i, Y_i),\ i = 1, \ldots, 17$, are used to fit the model

$$Y_i \;=\; \beta_0 \,+\, \beta_1 X_i \,+\, \beta_2 X_i^2 \,+\, \beta_3 X_i^3 \,+\, \beta_4 X_i^4 \,+\, \epsilon_i$$

using ordinary least squares. The *sequential* ("Type I") sums of squares are, in order,

$$R(\beta_1 \,|\, \beta_0) \;=\; 21.4 \qquad R(\beta_2 \,|\, \beta_0\,\beta_1) \;=\; 9.1 \qquad R(\beta_3 \,|\, \beta_0\,\beta_1\,\beta_2) \;=\; 3.2 \qquad R(\beta_4 \,|\, \beta_0\,\beta_1\,\beta_2\,\beta_3) \;=\; 1.3$$

and the residual (error) sum of squares is 36.3. Suppose that there are *exactly* 5 distinct values of $X_i$ represented in the data. Assume the errors are independent and identically normally distributed.

(a) Compute the coefficient of determination ($R^2$).

(b) Perform a lack-of-fit test for the simple linear regression of $Y_i$ on $X_i$ ($\alpha = 0.05$). Remember to state $H_0$ and $H_a$.

(c) Write out the (mean-corrected) ANOVA table that would be obtained for the *simple linear regression* of $Y_i$ on $X_i$ (sums of squares, degrees of freedom, mean squares).

(d) Select an appropriate polynomial degree by successively testing (and dropping) model terms, as in backward elimination, with significance level 0.05. Show your work and your decision at each step.

(e) Suppose $\widehat{\beta}_4 = 0.22$ is the ordinary least squares estimate of $\beta_4$. Compute the usual estimated standard error of $\widehat{\beta}_4$. (Hint: $t_\nu^2 \;=\; F_{1,\nu}$)

**7.** Consider the model

$$y_{ijk\ell} \;=\; \mu \,+\, \alpha_i \,+\, \beta_{ij} \,+\, \gamma_{ijk} \,+\, \epsilon_{ijk\ell} \qquad i = 1, \ldots, a \quad j = 1, \ldots, b \quad k = 1, \ldots, c \quad \ell = 1, \ldots, n$$

$$\beta_{ij} \sim \text{iid N}(0, \sigma_\beta^2), \qquad \gamma_{ijk} \sim \text{iid N}(0, \sigma_\gamma^2), \qquad \epsilon_{ijk\ell} \sim \text{iid N}(0, \sigma^2), \qquad \text{all independent,}$$

where $\alpha_1, \ldots, \alpha_a$ are fixed values summing to zero, $\sigma^2 > 0$, and $a \geq 2,\ b \geq 2,\ c \geq 2,\ n \geq 2$.

(a) Find the correlation between $y_{2121}$ and $y_{2122}$.

(b) In terms of the observations $y_{ijk\ell}$, write expressions for *mean* squares corresponding to the terms $\beta_{ij}$, $\gamma_{ijk}$, and $\epsilon_{ijk\ell}$. (You may use the dot-and-bar notation.)

(c) Give the expected values of the mean squares from the previous part.

(d) Write expressions for the unbiased ANOVA estimators of $\sigma_\beta^2$, $\sigma_\gamma^2$, and $\sigma^2$.

(e) Construct a test statistic for $H_0 : \sigma_\beta^2 = 0$ versus $H_a : \sigma_\beta^2 > 0$, and state the condition under which $H_0$ is rejected.

(f) Form an exact 95% two-sided confidence interval for $\alpha_1 - \alpha_2$ based on the $t$ distribution. Use the notation $t_{\alpha,\nu}$ for the value exceeded with probability $\alpha$ by a $t$-distributed random variable with $\nu$ degrees of freedom.

(g) Form an exact 95% two-sided confidence interval for $\sigma^2$ based on the chi-square distribution. Use the notation $\chi^2_{\alpha,\nu}$ for the value exceeded with probability $\alpha$ by a (central) chi-square random variable with $\nu$ degrees of freedom.

**8.** Twenty-four comparable plots of land are used as experimental units to study strategies for controlling blight in potatoes. Under investigation are two fungicides (F1 versus F2) and two application times (Early versus Late). The response is yield of blight-free potatoes from each plot. The experiment is conducted in a CRD with 5 treatment groups. Group descriptions, sizes, and response summary statistics are as follows:

| Treatment Group: | No Fungicide | F1, Early | F1, Late | F2, Early | F2, Late |
|---|---|---|---|---|---|
| Group Size | 8 | 4 | 4 | 4 | 4 |
| Sample Mean | 15 | 26 | 20 | 32 | 30 |
| Sample Variance | 5 | 8 | 10 | 11 | 10 |

(a) Write out a linear (cell) means model equation appropriate for analysis of this data. Clearly define your notation and specify any conditions on terms.

(b) Compute a corresponding ANOVA table (sums of squares, degrees of freedom, mean squares).

