May, 2010 PhD Qualifying Examination
Department of Statistics
University of South Carolina
Part II: Exam Day #2
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. One of the most popular problems in statistics is estimating p from an iid sample $Y_1, Y_2, ..., Y_n$ of Bernoulli(p) random variables. For this problem, the sufficient statistic is $T = \sum_{i=1}^{n} Y_i$, and define the sample proportion $\widehat{p} = T/n$. It is well known that

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$, where $\sigma^2 = p(1-p)$ and that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \tag{1}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile from the standard normal distribution, is an approximate $100(1-\alpha)$ percent confidence interval for p. The inconsistencies associated with this interval are well documented, so we will find an alternate interval for p using a variance stabilizing transformation. Suppose that h is a real differentiable function, and write $h(\widehat{p})$ in a first-order Taylor series expansion about p; i.e.,

$$h(\widehat{p}) \approx h(p) + h'(p)(\widehat{p} - p),$$

where $h'(p) = \partial h(p)/\partial p$.

(a) Show that $E[h(\widehat{p})] \approx h(p)$ and $\text{var}[h(\widehat{p})] \approx [h'(p)]^2\sigma^2/n$.
(b) Consider setting $\text{var}[h(\widehat{p})] = c_0$, a constant which is free of p. Show that the function

$$h(p) \equiv \sin^{-1}(\sqrt{p})$$

has variance which is free of p.
(c) Prove that

$$\sqrt{n}[h(\widehat{p}) - h(p)] \xrightarrow{d} \mathcal{N}(0, 1/4),$$

as $n \to \infty$, and use this fact to argue that

$$\left(\sin^2\left[h(\widehat{p}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right],\ \sin^2\left[h(\widehat{p}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right]\right) \tag{2}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p.
(d) If you wanted to perform a simulation study to compare the intervals (1) and (2), on what grounds would you compare them? List three factors.
(e) Describe two other confidence interval procedures you could use to estimate p (different than the two confidence intervals in this problem).

2. The time to failure of a machine component satisfies the condition that the conditional survivor function, given $Z = z$, is

$$S_{T|Z}(t|z) = \Pr\{T > t|Z = z\} = [1 - G(t)]^z,$$

where Z is a random variable with density function

$$g_Z(z) = \frac{\alpha^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp\{-\alpha z\} I\{z \geq 0\},$$

$G(\cdot)$ is some continuous distribution function, and $I\{\cdot\}$ is the usual indicator function.

(a) Obtain an expression of the marginal survivor probability function of T, defined via

$$S_T(t) = \Pr\{T > t\}.$$

(b) Specialize to the case where $G(t) = [1 - \exp(-\theta t)]I\{t \geq 0\}$. In this case, obtain the marginal density function of T, $f_T(t)$.

(c) The hazard rate function of T is defined to be

$$h(t) = \frac{f_T(t)}{S_T(t)}.$$

Obtain an expression of this hazard rate function.

(d) Sketch this function and describe its shape (that is, is it constant, increasing, or decreasing, etc.).

3. Let $X_1, \ldots, X_n$ $(n \geq 2)$ be a random sample from a distribution with pdf $f(x|\theta)$. It is known that this distribution can only be one of $N(0,\, \theta^2)$ and Laplace$(0,\, \theta/\sqrt{2})$. It is of interest to estimate θ and also to decide which of these two distributions the observed sample is from. In other words, one may view $f(x|\theta)$ as $g(x|\theta, j)$, where

$$\begin{aligned} g(x|\theta,1) &= \frac{1}{\sqrt{2\pi}\theta}\exp\{-x^2/(2\theta^2)\}, \quad -\infty < x < \infty \\ g(x|\theta,2) &= \frac{1}{\sqrt{2}\theta}\exp(-\sqrt{2}|x|/\theta), \quad -\infty < x < \infty \end{aligned}$$

and one is interested in making inference about (θ, j), where $\theta > 0$ and $j = 1$ or 2.

(a) Provide the second moment of the distribution identified by $g(x|\theta, 1)$. Repeat for $g(x|\theta, 2)$.

(b) Derive the MLE for $(\theta,\, j)$.

(c) Describe a size-α likelihood ratio test for testing $H_0 : j = 1$ versus $H_1 : j = 2$. Describe how you would approximate the critical value(s) for the test.

4. A researcher was studying the relationship between hours of sleep per day and brain weight, body weight, and gestation time. In the regression of Y on x_1, x_2, and x_3, where

$$\begin{aligned} Y &= \text{hours of sleep per day (SLEEP)} \\ x_1 &= \text{gestation time (GEST)} \\ x_2 &= \text{brain weight (BRAIN)} \\ x_3 &= \text{body weight (BODY)}, \end{aligned}$$

SAS PROC GLM produced the following output:

General Linear Model Procedure

Source	DF	Sum of Squares
Model	3	7.5231
Error	50	7.3624
Corrected Total	53	14.8855

Source	DF	Type I SS	Type III SS
GEST	1	_____	_____
BRAIN	1	0.3017	0.1316
BODY	1	_____	0.4249

Parameter	Estimate	Std Error of Estimate
INTERCEPT	15.2947	0.3160
GEST	-0.2685	0.0844
BRAIN	0.0793	0.0839
BODY	-0.1019	0.0600

(a) Write out a linear model to describe the data from this study. Define all quantities and state clearly your assumptions.

