First Year Examination
Department of Statistics, University of Florida
August 21, 2008, 1:00 - 5:00pm

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
- pdf = probability density function
- pmf = probability mass function
- $\mathbb{N} = \{1, 2, 3, \dots\}$
- $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$
- $\mathbb{R}^+ = (0, \infty)$
- N(μ, σ^2) = normal distribution with mean μ and variance σ^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. An experiment comparing three treatments (1, 2, and 3) is conducted in a completely randomized design, with $n = 5$ observations per treatment group, yielding the following summary statistics for the response variable in each treatment group:

Group	Treatment 1	Treatment 2	Treatment 3
Sample Mean	30	22	20
Sample *Standard Deviation*	10	4	10

Answer the following parts, assuming the analysis model

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \epsilon_{ij} \sim \text{iid N}(0, \sigma^2) \qquad i = 1, 2, 3 \qquad j = 1, \ldots, 5$$

where y_{ij} is the response of the j^{th} observation in treatment group i.

(a) Find the ordinary least squares estimates $\widehat{\mu}_1$, $\widehat{\mu}_2$, and $\widehat{\mu}_3$ of μ_1, μ_2, and μ_3. Also compute the usual unbiased estimate of σ^2.

(b) Assuming the restriction $\alpha_1 + \alpha_2 + \alpha_3 = 0$, compute the ordinary least squares estimates $\widehat{\alpha}_1$, $\widehat{\alpha}_2$, and $\widehat{\alpha}_3$ of α_1, α_2, and α_3.

(c) Write an expression for the (true) variance-covariance matrix of the vector $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3)'$ under the model above. Then estimate this matrix using your results from the preceding parts.

(d) Using the model parameters, form a contrast that compares the mean response to treatment 1 with the average of the mean responses to treatments 2 and 3, and is positive when the mean response to treatment 1 exceeds the mean responses to treatments 2 and 3. Is there sufficient evidence in these data to conclude that this contrast is *greater* than zero? (Perform a one-sided test at level $\alpha = 0.05$, under the model above, remembering to state the null and alternative hypotheses.)

(e) Suppose you are planning a new experiment of the same type, with the same three treatments and using the same analysis model equation, but with possibly some other number of observations n in each treatment group. Assuming you obtain the same estimate of σ^2 as you computed in part (a), determine the smallest value of n for which an individual 95% two-sided (equal-tailed) confidence interval for a pairwise comparison of treatment means would have a width no greater than 20.

2. Consider the following model

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

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

$$\textstyle\sum_{j=1}^{b} \beta_j = 0$$

which is often used for an analysis involving two crossed factors, one random and the other fixed.

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

(b) Find the *correlation* between y_{ijk} and $y_{ij'k}$ for $j \neq j'$.

(c) Form an expression, in terms of y_{ijk} values, for the F-test statistic that would be used to test whether this model could be reduced to an *additive* model. (Be sure to fully define any simplified notation you use.) State the null and alternative hypotheses, in terms of the model parameters. What are the degrees of freedom of the null F distribution?

(d) Form an expression, in terms of y_{ijk} values, for the usual unbiased estimator $\hat{\sigma}^2$ of σ^2. (Be sure to fully define any simplified notation you use.) Then find a nonzero constant c, possibly depending on model parameters, such that $c\hat{\sigma}^2$ has a (central) chi-square distribution.

(e) Using the chi-square random variable from the previous part, 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.)

3. Researchers identify four separate tracts of pine forest land at extreme risk for infestation by the mountain pine beetle. In each tract, they choose three healthy, mature lodgepole pine trees. One of these trees is sprayed with a chemical insecticide, another is partially coated with a protective resin, and the third is left untouched as a control. These experimental conditions are assigned to the trees according to a *randomized complete block (RCB) design*, with the tracts as blocks. After one year, the trees are harvested, and the researchers count separately the number of beetle larvae in each of two predetermined sections of trunk on each tree (the sections being of approximately the same size on every tree). The counts are listed in the table below (two counts for each tree):

	Chemical Spray		Resin Coat		Control	
Tract 1	150	130	160	130	230	220
Tract 2	50	70	60	40	160	100
Tract 3	20	80	30	30	120	80
Tract 4	110	110	120	110	180	150

(a) Answer the following very briefly:

(i) What are the *treatments*? How many are there?

(ii) What are the *experimental units*? How many are there (total)?

(iii) What are the *observational units* (also called *measurement units*)? How many are there (total)?

(b) Answer the following parts for the analysis of this data in a RCB design, *using the average of its two counts as the response for each tree*. (You may make all of the usual assumptions of the normal-theory analysis model ordinarily used for RCB designs.)

