First Year Examination
Department of Statistics, University of Florida

August 18, 2011, 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. 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
4. **Write your answers on the blank paper provided. Write only on one side of the paper, and start each question on a new page.**
5. **Put your number in the upper right-hand corner of every page you turn in. Do not write your name anywhere on your exam.**
6. While the 10 questions are equally weighted, some questions 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 are used throughout:

- iid = independent and identically distributed
- LRT = likelihood ratio test
- mgf = moment generating function
- ML = maximum likelihood
- MSE = mean squared error
- NP = Neyman-Pearson
- pdf = probability density function
- pmf = probability mass function
- $\mathbb{N} = \{1, 2, 3, \dots\}$

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 $\sigma^2 > 0$.

Poisson mass function: $X \sim \text{Poisson}(\lambda)$ means X has pmf

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

where $\lambda \geq 0$.

Beta density: $X \sim \text{Beta}(\alpha, \beta)$ 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 $\beta > 0$. Also, $\mathrm{E}(X) = \alpha/(\alpha+\beta)$.

1. Let Y and Z be two random variables with finite second moments.

(a) Use the function $h(t) = \mathrm{E}\Big\{\big[(Y - \mathrm{E}Y)t + (Z - \mathrm{E}Z)\big]^2\Big\}$ to prove that $-1 \leq \rho_{YZ} \leq 1$ where ρ_{YZ} denotes the correlation between Y and Z.

(b) Prove or disprove the following statement: Y and Z are independent if and only if $\mathrm{Cov}(Y, Z) = 0$.

Let $n \in \{2, 3, 4, \dots\}$ and suppose that $X_1, \dots, X_n$ are random variables such that $0 < \mathrm{Var}(X_i) < \infty$ for $i = 1, 2, \dots, n$.

(c) Show that

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j)\ .$$

(d) Now suppose that $X_1, \dots, X_n$ are iid and let $m \in \{1, 2, \dots, n-1\}$. Define $U = \sum_{i=1}^{m} X_i$ and $V = \sum_{i=1}^{n} X_i$ and show that the correlation of U and V can be expressed as a simple function of m and n.

2. Suppose that $X \sim \mathrm{Bernoulli}(p)$.

(a) Let $\hat{p}(X)$ be an estimator of p. Show that the MSE of $\hat{p}(X)$ takes the form $ap^2 + bp + c$, and identify a, b and c.

(b) Let $\mathcal{C}$ denote the class of estimators whose MSEs are linear functions of p. Which estimators are in $\mathcal{C}$?

(c) Suppose we put a $\mathrm{Beta}(\alpha, \beta)$ prior on p. Identify the (α, β) pairs that lead to Bayes estimators in the class $\mathcal{C}$.

Now consider a different family of priors for p that takes the form

$$\pi(p; d) = dI_{(0,0.5)}(p) + (2 - d)I_{(0.5,1)}(p)\ , \tag{1}$$

for $d \in (0, 2)$.

(d) Find the posterior density of p given the data x under the prior (1). (All integrals must be evaluated.)

(e) Are there any values of d that lead to Bayes estimators in $\mathcal{C}$?

3. Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ), where $n \geq 2$, and let $X = (X_1, \ldots, X_n)$.

(a) Find the ML estimator of λ, call it $\hat{\lambda}(X)$. Is the ML estimator unbiased?

(b) Find the Cramér-Rao lower bound for the variance of an unbiased estimator of $h(\lambda) = \lambda e^{-\lambda}$.

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

(d) Either *prove* that $\hat{h}(X)$ is the best unbiased estimator of $h(\lambda)$ or *find* the best unbiased estimator of $h(\lambda)$.

4. 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, β is an unknown regression parameter, and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathrm{N}(0, \sigma^2)$ with σ^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 $V \sim \mathrm{N}(\mu, \tau^2)$.

(b) Suppose that $W_1, \ldots, W_m$ are independent random variables such that $W_i \sim \mathrm{N}(\mu_i, \tau_i^2)$. 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 β, 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 $Z \sim \mathrm{N}(0, 1)$.)

5. Let $X_1, \dots, X_n$ be iid discrete uniform on $\{1, 2, \dots, \theta\}$; that is, the common mass function is given by

$$f(x|\theta) = \theta^{-1} I_{\{1,2,\dots,\theta\}}(x) \ ,$$

where $\theta \in \Theta = \mathbb{N}$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_0 < \theta_1$. The relevant sample space, $\mathcal{X}$, is given by

$$\mathcal{X} = \Big\{ (x_1, x_2, \dots, x_n) : x_i \in \{1, 2, \dots, \theta_1\} \ \text{ for each } \ i = 1, 2, \dots, n \Big\} \ .$$

There are θ_1^n elements in $\mathcal{X}$; that is, $\#(\mathcal{X}) = \theta_1^n$. Recall that a test is nothing more than a partition of $\mathcal{X}$ into the rejection region and the acceptance region. Since there are only a finite number of sample points, there are only a finite number of possible tests.

