First Year Examination
Department of Statistics, University of Florida
May 9, 2008, 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 "applied" questions and questions 6 through 10 are the "theory" questions. You must do exactly four of the applied questions and exactly four of the theory 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
- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- ML = maximum likelihood
- MSE = mean squared error
- pdf = probability density function
- $\mathbb{R}^+ = (0, \infty)$
- N$(\mu, \sigma^2)$ = normal distribution with mean $\mu$ and variance $\sigma^2$
- $H_0$ = Null hypothesis
- $H_a$ = Alternative hypothesis
- $\sim$ = "is distributed as"

You may use the following facts/formulas without proof:

**Normal density:** $X \sim \mathrm{N}(\mu, \sigma^2)$ means

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

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

**Beta density:** $X \sim \mathrm{Beta}(\alpha, \beta)$ means

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

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

**Gamma Density:** $X \sim \mathrm{Gamma}(\alpha, \beta)$ 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 $\beta > 0$. Also, $\mathrm{E}(X) = \alpha\beta$ and $\mathrm{Var}(X) = \alpha\beta^2$. The mgf is given by $m_X(t) = (1-\beta t)^{-\alpha}$ for $t < 1/\beta$.

**Iterated Expectation Formula:** $\mathrm{E}(X) = \mathrm{E}\left[\mathrm{E}(X|Y)\right]$.

**Iterated Variance Formula:** $\mathrm{Var}(X) = \mathrm{E}\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left[\mathrm{E}(X|Y)\right]$.

**1.** For each value of a variable $X$, some observations of a variable $Y$ are as shown in the following table:

| $X$ | $Y$ values | | | |
|---|---|---|---|---|
| 0 | 9 | 5 | 7 | 7 |
| 10 | 8 | 6 | | |
| 20 | 1 | 0 | 1 | 6 |

(a) Compute the ordinary least squares estimates of the intercept and slope parameters in the simple linear regression of $Y$ on $X$. (Be sure to identify which is which.)

(b) Compute an ANOVA table (corrected for the mean) for the simple linear regression. Based on the model, test whether there is any relationship between $Y$ and $X$ (making the usual assumptions about the data and using $\alpha = 0.05$). (Remember to state the null and alternative hypotheses.)

(c) Compute an unbiased estimate of the expected value of a new observation $Y$ at $X = 5$ that follows the same simple linear regression model. Compute a 95% symmetric two-sided confidence interval for this quantity.

(d) Test whether there is any evidence that a quadratic regression model would provide a better fit to the data ($\alpha = 0.05$). (Remember to state the null and alternative hypotheses.)

**2.** Consider the following model

$$y_{ijk} = \mu + \alpha_i + \eta_{k(i)} + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \qquad i = 1, \ldots, a$$

$$\eta_{k(i)} \sim \text{iid N}(0, \sigma_\eta^2), \qquad \epsilon_{ijk} \sim \text{iid N}(0, \sigma^2), \qquad \text{all independent} \qquad j = 1, \ldots, b$$

$$\textstyle\sum_i \alpha_i = \sum_j \beta_j = \sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0 \qquad k = 1, \ldots, n$$

which is often used for a split-plot design with one whole-plot factor ($A$) and one split-plot factor ($B$): $y_{ijk}$ is the response from the split plot that receives level $j$ of $B$, in the $k^{\text{th}}$ whole plot that receives level $i$ of $A$.

(a) Find the expected value and variance of $y_{ijk}$.

(b) Find the *correlation* between the responses from two different split plots in the same whole plot.

(c) *Derive* the expected value and variance of $\bar{y}_{ij\bullet} - \bar{y}_{i'j\bullet}$ for $i \neq i'$, where $\bar{y}_{ij\bullet} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$.

(d) *Derive* the expected value of

$$\sum_{i=1}^{a}\sum_{k=1}^{n}\left(\bar{y}_{i\bullet k} - \bar{y}_{i\bullet\bullet}\right)^2,$$

where $\bar{y}_{i\bullet k} = \frac{1}{b}\sum_{j=1}^{b} y_{ijk}$ and $\bar{y}_{i\bullet\bullet} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$.

