First Year Examination
Department of Statistics, University of Florida
May 12, 2006, 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
- *corrected total sum of squares* = total SS corrected for the mean
- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- ML = maximum likelihood
- pdf = probability density function
- pmf = probability mass function
- UMP = uniformly most powerful
- α = specified probability of Type I error
- $\mathbb{N} = \{1, 2, 3, \ldots\}$
- $\mathbb{R}^+ = (0, \infty)$
- $\mathrm{N}(\mu, \sigma^2)$ = normal distribution with mean μ and variance σ^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 $e^{bt} M_X(at)$.

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

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

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

where $\nu > 0$.

1. Suppose that X is a discrete random variable taking values in the non-negative integers. Assume that X has an mgf. For $r \in \mathbb{N}$, the rth factorial moment of X is defined as

$$\mathrm{E}^{(r)}(X) = \mathrm{E}\big[X(X-1)\dots(X-r+1)\big] \ .$$

Note that $\mathrm{E}^{(1)}(X) = \mathrm{E}(X)$.

(a) Express $\mathrm{Var}(X)$ in terms of factorial moments.

(b) The *probability generating function* of X is defined as

$$P_X(t) = \mathrm{E}\big(t^X\big) \ ,$$

for $|t| < h$ where $h > 1$ is the *radius of convergence*. Show that the probability generating function generates the factorial moments in the sense that

$$\frac{d^r}{dt^r} P_X(t)\Big|_{t=1} = \mathrm{E}^{(r)}(X) \ .$$

(c) Show that the factorial moments of the $\mathrm{Poisson}(1)$ distribution are all equal and find the value.

2. In this problem, we will do some calculations involving normal random variables.

(a) Find the mgf of a standard normal random variable and use this to find the distribution of $\sum_{i=1}^m W_i$, where $W_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ and the W_is are independent.

(b) Let $X_1, \dots, X_m$ be iid $\mathrm{N}\big(\mu_X, \sigma_X^2\big)$ and let $Y_1, \dots, Y_n$ be iid $\mathrm{N}\big(\mu_Y, \sigma_Y^2\big)$, and assume that the X_is are independent of the Y_is. As usual, let $\overline{X} = m^{-1}\sum_{i=1}^m X_i$ and $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$. Use the result from part (a) to show that, under $H_0 : \mu_X = \mu_Y$, we have

$$\overline{X} - \overline{Y} \sim \mathrm{N}\Big(0, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\Big) \ .$$

(c) Under H_0, an alternate sampling model (sometimes used in re-sampling experiments) is as follows. Assume that $X_1^*, \dots, X_m^*$ are iid with

$$X_1^* \sim \begin{cases} \mathrm{N}\big(0, \sigma_X^2\big) & \text{with probability } \frac{m}{m+n} \\ \mathrm{N}\big(0, \sigma_Y^2\big) & \text{with probability } \frac{n}{m+n} \end{cases} \ ,$$

and that $Y_1^*, \dots, Y_n^*$ are iid with

$$Y_1^* \sim \begin{cases} \mathrm{N}\big(0, \sigma_X^2\big) & \text{with probability } \frac{m}{m+n} \\ \mathrm{N}\big(0, \sigma_Y^2\big) & \text{with probability } \frac{n}{m+n} \end{cases} \ ,$$

and that the X_i^*s are independent of the Y_i^*s. Show that (surprisingly),

$$\mathrm{Var}\big(\overline{X}^* - \overline{Y}^*\big) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} \ .$$

(Note that m and n are flipped from part (b).)

3. Suppose that $X_1, \ldots, X_n$ are iid random variables such that

$$P(X_1 = x) = p(1-p)^x \text{ for } x = 0, 1, 2, \ldots$$

where $p \in (0, 1)$.

(a) Does the mgf of X_1 exist? If so, what is it?

(b) Suppose that $Y \sim \text{NB}(r, s)$; that is,

$$P(Y = y) = \binom{r+y-1}{y} s^r (1-s)^y \text{ for } y = 0, 1, 2, \ldots$$

where $s \in (0, 1)$ and $r \in \mathbb{N}$. Find the mgf of Y.

(c) Find the pmf of the random variable $Z = \sum_{i=1}^n X_i$.

(d) Find the ML estimator of $g(p) = p(1-p)$, call it $\widehat{g(p)}$.

(e) Is $\widehat{g(p)}$ the best unbiased estimator of $g(p)$? If not, find the best unbiased estimator of $g(p)$.

4. Suppose that $X_1, \ldots, X_n$ are iid with common pdf $f(x|\theta) = \theta^{-1} I_{[0,\theta]}(x)$, where $\theta \in \Theta = \mathbb{R}^+$.

(a) Find the ML estimator of θ.