(b) Fill in the three blanks in the Type I (sequential) and III (partial) SS output.

(c) Test the hypothesis that there is no additional relationship between BODY and SLEEP, given that GEST and BRAIN are in the model. Use a one-sided alternative that there is a negative relationship. Use $\alpha = 0.05$. State your conclusion.

(d) Test the hypothesis that BRAIN and BODY weight together have no effect on SLEEP, given that GEST is in the model. Use $\alpha = 0.05$. State your conclusion.

(e) In the light of your result from part (d), compute an estimate of $\text{var}(Y|x_1, x_2, x_3)$ when GEST $= 1$, BRAIN $= 3$, and BODY $= 70$.

5. You have been contacted by a zoologist, who is studying an enzyme found in the brain tissues of a particular species of snake. As he explains to you, when tissue from the snake comes in contact with a certain substrate (i.e., a substance upon which the enzyme will act), a kinetic reaction takes place. The reaction may be characterized by the rate at which a certain protein is produced; the greater the concentration of the substrate, the higher the rate of production. The zoologist is interested in the relationship between the rate of production and the concentration of substrate as a way of understanding the reaction. The zoologist shows you the data below. Tissue from several snakes was combined, samples of the combined tissue were subjected to difference concentrations of substrate, one sample per concentration, and the reaction rate was recorded for each.

Substrate concentration (μM)	Reaction rate (nM/mg/hour)
31.25	53.01
62.5	81.42
125	122.11
250	304.57
500	376.87
1000	414.13
2000	553.46

The zoologist would like to fit a formal model to his data to characterize the relationship. A standard model for this type of phenomenon is the so-called Michaelis-Menten model, which represents mean reaction rate as

$$E(Y|x) = f(x, \boldsymbol{\beta}) = \frac{\beta_1}{1 + \beta_2/x}, \tag{1}$$

where x denotes the substrate concentration. A more general version of this model is also sometimes considered:

$$E(Y|x) = f(x, \boldsymbol{\beta}) = \frac{\beta_1}{1 + (\beta_2/x)^{\beta_3}}, \tag{2}$$

where the additional parameter $\beta_3 > 0$, sometimes called the "Hill coefficient," from a modeling perspective, gives more flexibility in the shape of the relationship. The zoologist would like to fit whichever of models (1) or (2) is most appropriate for his data and then answer some specific questions, which may be summarized as follows:

- Is the more complicated model (2) required to describe my data, or does the simpler model (1) seem adequate?
- Under the appropriate model, I need a point and interval estimate of the mean reaction rate when the substrate concentration tends to infinity.
- Under the appropriate model, I need a point and interval estimate of the mean reaction rate that occurs at a substrate concentration of 750 μM.

Perform an analysis that addresses these questions. Make sure to hand in all of your code and pertinent output.

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

6. Consider a one-way fixed-effects experiment with $k > 1$ treatments. The following model assumptions are made for the k samples:

$$\begin{aligned}
&\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n} \sim \text{iid } \mathcal{N}(\mu_1, c_1\sigma^2) \\
&\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n} \sim \text{iid } \mathcal{N}(\mu_2, c_2\sigma^2) \\
&\qquad\qquad \vdots \\
&\text{Sample } k\text{:} \quad Y_{k1}, Y_{k2}, ..., Y_{kn} \sim \text{iid } \mathcal{N}(\mu_k, c_k\sigma^2).
\end{aligned}$$

The parameters $\mu_1, \mu_2, ..., \mu_k$ and σ^2 are unknown. Note that the design is balanced; i.e., the number of replications n is the same for each treatment. We make the additional assumptions:

- the samples are independent
- the constants $c_1, c_2, ..., c_k$ are known.

(a) Write a statistical model to describe the data from this experiment. Describe what each term in your model means.
(b) Show that

$$\widehat{\sigma}^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} c_i^{-1} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i+})^2,$$

where $\overline{Y}_{i+} = n^{-1} \sum_{j=1}^{n} Y_{ij}$, is an unbiased estimator of σ^2.
(c) Derive a $100(1-\alpha)$ percent confidence interval for $\theta = a_1\mu_1 + a_2\mu_2 + \cdots + a_k\mu_k$, where $a_1, a_2, ..., a_k$ are constants.
(d) To determine diet quality, male weanling rats were fed diets with $k = 3$ protein levels (low, medium, and high). Each of 15 rats was randomly assigned to one of the three diets, and their weight gain in grams was recorded. The data are below:

Low	Medium	High
3.89	8.54	20.39
4.87	9.32	24.22
3.26	8.76	30.91
2.70	11.30	22.78
4.82	10.45	26.33

Assume that $c_1 = 1$, $c_2 = 2$, and $c_3 = 15$. Based on your result in part (c), find a 90 percent confidence interval that allows you to compare the medium diet group to the average of the low and high diet groups.

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