**3.** Twenty athletes are randomly split into 4 groups, with 5 athletes in each (using a method that makes all such splits equally likely). Three groups are each assigned to drink one of three different brands of "energy drink", and the fourth group is given none. An hour later, all athletes compete in a 2 mile running race, and their times $Y$ (minutes) to complete the race are recorded. On the previous day, all athletes had run the same 2 mile race without consuming energy drinks, and their completion times $X$ (minutes) were recorded. Analysis of the variance of $Y$ in terms of the group and the covariate $X$ (but *not* their interaction) produces the following sequential (Type I) sums of squares:

| Source | Sequential SS |
|---|---|
| Group | 8.13 |
| Previous Day's Time ($X$) | 16.16 |

| Source | Sequential SS |
|---|---|
| Previous Day's Time ($X$) | 17.63 |
| Group | 6.66 |

(Note the different order of the sources in the two tables.)

The *sum of the squared residuals* is $4.13$.

(a) Name the type of *design* that this experiment uses.

(b) For this experiment, why is it unnecessary to test whether the treatment affects the covariate $X$? Explain *briefly*.

(c) Briefly explain the purpose of including $X$ as a covariate in this experiment.

(d) Write a linear model equation for the analysis using the covariate, clearly defining all symbols and stating any conditions they satisfy.

(e) Test whether energy drink consumption affects running time. (Your test *should* account for the effect of the covariate $X$). Clearly state your conclusion. Use $\alpha = 0.05$.

(f) Perform the same test, but this time *ignoring* the covariate (i.e. based on the simple one-way analysis of variance, as if $X$ had not been recorded). Clearly state your conclusion. Use $\alpha = 0.05$.

**4.** A medical study seeks to understand the effects of a new prescription drug and a nutritional supplement for treating a rare disorder, based on a single response variable. The study uses four treatment conditions: (1) placebo, (2) drug only, (3) supplement only, and (4) both drug and supplement. Human patients enter the study at different times. Once 4 patients have entered the study, the 4 treatments are randomly assigned to these patients (each patient receiving a different treatment). This process continues for each new block of 4 consecutive patients, until a total of 24 patients have entered the study. Let

$$y_{ij} = \text{ response of patient from } j^{\text{th}} \text{ block who is assigned treatment condition } i$$

for $i = 1, \ldots, 4$ and $j = 1, \ldots, 6$. Suppose

$$\bar{y}_{1\bullet} = 20, \quad \bar{y}_{2\bullet} = 15, \quad \bar{y}_{3\bullet} = 20, \quad \bar{y}_{4\bullet} = 5,$$
$$\sum_{j=1}^{6}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2 = 79, \qquad \sum_{i=1}^{4}\sum_{j=1}^{6}(y_{ij} - \bar{y}_{\bullet\bullet})^2 = 1531$$

where

$$\bar{y}_{i\bullet} = \frac{1}{6}\sum_{j=1}^{6} y_{ij} \qquad \bar{y}_{\bullet j} = \frac{1}{4}\sum_{i=1}^{4} y_{ij} \qquad \bar{y}_{\bullet\bullet} = \frac{1}{24}\sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}$$

and suppose that, within each treatment condition, $y_{ij}$ is normally distributed, with the same variance for all responses, regardless of treatment condition.

(a) Write a linear model equation appropriate for analyzing the responses $y_{ij}$ from this experiment, clearly specifying the conditions on all terms, if any.

(b) Write an ANOVA table appropriate for your model.

(c) Test whether there are any differences between the treatment conditions ($\alpha = 0.05$). (Remember to state the null and alternative hypotheses in terms of your model.)

(d) Test whether the effects of the drug and supplement on the responses $y_{ij}$ could be additive, i.e. whether or not there appears to be interaction between the effects of the drug and supplement ($\alpha = 0.05$). (Remember to state the null and alternative hypotheses in terms of your model.)

(e) Can you assert that the response mean is significantly *smaller* when the drug is given *without* the supplement versus when only a placebo is given? Test at level $\alpha = 0.05$. (Remember to state the null and alternative hypotheses in terms of your model.)

