May, 2010 PhD Qualifying Examination
Department of Statistics
University of South Carolina
Part I: Exam Day #1
9:00AM–1:00PM

Instructions: **Choose 2 problems** from problems 1, 2 and 3; and **choose 2 problems** from problems 4, 5 and 6. Indicate clearly which problems you have chosen to be graded. Use separate sheets of paper for each problem. You are allowed to use the computers and the statistical software in the examination room. However, you are **not** allowed to use the Internet, except for the Help Files of the statistical software. Data sets that are needed in some of the problems are contained in the accompanying CD. Provide details in your solutions. You have **four hours** to complete this examination. Good luck.

1. Let X_1 and X_2 be independent random variables with common probability mass function

$$f_X(x|\theta) = -\frac{\theta^x}{x \log(1-\theta)},$$

where $x = 1, 2, ...,$ and where $\theta \in (0, 1)$. Take $\mathbf{X} = (X_1, X_2)'$.

(a) Find the mean and variance of X_1.

(b) Let $I(\cdot)$ denote the usual indicator function. Show that $U(\mathbf{X}) = U = -I(X_1 = 1)$ is an unbiased estimator of $\tau(\theta) = \theta/\log(1-\theta)$.

(c) Argue that $T(\mathbf{X}) = T = X_1 + X_2$ is a complete sufficient statistic for θ.

(d) Find the uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

2. Peter and Paul each has a box containing balls numbered $1, 2, 3, \ldots, N$. Each of them will draw N balls from their respective boxes in succession, and we will say that there is a **match** if at any of the N draws, the numbers in their chosen balls at that draw coincide. For example, if at the 10th draw both of them get a ball numbered 21, then there is a match.

(a) If the drawing of the balls is *with replacement*, what is the probability that a match occurs? Provide a general formula, and evaluate it when $N = 5$.

(b) What happens to the probability in (a) when you let $N \to \infty$?

(c) If the drawing of the balls is *without replacement*, what is the probability that a match occurs? Provide a general formula, and evaluate it when $N = 5$.

(d) What happens to the probability in (c) when you let $N \to \infty$?

(e) Compare the limiting probabilities in (b) and (d). Are they the same? Comment on your results.

3. Let $X_1, \ldots, X_n$ be a random sample from a uniform$(0, \theta)$ distribution, where $\theta > 0$. In turn, suppose that the prior distribution on θ is lognormal(μ_0, σ_0^2), where $\sigma_0 > 0$.

(a) Find the posterior distribution of $\log \theta$.

(b) Suppose that one defines the Bayes estimate for θ as the value that maximizes the posterior distribution of θ. Find this Bayes estimator.

(c) Is the Bayes estimator in part (b) a consistent estimator for θ? If so, prove it. If not, explain why not.

4. Consider the following sequence of 200 zeros and ones:

```
0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 0 0 0 0 0 0 1 1 1 0
0 1 1 0 1 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0 0 0 1 0 0
1 0 1 1 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 0 0 0 0 0
0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0
1 1 1 0 1 1 1 1 0 0 1 0 1 1 1 0 1 1 1 0 1 0 0 1 0
1 1 0 0 1 1 1 0 0 0 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0
1 1 1 1 1 1 0 0 1 1 0 0 0 1 1 0 0 1 0 0 1 1 1 0 0
0 1 0 1 1 1 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0 0 1 0 0
```

Suppose that someone tells you this sequence of 0's and 1's (starting with the first row, read from left to right) is the result of 200 successive independent tosses of a fair coin, with `0` = Tail and `1` = Head. Discuss how you would proceed to *statistically* validate (or invalidate) the claim of this individual. Explain clearly how you would implement your approach.

Note: This data set can be found on the accompanying data disk. It is called `coin.xls`.

5. Consider the two-factor (fixed-effects) ANOVA model for an $a \times b$ factorial experiment with only one observation per cell, represented by

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, 2, \ldots, a, j = 1, 2, \ldots, b,$$

and $\epsilon_{ij} \sim$ i.i.d. N$(0, \sigma^2)$, where σ^2 is *unknown*. Note that under this model, there is no $A \times B$ interaction.

(a) Consider the single observation in cell (i, j): Y_{ij}. What is $E(Y_{ij})$? What is $var(Y_{ij})$?

(b) Now consider the fitted value defined as $\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$ for this model. Write $\hat{Y}_{ij}$ explicitly as a function of the ab Y_{ij} values.

(c) Show that $\hat{Y}_{ij}$ is an unbiased estimator of the population cell mean μ_{ij}.

(d) For the special case of $a = 2, b = 3$, derive $var(\hat{Y}_{11})$. Based on this, also indicate a formula for $var(\hat{Y}_{ij})$ for $a = 2$ and $b = 3$.

(e) Based on your answers to (a), (c), and (d), discuss whether the single observation or the fitted value is a better estimator of the population cell mean μ_{ij}.

(f) In light of the model and your conclusion in part (e), suggest a *complete* formula for a $100(1 - \alpha)\%$ confidence interval for the population cell mean μ_{ij} when $a = 2$ and $b = 3$. Make sure to carefully define and express *all* terms in your formula so that a practitioner could compute the CI directly, having only the observed data (and any necessary tables of critical values).

6. Consider the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where the *only* assumptions are that ϵ has mean vector $\mathbf{0}$ and covariance matrix Σ, where $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. The vector of weighted-least-squares (WLS) estimates $\mathbf{b}_w$ can be written as

$$\mathbf{b}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y},$$

where $\mathbf{Y}$ is the vector of observed response values, $\mathbf{X}$ is the usual model matrix (including a column of 1's) and $\mathbf{W}$ is a diagonal matrix with diagonal elements $w_1, w_2, \ldots, w_n$.

(a) Suppose there is only one predictor variable, say, X. Show via explicit calculations that the WLS estimated intercept and slope are

$$b_{w0} = \frac{\sum w_i X_i^2 \sum w_i Y_i - \sum w_i X_i \sum w_i X_i Y_i}{\sum w_i \sum w_i X_i^2 - (\sum w_i X_i)^2}$$

and

$$b_{w1} = \frac{\sum w_i \sum w_i X_i Y_i - \sum w_i X_i \sum w_i Y_i}{\sum w_i \sum w_i X_i^2 - (\sum w_i X_i)^2}$$

(b) If (in simple linear regression) $w_i = 1/\sigma_i^2$ and the error variances σ_i^2 are known merely to be proportional to the predictor values X_i, simplify the formulas for b_{w0} and b_{w1} as much as possible.

(c) Assume this model stated in (b) applies to an industrial experiment with 9 observations in which the predictor is set to the following values: $X_1 = X_2 = X_3 = 1$, $X_4 = X_5 = X_6 = 2$, $X_7 = X_8 = X_9 = 3$. Write the vector $\mathbf{b}_w$ as a function only of $Y_1, Y_2, \ldots, Y_9$.

(d) Suggest a method to formally test whether Y and X have a linear association in the model in part (c). Outline your testing procedure as completely as possible.