(i) Compute an ANOVA table (relevant sources, degrees of freedom, sum of squares, mean squares). Then use it to perform a test of whether there are any treatment effects ($\alpha = 0.05$).

(ii) Using the *Bonferroni* method, form 95% simultaneous two-sided (equal-tailed) confidence intervals for *all pairwise comparisons* between the treatment means. Based on these intervals, what do you conclude about the effectiveness of the spray and the resin against beetle infestation?

(c) List two major assumptions of the usual analysis model for data from a RCB design that are probably not satisfied for the data analysis of part (b).

4. Consider a linear model in the general matrix formulation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ is the column vector of dependent variables, $\boldsymbol{X}$ is a known constant matrix with full column rank, $\boldsymbol{\beta}$ is the column vector of regression parameters, and the error vector $\boldsymbol{\epsilon}$ has mean vector $\mathbf{0}$ (all zeros) and variance-covariance matrix $\boldsymbol{V}\sigma^2$.

(a) Write the expression for the *ordinary least squares* estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, in terms of $\boldsymbol{Y}$ and $\boldsymbol{X}$.

(b) Write an expression for the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$, assuming $\boldsymbol{V}$ is the identity matrix.

(c) Derive an expression for the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ that is correct for any $\boldsymbol{V}$.

(d) Is $\widehat{\boldsymbol{\beta}}$ unbiased if $\boldsymbol{V}$ is the identity matrix? Is the same necessarily true if $\boldsymbol{V}$ is *not* the identity matrix? Justify your answers.

(e) Write the expression for the *generalized least squares* estimate $\widehat{\boldsymbol{\beta}}_G$ of $\boldsymbol{\beta}$, assuming the matrix $\boldsymbol{V}$ is known.

(f) Assuming $\boldsymbol{\epsilon}$ has a multivariate normal distribution, specify the distribution of $\widehat{\boldsymbol{\beta}}_G$. (Name the distribution and give its parameter values.)

(g) Under what general assumption about the matrix $\boldsymbol{V}$ would *weighted least squares* estimation of $\boldsymbol{\beta}$ be most appropriate? What would this assumption imply about the errors (the elements of $\boldsymbol{\epsilon}$)?

5. The absolute pressure Y (in kPa) inside a sealed, rigid container of gas is measured (with error) once at each of 20 (exact) temperatures X (in °C), yielding a set of data pairs (X_i, Y_i), $i = 1, \ldots, 20$. Summary statistics for the data are as follows:

$$\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X_i = 10.5 \qquad \sum_{i=1}^{20}(X_i - \overline{X})^2 = 665.0$$

$$\overline{Y} = \frac{1}{20}\sum_{i=1}^{20} Y_i = 101.8 \qquad \sum_{i=1}^{20}(Y_i - \overline{Y})^2 = 72.3$$

$$\sum_{i=1}^{20}(X_i - \overline{X})(Y_i - \overline{Y}) = 216.79$$

Answer the following, expressing all decimal numbers to at least three significant figures.

(a) Write a model equation for the simple linear regression of pressure (kPa) on temperature (°C). Estimate the slope and intercept parameters using ordinary least squares, remembering to write their units.

(b) Compute an ANOVA table (corrected for the mean) for the simple linear regression of part (a). Also compute the coefficient of determination (R^2).

(c) Compute estimates of the variances of your slope and intercept parameter estimates in part (a). Also compute an estimate of their covariance. (Your estimates should be unbiased under the usual simple linear regression model assumptions.)

(d) A kelvin (K) is a unit of *absolute* temperature with the same unit length as a degree Celsius (°C), but with 0°C corresponding to 273.15 K. That is, [K] = [°C] + 273.15.

According to the "ideal gas law," the absolute pressure in the container should be directly proportional to the *absolute* temperature measured in kelvin (K). (That is, there exists a constant c such that the relationship between absolute temperature T and true absolute pressure P is $P = cT$.) Test whether these data are consistent with the ideal gas law (except for random error), assuming the simple linear regression model is correct (and the errors satisfy the usual assumptions). State the null and alternative hypotheses and use level $\alpha = 0.05$.

6. Let (X, Y) be a bivariate random vector with joint pdf $f_{X,Y}(x, y)$. Assume that

$$\big\{(x, y) \in \mathbb{R} \times \mathbb{R} : f_{X,Y}(x, y) > 0\big\} = (1, \infty) \times (1, \infty) \; .$$

Now consider another bivariate random vector, (U, V), defined in terms of (X, Y) as follows: $U = \frac{X}{X+Y}$ and $V = X + Y$.