(a) Define a set as follows

$$S = \Big\{ (x_1, x_2, \dots, x_n) \in \mathcal{X} : \max_{1 \le i \le n} x_i > \theta_0 \Big\} \ .$$

Let $P_{\theta_0}(R)$ and $P_{\theta_1}(R)$ denote the size and power of the test with rejection region R. Show that $P_{\theta_0}(R)$ and $P_{\theta_1}(R)$ can both be written as simple functions of n, θ_0, θ_1, R and S.

(b) *Without appealing to the NP Lemma*, prove that if $S \subset R$, then R is a most powerful test of its size.

(c) Suppose that the set S

Q.6. Consider the (scalar based) simple linear regression model through the origin:

$$Y_i = \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim NID\left(0,\sigma^2\right) \quad i = 1,\ldots,n$$

$$\text{Define:} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad \text{and} \qquad \tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$$

p.1.a. Derive: $E\left(\hat{\beta}_1\right)$ and $V\left(\hat{\beta}_1\right)$ SHOW ALL WORK.

p.1.b. Derive: $E\left(\tilde{\beta}_1\right)$ and $V\left(\tilde{\beta}_1\right)$ SHOW ALL WORK.

p.1.c. Which estimator of β_1 has the smaller variance? Why? SHOW ALL WORK. (Hint: $S_{XX} = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 > 0$)

p.1.d. Derive: $E\left[SSE\left(\hat{\beta}_1\right)\right] = E\left[\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 X_i\right)^2\right]$

Q.7. A balanced 2-Way Analysis of Variance was conducted as a Completely Randomized Design (where each experimental unit receives exactly 1 combination of the levels of fixed factors A and B). The following table gives the mean and (standard deviation) of the *r* = 4 replicates for each of the *ab* = 2(3) = 6 treatments, for the model:

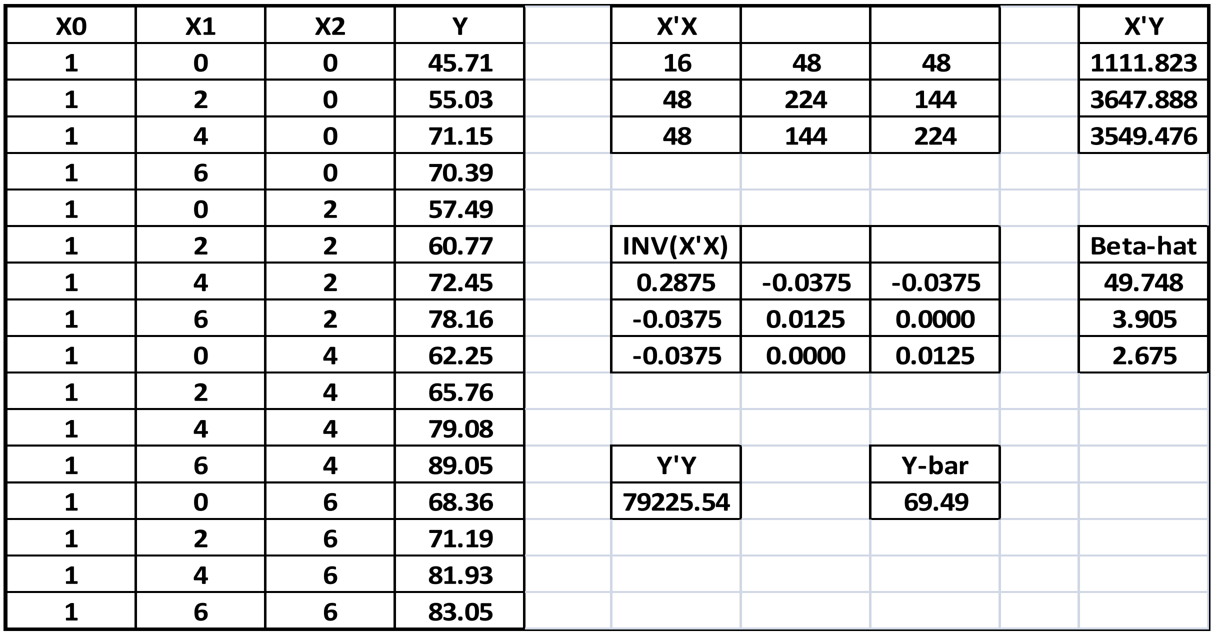
$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \quad i = 1,2;\ j = 1,2,3;\ k = 1,\ldots,4$$

$$\sum_{i=1}^{2}\alpha_i = \sum_{j=1}^{3}\beta_j = \sum_{i=1}^{2}(\alpha\beta)_{ij} = \sum_{j=1}^{3}(\alpha\beta)_{ij} = 0 \quad e_{ijk} \sim NID\left(0,\sigma^2\right)$$

Q.8. A multiple regression model relating *Y* to 2 independent variables (X_1 and X_2) is fit, based on *n*=16 observations. The model fit is:

Scalar-form: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad \varepsilon_i \sim NID\left(0,\sigma^2\right) \quad i = 1,\ldots,n$

Matrix-Form: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0},\sigma^2\mathbf{I}\right)$



X0	X1	X2	Y		X'X				X'Y
1	0	0	45.71		16	48	48		1111.823
1	2	0	55.03		48	224	144		3647.888
1	4	0	71.15		48	144	224		3549.476
1	6	0	70.39						
1	0	2	57.49						
1	2	2	60.77		INV(X'X)				Beta-hat
1	4	2	72.45		0.2875	-0.0375	-0.0375		49.748
1	6	2	78.16		-0.0375	0.0125	0.0000		3.905
1	0	4	62.25		-0.0375	0.0000	0.0125		2.675
1	2	4	65.76						
1	4	4	79.08						
1	6	4	89.05		Y'Y		Y-bar		
1	0	6	68.36		79225.54		69.49		
1	2	6	71.19						
1	4	6	81.93						
1	6	6	83.05						

Compute or obtain the following quantities from the matrix results above:

p.8.i. Total Sum of Squares (Uncorrected) and corresponding degrees of freedom.

p.8.ii. Sum of Squares for the Intercept (μ) and corresponding degrees of freedom.

p.8.iii. Total Sum of Squares (Corrected) and corresponding degrees of freedom.

p.8.iv. Sum of Squares for Regression and corresponding degrees of freedom.

p.8.v. Sum of Squares for Residual and corresponding degrees of freedom.

p.8.vi. Test Statistic and Rejection Region for testing: $H_0 : \beta_1 = \beta_2 = 0$ (α = 0.05 significance level)

p.8.vii. Test Statistic and Rejection Region for testing: $H_0 : \beta_1 = 0$ (α = 0.05 significance level, 2-sided)

p.8.viii. 95% Confidence Interval for β_2

p.8.ix. Test Statistic and Rejection Region for testing: $H_0 : \beta_1 = \beta_2$ (α = 0.05 significance level, 2-sided)

p.8.x. Coefficient of Determination (R^2)

p.8.xi. Fitted value when $X_1 = X_2 = 3$

Q.9. A study was conducted to measure the variation in endurance within and across females who play college soccer at major U.S. colleges and universities. A random sample of t = 10 players are sampled (from the population of a female college soccer players), and each player is measured r = 3 times, where the response y is the time to exhaustion on a treadmill at a 10-degree slope at 10 kilometers per hour. The model fit is:

$$y_{ij} = \mu + a_i + e_{ij} \quad i = 1,...,t;\ j = 1,...,r$$

$$a_i \sim NID\left(0, \sigma_a^2\right) \quad e_{ij} \sim NID\left(0, \sigma^2\right) \quad \left\{a_i\right\} \perp \left\{e_{ij}\right\}$$

p.9.a. Derive $E\left(y_{ij}\right), \quad V\left(y_{ij}\right), \quad Cov\left(y_{ij}, y_{ij'}\right), \quad Cov\left(y_{ij}, y_{i'j}\right) \quad Cov\left(y_{ij}, y_{i'j'}\right) \quad i \neq i'\ j \neq j'$

p.9.b. Derive the expected values of the Among and Within players mean squares, where the respective sums of squares are (show all work):

$$SSA = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}\right)^2 \qquad SSW = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \overline{y}_{i\bullet}\right)^2$$

p.9.c. For this study, SSA = 2700 and SSW = 400. Give point estimates of σ_a^2 and σ^2

p.9.d. The intra-class correlation is defined as: $\rho_I = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}$. Give an estimate of this correlation based on this data (note that this will not necessarily be an unbiased estimate).

Q.10. A randomized complete block design is conducted to compare *t* = 3 treatments in *r* = 6 blocks. The model is fit, based on fixed treatments and random blocks (note that standard errors of individual treatment means depend on whether blocks are fixed or random, but standard errors of differences of pairs of means do not). The following model is fit:

$$y_{ij} = \mu + \tau_i + b_j + e_{ij} \quad i = 1,...,t;\ j = 1,...,r$$

$$\sum_{i=1}^{t} \tau_i = 0 \quad b_j \sim NID\left(0, \sigma_b^2\right) \quad e_{ij} \sim NID\left(0, \sigma^2\right) \quad \left\{b_j\right\} \perp \left\{e_{ij}\right\}$$

The following table gives experimental results:

p.10.a. Compute the Treatment Sum of Squares and its corresponding degrees of freedom.

p.10.b. Compute the Block Sum of Squares and its corresponding degrees of freedom.

p.10.c. Compute the Error Sum of Squares and its corresponding degrees of freedom.

p.10.d. Test: $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ at the α = 0.05 significance level.

p.10.e. Use Bonferroni's method to obtain simultaneous 95% confidence intervals for: $\tau_1 - \tau_2, \quad \tau_1 - \tau_3, \quad \tau_2 - \tau_3$