First Year Examination
Department of Statistics, University of Florida

May 6, 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
- pdf = probability density function

You may use the following facts/formulas without proof:

Beta density: $X \sim$ Beta(α, β) 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(α, β) 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 $\beta > 0$. The mgf of X is given by $M_X(t) = (1-\beta t)^{-\alpha}$. Also, $Y \sim$ Exp(β) is equivalent to $Y \sim$ Gamma$(1, \beta)$.

Distributional Result: If $X \sim$ Gamma(α_x, β) and $Y \sim$ Gamma(α_y, β) and X and Y are independent, then $X/(X+Y) \sim$ Beta(α_x, α_y).

1. Suppose that (X, Y) is a continuous bivariate random vector with support $\mathcal{A} = (0, 1) \times (0, 1)$; that is, the support is the unit square centered at the point $(1/2, 1/2)$. Set $U = 2X + Y$ and $V = X - 2Y$.

(a) What is the (marginal) support of the random variable U?

(b) What is the (marginal) support of the random variable V?

(c) Neatly sketch a graph showing the (joint) support of the bivariate random vector (U, V). (Hint: Think about indicator functions.) Clearly label the axes as well as any functions that you draw.

For the remaining parts of this problem, assume that X and Y are iid Uniform$(0, 1)$.

(d) Find $f_{U,V}(u, v)$.

(e) Find the marginal pdf of U.

(f) Find the conditional pdf of V given $U = u$.

2. 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 $g : \mathbb{R} \to [0, \infty)$ be another pdf such that $\mathsf{X} \subset \{x \in \mathbb{R} : g(x) > 0\}$. Suppose that we have a positive number M such that, for all $x \in \mathsf{X}$, $f(x) \le Mg(x)$. Consider Algorithm I:

```
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 Algorithm I is a draw from f. (Hint: The fact that g has a larger support than f doesn't really change anything.)

Now suppose that $h : \mathbb{R} \to [0, \infty)$ is a pdf and that $S \subset \mathbb{R}$ is such that $0 < \int_S h(y)\, dy < 1$. Consider a truncated version of h given by

$$h^*(x) = c\, h(x)\, I_S(x) \ ,$$

where $c = \left[\int_S h(y)\, dy \right]^{-1}$ is a known normalizing constant. Consider Algorithm II:

```
1. Draw $Y \sim h(\cdot)$.
2. If $Y \in S$, then accept $Y$ and stop; otherwise, return to Step 1.
```

(b) Prove that the output of Algorithm II is a draw from h^* by showing that Algorithm II is a special case of Algorithm I.

Suppose we want to simulate from the pdf given by

$$de^{-x} I_{S_e}(x) + de^{-x/2} I_{S_o}(x) \ , \tag{1}$$

where $S_e = \cup_{i=0}^{\infty} \big(2i, 2i + 1\big]$, $S_o = \cup_{i=0}^{\infty} \big(2i + 1, 2(i + 1)\big]$, and d is a normalizing constant.

(c) Find a closed-form expression for d.

(d) Show that the pdf in (1) can be written as a mixture of two pdfs whose supports are mutually exclusive.

(e) Construct an algorithm for simulating random variables from the pdf in (1). (As usual, the only random variables that you are allowed to use in your algorithm are those from the Uniform$(0, 1)$ distribution.)

3. Suppose that X and Y are identically distributed random variables and that their common distribution is Bernoulli(p), where $p \in (0,1)$. Suppose that $P(X = Y = 1) = \theta$.

(a) Show that the Bernoulli(p) family is complete when the parameter space is $(0,1)$.

(b) Prove that $\theta \leq p$.

(c) Is θ bounded below?

(d) Find the correlation of X and Y.

(e) Suppose that $g(X)$ is a random variable that has mean 0 and variance 1. Identify the function $g(\cdot)$.

Let U and V be two random variables. The *maximal correlation* of U and V is defined as

$$\sup_{g,h} \operatorname{cov}\big(g(U), h(V)\big) ,$$

where the $\sup$ ranges over all functions g and h such that $g(U)$ and $h(V)$ both have mean 0 and variance 1.

(f) Find the maximal correlation of X and Y, and compare it to the correlation of X and Y.

4. Let $X_1, \dots, X_n$ be iid Exp(λ), where $n \geq 2$. As usual, let $X = (X_1, \dots, X_n)$. Also, let $h(\lambda) = e^{-\lambda}$, which is the probability that an Exp(λ) random variable exceeds λ^2.

(a) What is the distribution of $Y = X_2 + X_3 + \cdots + X_n$?

(b) Write down the joint pdf of X_1 and Y.

(c) Find the joint pdf of X_1 and $X_1 + Y$.

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

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

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

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

5. Suppose that $X_1, \dots, X_n$ are iid Beta$(\gamma, 1)$ and that $Y_1, \dots, Y_m$ are iid Beta$(\theta, 1)$. As usual, let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$. Assume further that X and Y are independent. We will consider testing $H_0 : \gamma = \theta$ versus $H_1 : \gamma \neq \theta$ using the statistic

$$T = \frac{\sum_{i=1}^n \log X_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i} \; .$$

(a) *Derive* the distribution of $Z = -\gamma \log X_1$.

(b) *Identify* the distribution of T under H_0.

(c) Show that the ML estimator of γ is $\hat{\gamma}(X)$ where

$$\hat{\gamma}(x) = \frac{-n}{\sum_{i=1}^n \log x_i} \; .$$

Obviously, the ML estimator of θ is $\hat{\theta}(Y)$ where $\hat{\theta}(y) = -m / \sum_{i=1}^m \log y_i$.

(d) Construct the LRT statistic for testing $H_0 : \gamma = \theta$ against $H_1 : \gamma \neq \theta$.

(e) Use the fact that $\prod_{i=1}^n x_i = \exp\left\{ \sum_{i=1}^n \log x_i \right\}$ to simplify the following expression

$$\frac{\left[\left(\prod_{i=1}^n x_i \right) \left(\prod_{i=1}^m y_i \right) \right]^{\hat{\gamma}_0(x,y)-1}}{\left(\prod_{i=1}^n x_i \right)^{\hat{\gamma}(x)-1} \left(\prod_{i=1}^m y_i \right)^{\hat{\theta}(y)-1}} \; ,$$

where

$$\hat{\gamma}_0(x,y) = \frac{-(n+m)}{\sum_{i=1}^n \log x_i + \sum_{i=1}^m \log y_i} \; .$$

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

(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 $n = 23$ and $m = 12$. Explain exactly how you would find the rejection region of the size 0.10 LRT in terms of T.

Q.6. A study was conducted that measured total lipids (Y) among meats of 3 types (red meat, poultry, and fish). Red meat was made up of 2 varieties (beef and pork); poultry consisted of 3 varieties (chicken, duck, and ostrich); fish had 2 varieties (whiting and mackerel). For the purposes of this study, consider both meat type (Factor A) and variety (Factor B) to be fixed factors. Consider the following model:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijk} \quad i = 1,2,3 \quad j = 1,\ldots,b_i \quad k = 1,2,3 \quad \sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{b_i} \beta_{j(i)} = 0 \quad e_{ijk} \sim NID\left(0,\sigma^2\right)$$