(a) Let $f_{U,V}(u, v)$ denote the joint pdf of (U, V) and let $S_{U,V}$ denote the support of (U, V); that is,

$$S_{U,V} = \big\{(u, v) \in \mathbb{R} \times \mathbb{R} : f_{U,V}(u, v) > 0\big\} \; .$$

Draw a graph of $S_{U,V}$. (Hint: If you know that $v = 7$, what can you say about u?)

(b) Now find $f_{U,V}(u, v)$ assuming that X and Y are iid and that the pdf of X is $f_X(x) = x^{-2} I_{(1,\infty)}(x)$.

(c) Find closed form expressions for the marginal pdfs of U and V.

(d) Find $P(U > 1/2 \mid V > \pi)$. (You do not have to evaluate the integrals.)

(e) Find $P(U > 1/2 \mid V = \pi)$. (You do not have to evaluate the integrals.)

7. Let $W_1, \ldots, W_k$ be unbiased estimators of a parameter θ with $\text{Var}(W_i) = \sigma_i^2$ and $\text{Cov}(W_i, W_j) = 0$ if $i \neq j$. In this question, we will consider estimating θ with estimators of the form $\sum_{i=1}^k a_i W_i$ where $a_1, \ldots, a_k$ are constants.

(a) Show that $\sum_{i=1}^k a_i W_i$ is unbiased for θ if and only if $\sum_{i=1}^k a_i = 1$.

(b) Find the variance of $\sum_{i=1}^k a_i W_i$.

(c) Let Y be a discrete random variable that takes values in the set $\big\{a_1\sigma_1^2, a_2\sigma_2^2, \ldots, a_k\sigma_k^2\big\}$ with probabilities given by

$$P\big(Y = a_i\sigma_i^2\big) = \frac{c}{\sigma_i^2} \; ,$$

for $i = 1, 2, \ldots, k$, where c is a normalizing constant. Find the value of c.

(d) Find the variance of Y using the formula $\text{Var}(Y) = \text{E}Y^2 - (\text{E}Y)^2$.

(e) *Use the calculation in part (d)* to get a lower bound on the variance of unbiased estimators of the form $\sum_{i=1}^k a_i W_i$.

(f) Find an estimator that achieves the minimum variance identified in part (e).

8. Suppose $X_1, \ldots, X_n$ are iid Poisson(λ); that is,

$$P_\lambda(X_1 = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

for $x \in \mathbb{Z}_+$ and $\lambda \in \mathbb{R}^+$.

(a) Derive the expectation and variance of X_1.

(b) The probability that X_1 lands in the set $\{0, 1\}$ is $g(\lambda) = (\lambda+1)e^{-\lambda}$. Derive the ML estimator of $g(\lambda)$.

(c) Now consider a Bayesian approach using a Gamma(α, β) prior for λ. Find the posterior density of λ, call it $\pi(\lambda|x_1, \ldots, x_n)$.

(d) Find the posterior expectation of $g(\lambda)$, which is a Bayesian estimator of $g(\lambda)$.

(e) Thus far, we have implicitly assumed that we are presented with all n observations at the same time. Imagine now that you receive only one new observation each day. In other words, on day one, you are given the first observation, $X_1 = x_1$. Then on day two, you are given the second observation, $X_2 = x_2$. This continues for n days. Suppose you update the posterior distribution of λ each day. Specifically, after observing $X_1 = x_1$ on day one, you compute the posterior distribution, call it $\pi_1(\lambda|x_1)$, using the Gamma(α, β) prior for λ. Then, after observing $X_2 = x_2$ on day two, you compute the new posterior, call it $\pi_2(\lambda|x_1, x_2)$, using $\pi_1(\lambda|x_1)$ as the prior. You continue in this way for n days, always using the posterior from the previous day as the new prior. Show that $\pi_n(\lambda|x_1, \ldots, x_n)$ is the same as $\pi(\lambda|x_1, \ldots, x_n)$.

(f) Prove or disprove the following statement: "The result in part (e) that $\pi_n(\lambda|x_1, \ldots, x_n) = \pi(\lambda|x_1, \ldots, x_n)$ would fail if a non-conjugate prior were used for λ."

9. 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 $g(p)$.

10. Suppose the random variables Y_{ij}, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ satisfy the *oneway ANOVA assumptions*; that is,

$$Y_{ij} = \theta_i + \varepsilon_{ij} \ ,$$

where the ε_{ij} are independent with $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ and $\theta_1, \theta_2, \ldots, \theta_k$ and σ^2 are all unknown parameters. Consider testing $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$ versus H_1 : Not H_0.

(a) Write down the usual "F-test." (Carefully define any notation that you introduce.)

(b) Derive the LRT.

(c) Prove or disprove the following statement: "The LRT is equivalent to the F-test."