**5.** Consider the linear model

$$Y \;=\; \beta_0 \,+\, \beta_1 X_1 \,+\, \beta_2 X_2 \,+\, \beta_3 X_3 \,+\, \epsilon, \qquad \epsilon \;\sim\; \mathrm{N}(0, \sigma^2).$$

Suppose you have $n$ independent observations from this model, with the variable values for the $i^{\text{th}}$ observation denoted $(Y_i, X_{i1}, X_{i2}, X_{i3})$. If $\boldsymbol{X}$ is the matrix whose $i^{\text{th}}$ row is $[\, 1 \; X_{i1} \; X_{i2} \; X_{i3} \,]$ and $\boldsymbol{Y}$ is the column vector whose $i^{\text{th}}$ element is $Y_i$, summary statistics are as follows:

$$\boldsymbol{X'X} = \begin{bmatrix} 25 & 0 & 0 & 0 \\ 0 & 4 & -4 & -3.6 \\ 0 & -4 & 104 & -6.4 \\ 0 & -3.6 & -6.4 & 5.24 \end{bmatrix} \qquad (\boldsymbol{X'X})^{-1} = \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 1.26 & 0.11 & 1 \\ 0 & 0.11 & 0.02 & 0.1 \\ 0 & 1 & 0.1 & 1 \end{bmatrix} \qquad \boldsymbol{X'Y} = \begin{bmatrix} 55 \\ 4.4 \\ -34.4 \\ 4.54 \end{bmatrix}$$

$$\boldsymbol{Y'Y} = 182.52$$

(a) Find $n$ and the average of the values $Y_1, \ldots, Y_n$.

(b) Note the zeros in the matrix $\boldsymbol{X'X}$. What do they tell you about the independent variables?

(c) Compute the least squares estimates of $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$.

(d) Compute the sum of the squared residuals and the coefficient of determination ($R^2$).

(e) Find the *partial* sums of squares for $X_1$, $X_2$, and $X_3$. (Hint: Recall the relationship between $t$ and $F$ statistics for testing whether $\beta_j = 0$ ?)

(f) Suppose *backward elimination* is applied (until all variables are dropped). Which would be the first independent variable to be dropped? Explain briefly.

(g) Suppose *forward selection* is applied (until all variables have been added). Which would be the first independent variable to be added? Explain briefly.

(h) Test $\mathrm{H}_0 : \; \beta_1 = \beta_3$ versus the alternative $\mathrm{H}_a : \; \beta_1 \neq \beta_3$. Use $\alpha = 0.05$.

**6.** Let $X_1, \ldots, X_n$ be iid Uniform$(0, \theta)$ where $\theta > 0$; that is, the common pdf is

$$f(x|\theta) = \theta^{-1} I_{(0,\theta)}(x) \; .$$

(a) Is $\theta$ a location parameter, scale parameter, or neither of these? Carefully explain your answer.

(b) Show that $X_{(n)}$ is a complete sufficient statistic.

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

(d) Find the ML estimator of $1/\theta$.

(e) Find the pdf of $\hat{\theta}(X)/\theta$.

(f) Show that $n\Big(1 - \frac{\hat{\theta}(X)}{\theta}\Big)$ converges in distribution and find the limiting distribution.

(g) Find the method of moments estimator of $\theta$.

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

**7.** Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent random variables such that $Y_i \sim \text{Gamma}(\alpha, \beta_i)$ with $\alpha > 1$. In this problem, we consider the following expectation

$$\text{E}\left[\frac{1}{\sum_{i=1}^n Y_i}\right] . \tag{1}$$

(a) Give an exact formula for (1) in the special case where the $\beta_i$s are all equal with common value $\beta$.

When the $\beta_i$s are not equal, exact calculation is impossible so we develop upper and lower bounds.

(b) Use Jensen's inequality to derive a simple lower bound for (1).

(c) The upper bound requires more work. Let $x_1, \ldots, x_n$ be positive numbers and define the arithmetic mean $(m_A)$, geometric mean $(m_G)$, and harmonic mean $(m_H)$ of these numbers as follows

$$m_A = \frac{1}{n}\sum_{i=1}^n x_i \ , \qquad m_G = \left[x_1 x_2 \cdots x_n\right]^{\frac{1}{n}} \ , \qquad \text{and} \qquad m_H = \frac{1}{\frac{1}{n}\sum_{i=1}^n \frac{1}{x_i}} \ .$$

Let $X$ be a discrete random variable that is uniform on the set $\{x_1, x_2, \ldots, x_n\}$. Use $X$ together with Jensen's inequality to prove that $m_H \leq m_G \leq m_A$.