(b) Fix $\alpha \in (0, 1)$ and $\theta_0 \in \mathbb{R}^+$. Find the level-$\alpha$ LRT of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

(c) Find the power function of this LRT. (Hint: Consider the two different cases $\theta \in (0, \theta_0]$ and $\theta > \theta_0$.)

(d) Is this LRT unbiased?

(e) Find a UMP level-α test of $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$.

(f) Find a UMP level-α test of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

(g) Prove or disprove the following statement: The LRT developed in part (b) is a UMP level-α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$?

5. Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{N}(\theta, \sigma^2)$ and that the prior on (θ, σ^2), which we denote by $\pi(\theta, \sigma^2)$, is characterized by $\theta|\sigma^2 \sim \mathrm{N}(\mu, \tau\sigma^2)$ and $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$ where $\mu \in \mathbb{R}$ and $\tau, \alpha, \beta \in \mathbb{R}^+$. Let $x = (x_1, \ldots, x_n)$ denote the observed data. In this question, we will demonstrate that $\pi(\theta, \sigma^2)$ is a conjugate prior by showing that the posterior density, $\pi(\theta, \sigma^2|x)$, can be written in the same form as the prior.

(a) Write down the prior density, $\pi(\theta, \sigma^2)$.

(b) Show that

$$\sum_{i=1}^{n}(x_i - \theta)^2 = (n-1)S^2 + n(\overline{x} - \theta)^2$$

where, as usual, $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ and $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$.

(c) Use the fact that

$$-\frac{1}{2\tau\sigma^2}(\theta - \mu)^2 - \frac{n}{2\sigma^2}(\overline{x} - \theta)^2 = -\frac{(n\tau + 1)}{2\tau\sigma^2}\Big(\theta - \frac{\mu + n\tau\overline{x}}{n\tau + 1}\Big)^2 - \frac{n}{2\sigma^2(n\tau + 1)}\big(\overline{x} - \mu\big)^2$$

to show that $\theta|\sigma^2, x \sim \mathrm{N}(\mu', \tau'\sigma^2)$ and $\sigma^2|x \sim \mathrm{IG}(\alpha', \beta')$, where μ', τ', α' and β' are (potentially) functions of x, n, μ, τ, α and β that you must identify.

(d) Find $\pi(\theta|x)$. What type of density is this? (Hint: First, answer the question for $\pi(\theta)$ and then use the conjugacy results.)

6. Suppose independent data pairs $(x_{ij}, y_{ij}),\ i = 1, \ldots, 6,\ j = 1, \ldots, 4$ are observed, where index i represents a treatment group, and index j represents a replication (within each treatment group).

The following four models are fit to the data (using ordinary least squares), with the resulting residual (error) sums of squares as specified:

Model 1:	$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$	$\mathrm{SS(Res)} = 750$
Model 2:	$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}$	$\mathrm{SS(Res)} = 600$
Model 3:	$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$	$\mathrm{SS(Res)} = 400$
Model 4:	$y_{ij} = \mu + \beta x_{ij} + \epsilon_{ij}$	$\mathrm{SS(Res)} = 800$

The corrected total sum of squares is 1000. In these models, $\sum_{i=1}^{6}\alpha_i = 0$, parameters $\mu, \beta, \beta_1, \ldots, \beta_6$ are unrestricted, and ϵ_{ij} is the error term.

Perform the following F-tests at level $\alpha = 0.05$, clearly stating the null and alternative hypotheses in terms of the parameters. You may assume that the usual normal-theory conditions are valid in each case.

(a) Test for treatment effects on the response variable y_{ij} as a simple one-way ANOVA (ignoring x_{ij}).

(b) Test whether the slope parameter in the simple linear regression of y_{ij} on x_{ij} is nonzero, ignoring all effects due to different treatment groups.

(c) Test whether different treatment groups have different slope parameters for the regression of y_{ij} on x_{ij}, allowing for separate intercepts for each group.

(d) Perform a classical analysis of covariance (ANCOVA) test for treatment effects on y_{ij} (adjusting for a linear effect in the covariate x_{ij}).

7. Suppose data pairs (x_i, y_i) are observed, for $i = 1, \ldots, n$. Let $\boldsymbol{Y}$ be the $n \times 1$ vector with y_i in position i, and let $\boldsymbol{X}$ be the $n \times 3$ matrix having row i equal to $[\ 1 \ \ x_i \ \ x_i^2\ ]$. Suppose we compute that

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 20 & 0 & 30 \\ 0 & 30 & -30 \\ 30 & -30 & 90 \end{bmatrix} \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{30}\begin{bmatrix} 6 & -3 & -3 \\ -3 & 3 & 2 \\ -3 & 2 & 2 \end{bmatrix} \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} 260 \\ -30 \\ 480 \end{bmatrix} \quad \boldsymbol{Y}'\boldsymbol{Y} = 4670$$

(a) Find n and the average of the values $y_1, \ldots, y_n$.