(c) Form a contrast representing interaction between fungicide type and application time, and perform a test for interaction using this contrast ($\alpha = 0.05$). Remember to state $H_0$ and $H_a$.

(d) Form Bonferroni 95% simultaneous two-sided confidence intervals for the pairwise differences in mean yield for each active treatment condition versus the "No Fungicide" condition. What do you conclude from these?

(e) Suppose this experiment had been conducted in a *balanced* CRD, with (possibly) a different number of plots. Find the minimum number of plots per treatment group such that none of the intervals of the previous part is wider than it is under the current design, assuming the same estimate of error variance. What are the new widths? Show your work.

**9.** The calorie count $y_{ij}$ of a pork ($i = 1$) or poultry ($i = 2$) sausage is modeled versus sodium content $x_{ij}$ as

$$y_{ij} \;=\; \beta_{i0} \;+\; \beta_{i1} x_{ij} \;+\; \epsilon_{ij} \qquad\qquad \epsilon_{ij} \sim \text{iid N}(0, \sigma^2) \qquad\qquad i = 1, 2 \qquad j = 1, \ldots, 14$$

The ordinary least squares estimates, their usual (unbiased) estimated variance-covariance matrix, and the usual (unbiased) estimate of $\sigma^2$ are, respectively,

$$\widehat{\boldsymbol{\beta}} \;=\; \begin{bmatrix} \widehat{\beta}_{10} \\ \widehat{\beta}_{11} \\ \widehat{\beta}_{20} \\ \widehat{\beta}_{21} \end{bmatrix} \;=\; \begin{bmatrix} 84.6 \\ 0.175 \\ 21.8 \\ 0.225 \end{bmatrix} \qquad\qquad s^2(\widehat{\boldsymbol{\beta}}) \;=\; \begin{bmatrix} 450 & -1.08 & 0 & 0 \\ -1.08 & 0.0028 & 0 & 0 \\ 0 & 0 & 1350 & -2.88 \\ 0 & 0 & -2.88 & 0.0063 \end{bmatrix} \qquad\qquad s^2 \;=\; 468$$

Perform the following parts under the usual model assumptions.

(a) Test whether the expected increase in calories per unit increase in sodium is *less* for pork than for poultry sausage. State $H_0$ and $H_a$ and use $\alpha = 0.05$.

(b) Test $H_0 : \beta_{10} = \beta_{20},\ \beta_{11} = \beta_{21}$ versus the general alternative (using $\alpha = 0.05$).

(c) Form a 95% two-sided *prediction* interval for the calorie count of a poultry sausage with a sodium content of 350.

(d) Consider the alternative parameterization

$$y_{ij} \;=\; \beta_0 \;+\; \beta_1 x_{ij} \;+\; \beta_2 w_{ij} \;+\; \beta_3 x_{ij} w_{ij} \;+\; \epsilon_{ij}$$

where $w_{ij}$ equals 1 if $i = 1$ and equals 0 if $i = 2$. Compute the ordinary least squares estimates of $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$, and an unbiased estimate of the error variance.

(e) Compute the overall sample average of the $y_{ij}$ values ($\overline{y}_{\bullet\bullet}$).

**10.** Consider a linear model in the general matrix formulation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y} = [\, y_1 \ \cdots \ y_n \,]'$ is the column vector of dependent variables, $\boldsymbol{X}$ is a known $n \times p'$ constant matrix with rank $p'$, $\boldsymbol{\beta}$ is the column vector of regression parameters, and $\boldsymbol{\epsilon}$ is the vector of errors.

Let $e_i$ be the ordinary least squares residual for $y_i$, and let $h_{ii}$ be the corresponding *leverage* value. Provided $h_{ii} < 1$ for all $i$, and $n > p'$, the *standardized residuals* are

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} \qquad i = 1, \ldots, n \qquad \text{where} \quad s = \sqrt{\text{mean square for residual (error)}} > 0.$$

(a) Define the leverage $h_{ii}$ (in terms of the model).

(b) Suppose $\boldsymbol{Y}$ is replaced by $a\boldsymbol{Y}$ for some known constant $a \neq 0$. Do $e_i$ and $r_i$ change? If so, how?

(c) Show that $r_i^2 \le (n - p')/(1 - h_{ii})$ for all $i$. Can this ever hold with equality? When, or why not?

(d) Is the residual $e_i$ normally distributed under the usual normal-theory model assumptions on $\boldsymbol{\epsilon}$? If so, give its expected value and variance. If not, briefly explain how you know.

(e) Answer the question of the previous part for $r_i$ instead of $e_i$.

(f) Suppose the model is such that

$$E(y_i) = \begin{cases} \mu_1 & i = 1, \ldots, n_1 \\ \mu_2 & i = n_1 + 1, \ldots, n \end{cases}$$

for unknown constants $\mu_1$ and $\mu_2$. Using *only* the values $y_1, \ldots, y_n$, $n_1$, $n$, and explicit numbers, write expressions for $e_i$ and $h_{ii}$ (for all $i$).