The following table of means and standard deviations (SD), based on 3 replicates per variety type was given by the authors.

Meat	Variety	Variety Mean	Meat Mean	Overall Mean	Variety SD
Red	Beef	11.3	10.5	10.7	3.6
Red	Pork	9.7	10.5	10.7	3.1
Poultry	Chicken	7.7	10.3	10.7	2.3
Poultry	Duck	14.7	10.3	10.7	4.8
Poultry	Ostrich	8.5	10.3	10.7	1.9
Fish	Whiting	3.8	11.5	10.7	1.2
Fish	Mackerel	19.2	11.5	10.7	5.5

p.6.a. Obtain the Analysis of Variance, including all sources of variation, degrees of freedom, sums of squares, and mean squares.

p.6.b. Test H$_0$: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ vs H$_A$: Not all α_i = 0 at the 0.05 significance level by computing the test statistic, stating clearly the rejection region, and giving the appropriate conclusion.

p.6.c. Test H$_0$: $\beta_{1(1)} = \ldots = \beta_{2(3)} = 0$ vs H$_A$: Not all $\beta_{j(i)}$ = 0 at the 0.05 significance level by computing the test statistic, stating clearly the rejection region, and giving the appropriate conclusion.

p.6.d. Compute Bonferroni's and Tukey's minimum significant differences for comparing all pairs of varieties with an experimentwise error rate of 0.05.

Q.7. An experiment was conducted, relating the percentage of female offspring (Y) of black rockfish to the temperature in the breeding tank (X). Since the response is a percentage (and for computational ease), fit the regression model based on the following transformations: $Y' = 100\sqrt{\sin^{-1}\left(\frac{Y}{100}\right)}$ $\quad X' = \frac{X - \bar{X}}{4}$

Consider the following 2 Models (with error terms assumed to be independent, N(0,σ^2):

Model 1: $Y_i' = \beta_0 + \beta_1 X_i' + \varepsilon_i$ and Model 2: $Y_i' = \beta_0 + \beta_1 X_i' + \beta_2 \left(X_i'\right)^2 + \varepsilon_i$

				X2'X2		
				5	0	10
				0	10	0
X=Temp(Cel)	Y=%Female	X'	Y'	10	0	34
10	45.2	-2	68			
14	46.2	-1	69	INV(X2'X2)		
18	50	0	72	0.486	0.000	-0.143
22	63	1	83	0.000	0.100	0.000
26	82.5	2	98	-0.143	0.000	0.071

p.7.a. For model 1, obtain the following matrices, vectors, and scalars (use “centered” X-values)

$$Y, \quad X, \quad X'X, \quad X'Y, \quad \left(X'X\right)^{-1}, \quad \hat{\beta}, \quad \hat{Y}, \quad SS_{\text{Reg}}, \quad SS_{\text{Res},} \quad s^2, \quad \hat{V}\left(\hat{\beta}\right)$$

p.7.b. For Model 2, obtain the following matrices, vectors, and scalars (use “centered” X-values)

$$X, \quad X'Y, \quad \hat{\beta}, \quad \hat{Y}, \quad SS_{\text{Reg}}, \quad SS_{\text{Res}}, \quad s^2, \quad \hat{V}\left(\hat{\beta}\right)$$

p.7.c.i. Obtain a 95% confidence interval for β_2.

p.7.c.ii. Conduct the t-test for testing H_0: $\beta_2 = 0$ versus H_A: $\beta_2 \neq 0$. at the 0.05 significance level by computing the test statistic, stating clearly the rejection region, and giving the appropriate conclusion.

Q.8. Consider the Randomized Complete Block Design with *t* treatments and *r* blocks, and no interaction between treatments and blocks. Assume fixed treatment effects for each model, with fixed blocks for Model 1 and random blocks for Model 2.

$$\text{Model 1: } y_{ij} = \mu + \tau_i + \beta_j + e_{ij} \quad i=1,...,t;\ j=1,...,r \quad \sum_{i=1}^{t}\tau_i = \sum_{j=1}^{r}\beta_j = 0 \quad e_{ij} \sim NID\left(0,\sigma^2\right)$$

$$\text{Model 2: } y_{ij} = \mu + \tau_i + b_j + e_{ij} \quad i=1,...,t;\ j=1,...,r \quad \sum_{i=1}^{t}\tau_i = 0 \quad b_j \sim NID\left(0,\sigma_b^2\right) \quad e_{ij} \sim NID\left(0,\sigma^2\right) \quad \left\{b_j\right\} \perp \left\{e_{ij}\right\}$$

p.8.a. For each model, give the following values:

$$E\left(y_{ij}\right), \quad V\left(y_{ij}\right), \quad Cov\left(y_{ij}, y_{ij'}\right), \quad Cov\left(y_{ij}, y_{i'j}\right), \qquad Cov\left(y_{ij}, y_{i'j'}\right)$$

p.8.b. For each model, derive the following values (show all work):

$$V\left(\overline{y}_{i\bullet}\right), \quad Cov\left(\overline{y}_{i\bullet}, \overline{y}_{i'\bullet}\right), \quad V\left(\overline{y}_{i\bullet} - \overline{y}_{i'\bullet}\right)$$

p.8.c. Suppose an experiment consists of *t* = 4 treatments and *r* = 8 blocks, with:

$$\sum_{i=1}^{t}\left(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}\right)^2 = 3000 \quad \sum_{j=1}^{r}\left(\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet}\right)^2 = 14000 \quad \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \overline{y}_{\bullet\bullet}\right)^2 = 84200$$