(d) Use the inequality $m_H \leq m_A$ to get an upper bound on (1) that is similar in form to the result from part (a).

**8.** Let $X$ and $Y$ be jointly continuous random variables with joint pdf given by

$$f(x,y) = \begin{cases} y\, e^{-y(x+1)} & x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the probability that $X$ exceeds $Y$? (You do not have to evaluate the integrals.)

(b) Prove or disprove the following statement: $X$ and $Y$ are independent.

(c) Find the marginal density of $X$ and use it to calculate

$$\text{E}\left[\frac{1}{(X+1)^p}\right]$$

for $p \in \mathbb{R}$.

(d) Find the conditional density of $Y$ given $X = x$.

(e) Show that, conditional on $X = x$, the probability that $Y$ exceeds 2 is no more than $\frac{1}{x+1}$.

(f) Find the conditional mean and variance of $Y$ given $X = x$.

(g) *Use the results from (f) together with the iterated variance formula* to calculate the (marginal) variance of $Y$.

(h) Find $P(X > 1 \mid Y > \pi)$. (You do not have to evaluate the integrals.)

(i) Find $P(X > 1 \mid Y = \pi)$. (You do not have to evaluate the integrals.)

**9.** Suppose that $X \sim \text{Bernoulli}(p)$ and that the prior density for $p$ is $\text{Beta}(\alpha, \beta)$.

(a) Find the posterior density of $p$ given $x$, call it $\pi(p|x)$.

(b) Find the Bayes estimator of $p$, call it $\tilde{p}(X)$, and show that it can be written as a weighted average of the ML estimator of $p$, which is $\hat{p}(X) = X$, and the prior mean of $p$.

(c) Find the MSE of $\tilde{p}(X)$ and the MSE of $\hat{p}(X)$.

(d) Suppose that $\alpha = \beta = 1$. Does either of these estimators dominate the other in terms of MSE? In other words, is one of the MSE functions uniformly below the other?

Now consider an alternative prior that is a mixture of $k$ beta densities; that is,

$$\pi^*(p) = \sum_{i=1}^{k} r_i \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} p^{\alpha_i - 1} (1-p)^{\beta_i - 1} I_{(0,1)}(p) \ ,$$

where, for each $i \in \{1, \ldots, k\}$, $\alpha_i, \beta_i \in \mathbb{R}^+$, $r_i \in (0,1)$ and $\sum_{i=1}^{k} r_i = 1$.

(e) Find the exact posterior density of $p$ given $x$, call it $\pi^*(p|x)$.

(f) Is $\pi^*(p)$ a conjugate prior? (A yes/no answer is not sufficient here.)

**10.** 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, $\beta$ is an unknown regression parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\text{N}(0, \sigma^2)$ with $\sigma^2$ known. In this question, we develop the LRT of $H_0 : \beta = 0$ against $H_a : \beta \neq 0$. We start with some preliminary results.

(a) Derive the mgf of $V \sim \text{N}(\mu, \tau^2)$.

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

Now back to the original problem.

(c) Find the ML estimator of $\beta$, call it $\hat{\beta}(Y)$.

(d) Construct the LRT statistic for testing $H_0 : \beta = 0$ against $H_a : \beta \neq 0$.

(e) Show that the LRT statistic can be written in such a way that it involves the data, $Y$, only through $T = \hat{\beta}^2(Y)$.

(f) Find the distribution of $T = \hat{\beta}^2(Y)$ under $H_0$.

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

(h) Suppose that $n = 100$, $\sum_{i=1}^{100} x_i^2 = 10$ and $\sigma^2 = 5$. Give the *exact* rejection region of the size $0.10$ LRT in terms of $T$. (You may use the fact that $P(Z > 1.645) = 0.05$ when $Z \sim \text{N}(0,1)$.)