(b) Find the ordinary least squares estimates $\widehat{\beta}_{0,Q}, \widehat{\beta}_{1,Q}, \widehat{\beta}_{2,Q}$ of $\beta_{0,Q}, \beta_{1,Q}, \beta_{2,Q}$ in the quadratic model

$$y_i = \beta_{0,Q} + \beta_{1,Q}\, x_i + \beta_{2,Q}\, x_i^2 + \epsilon_i.$$

Also, find the residual sum of squares for this model.

(c) Find an unbiased estimate of the variance of $\widehat{\beta}_{1,Q} - \widehat{\beta}_{2,Q}$.

(d) Find the ordinary least squares estimates $\widehat{\beta}_{0,L}, \widehat{\beta}_{1,L}$ of $\beta_{0,L}, \beta_{1,L}$ in the simple linear model

$$y_i = \beta_{0,L} + \beta_{1,L}\, x_i + \epsilon_i.$$

Also, find the residual sum of squares for this model.

(e) Test whether the simple linear model of part (d) is adequate relative to the quadratic model of part (b). Use $\alpha = 0.05$. Be sure to state the null and alternative hypotheses.

8. The entries of three finalists in a chili cooking competition are rated by each of five judges. Each judge tastes the entries in a random order and assigns a numerical rating to each entry (larger ratings being better), as given in the following table:

	Entry 1	Entry 2	Entry 3
Judge 1	50	40	60
Judge 2	70	50	60
Judge 3	70	60	80
Judge 4	45	10	35
Judge 5	65	15	40

In the following analyses, regard the judges as blocks.

(a) Find the corrected total sum of squares and the sums of squares for the entries and for the judges (blocks).

(b) Perform an F-test to determine whether there are any statistically significant differences among the mean ratings of the entries ($\alpha = 0.05$).

(c) Form Bonferroni simultaneous 95% two-sided confidence intervals for all pairwise differences between the mean ratings of the entries.

9. Consider the linear model in the general matrix formulation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ 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, $\boldsymbol{\beta}$ is the vector of regression parameters, and the error vector $\boldsymbol{\epsilon}$ has a multivariate normal distribution with mean vector zero and variance-covariance matrix $\boldsymbol{I}\sigma^2$. ($\boldsymbol{I}$ = identity matrix)

In the following, carefully define any matrix notation you use that is not introduced above.

(a) Form the vector of ordinary least squares residuals, in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

(b) Form the residual sum of squares, SS(Res), in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

(c) Using SS(Res), form a random variable having a (central) chi-square distribution. How many degrees of freedom does it have?

(d) Using the chi-square random variable from part (c), derive a $(1-\alpha)100\%$ two-sided (equal-tailed) confidence interval for σ^2. (Use the notation $\chi^2_{\varepsilon,\nu}$ to denote the value exceeded with probability ε by a chi-square random variable with ν degrees of freedom.)

10. 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 $i = 1,2,3,\ j = 1,2,3,4,5,\ k = 1,2,3,4,5,6$. The data are analyzed using the model equation

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

where the terms ϵ_{ijk} are iid $\mathrm{N}(0,\sigma^2)$. The sums of squares for the effects and error are as follows:

$$\text{SS(A)} = 120 \qquad \text{SS(B)} = 300 \qquad \text{SS(AB)} = 150 \qquad \text{SS(Error)} = 800$$

Consider the following two separate sets of conditions on the model terms:

Condition Set 1: $\sum_{i=1}^{3}\alpha_i = \sum_{j=1}^{5}\beta_j = 0, \qquad \sum_{i=1}^{3}\alpha\beta_{ij} = 0, \quad \text{all } j, \qquad \sum_{j=1}^{5}\alpha\beta_{ij} = 0, \quad \text{all } i$

Condition Set 2: $\alpha_i \sim \mathrm{N}(0,\sigma^2_\alpha), \quad \beta_j \sim \mathrm{N}(0,\sigma^2_\beta), \quad \alpha\beta_{ij} \sim \mathrm{N}(0,\sigma^2_{\alpha\beta}),$
$\alpha_i, \beta_j, \alpha\beta_{ij}, \epsilon_{ijk}$ independent, all i,j,k

Answer the following questions, first under Condition Set 1, then under Condition Set 2.

(a) Perform the F-test for interaction between factors A and B, clearly stating the null and alternative hypotheses in terms of the parameters ($\alpha = 0.05$).

(b) Perform the F-test for the main effect of factor A, clearly stating the null and alternative hypotheses in terms of the parameters ($\alpha = 0.05$).

(c) In terms of the parameters, find (i) the mean of y_{111} and (ii) the correlation between y_{111} and y_{112}.