p.8.c.i. Obtain the Analysis of Variance, including all sources of variation, degrees of freedom, sums of squares, and mean squares (this will be the same for each model).

p.8.c.ii. Test H$_0$: $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ at the 0.05 significance level (this will be the same for each model).

p.8.c.iii. Give Tukey's and Bonferroni's minimum significant differences for all pairwise comparisons among treatments, with an experiment-wise error rate of 0.05 for each model, based on your results from p.8.b.

Q.9. For the simple regression model (scalar form): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n \quad \varepsilon_i \sim NID\left(0, \sigma^2\right)$

we get: $\overset{\wedge}{\beta}_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{XX}} \right) Y_i, \quad \overline{Y} = \sum_{i=1}^{n} \left(\frac{1}{n} \right) Y_i, \quad \overset{\wedge}{\beta}_0 = \overline{Y} - \overset{\wedge}{\beta}_1 \overline{X}$

p.9.a. Derive $E\left(\overset{\wedge}{\beta}_1\right), \quad E\left(\overline{Y}\right), \quad E\left(\overset{\wedge}{\beta}_0\right)$

p.9.b. Derive $V\left(\overset{\wedge}{\beta}_1\right), \quad V\left(\overline{Y}\right), \quad Cov\left(\overset{\wedge}{\beta}_1, \overline{Y}\right), \quad V\left(\overset{\wedge}{\beta}_0\right), \quad Cov\left(\overset{\wedge}{\beta}_1, \overset{\wedge}{\beta}_0\right)$

p.9.c. Derive $E\left(\hat{Y}_0\right), \quad V\left(\hat{Y}_0\right)$ $\quad$ where $\hat{Y}_0 = \overset{\wedge}{\beta}_0 + \overset{\wedge}{\beta}_1 X_0$ for known X_0

Q.10. A study was conducted to maximize sound transmission loss (y) in non-woven fabrics. Each of 3 factors was set at 5 levels, so that a polynomial response surface could be fit to the data as a response surface.

Description: 3-Factor Experiment conducted to maximize sound transmission
loss in non-woven fabrics. 20 runs.
Factors/Levels:
X_1= Punch Density (Punches/cm^2) -1.682=70,-1=106.5,0=160,1=213.5,1.682=250
X_2= Needle Penetration (mm) -1.682=18,-1=16.4,0=14,1=11.6,1.682=10
X_3 = Mass/Unit Area (g/cm^2) -1.682=300,-1=442,0=650,1=858,1.682=1000

The following 3 models are fit to the coded (-1.682 to 1.682) data, based on n=20 observations:

$$\text{Model 1: } E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3$$

$$\text{Model 2: } E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_{11}X_1^2+\beta_{22}X_2^2+\beta_{33}X_3^2$$

$$\text{Model 3: } E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_{11}X_1^2+\beta_{22}X_2^2+\beta_{33}X_3^2+\beta_{12}X_1X_2+\beta_{13}X_1X_3+\beta_{23}X_2X_3$$

ANOVA	Model 1		Model 2		Model3	
	df	*SS*	*df*	*SS*	*df*	*SS*
Regression		5.41		172.37		178.66
Residual		193.86		26.90		20.60
Total		199.27		199.27		199.27

p.10.a. Controlling for linear and quadratic terms for each factor, test whether there are any interactions among the 3 factors' linear effects. That is, test H_0: $\beta_{12}=\beta_{13}=\beta_{23}=0$ at the 0.05 significance level by computing the test statistic, stating clearly the rejection region, and giving the appropriate conclusion.

p.10.b. Compute R(β_{11}, β_{22}, β_{33} | β_0, β_1, β_2, β_3). Use this to test whether any of the factors' effects are quadratic (given the linear effects, but not interactions). That is, test H_0: $\beta_{11}=\beta_{22}=\beta_{33}=0$ at the 0.05 significance level by computing the test statistic, stating clearly the rejection region, and giving the appropriate conclusion.

p.10.c. The following output gives the regression coefficients and t-tests for Model 2.

p.10.c.i. Give the predicted value for the 6 observations set at the coded values $X_1 = X_2 = X_3 = 0$.

p.10.c.ii. Pure Error is obtained by taking the error sum of squares for these 6 observations around their predicted value (all other observations are at distinct levels of the factors). The 6 observed values are: 16.9, 15.1, 14.5, 15.7, 14.3, 12.4. Compute SS(Pure Error) and SS(Lack-of-Fit) for model 2. Conduct the F-test for Lack-of-Fit for Model 2. Note the degrees of freedom for Pure Error are 6-1 = 5.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	14.79	0.59	25.20	0.0000
punch	-0.49	0.39	-1.26	0.2312
needle	0.12	0.39	0.31	0.7582
mass	0.38	0.39	0.97	0.3500
punch2	1.51	0.38	3.99	0.0015
needle2	-2.76	0.38	-7.28	0.0000

mass2	0.74	0.38	1.96	0.0719