January 2011 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 *official* documentation (official help files) of the statistical software. You may also view the particular web pages specified within the exam, in order to use the data sets that are needed in some of the problems. Provide details in your solutions. You have **four hours** to complete this examination. Good luck.

1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution with pdf given by $f(x|\theta) = \theta^{-c} c x^{c-1} e^{-(x/\theta)^c} I(x > 0)$, where $c > 0$ is known.

(a) Find the MLE for θ.

(b) Find the UMVUE for θ.

(c) Find the uniformly most powerful (UMP) test of size α for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a positive constant.

2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution specified by the following probability density function (pdf),

$$f(x|\theta) = \theta^{-1} x^{(1-\theta)/\theta} I(0 \le x \le 1), \qquad \text{where } \theta > 0.$$

(a) Show that $T(\mathbf{X}) = -2\sum_{i=1}^{n} \log X_i$ is a minimal sufficient statistic for θ.

(b) Find the distribution of $Y = -2\log X_1$.

(c) Find a two-sided 95% confidence interval for θ based on T.

(d) Argue or prove that the expected length of your confidence interval in part (c) converges to zero as $n \to \infty$.

3. Let $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$ with X and Y independent.

(a) Characterize the collection of all real constants c_1, c_2, c_3, c_4 such that the random variables

$$L_1 = c_1 X + c_2 Y \quad \text{and} \quad L_2 = c_3 X + c_4 Y$$

are independent.

(b) Are the random variables

$$V = \frac{(X+Y)}{\sqrt{2}\sigma} \quad \text{and} \quad W = \frac{(X-Y)}{\sqrt{2}\sigma}$$

independent?

(c) Under the case $\mu_1 = \mu_2$, determine the marginal distributions of V and W.

(d) Under the case $\mu_1 = \mu_2$, obtain the conditional distribution of

$$P = \frac{XY}{\sigma^2}$$

given that

$$T = \frac{(X+Y)^2}{2\sigma^2} = t.$$

4. The following (edited) output is from a SAS run of PROC GLM to analyze a randomized complete block design. The response variable StemLength measures the length of flowers grown in soil for a fixed length of time. The treatment factor, Type, represents which type of soil was used for the flower. The blocking factor, Region, represents the region of the experimental field in which the flower was planted. Assume that the usual ANOVA model assumptions hold.

```
                                     The GLM Procedure
Dependent Variable: StemLength

                R-Square     Coeff Var      Root MSE    StemLength Mean
                0.878079      3.939745      1.282668           32.55714

    Source                  DF        Type I SS      Mean Square     F Value     Pr > F
    Region                   2       39.0371429       19.5185714       11.86     0.0014
    Type                     6      103.1514286       17.1919048       10.45     0.0004

    Source                  DF      Type III SS      Mean Square     F Value     Pr > F
    Region                   2       39.0371429       19.5185714       11.86     0.0014
    Type                     6      103.1514286       17.1919048       10.45     0.0004

                                         Sum of
    Source                  DF          Squares      Mean Square     F Value     Pr > F
    Corrected Total         20      161.9314286
```

(a) Carefully state the model equation and the ANOVA model assumptions for this analysis.

(b) Does this represent a balanced design or an unbalanced design? How do you know?

(c) Suppose the researcher wanted to formally test the null hypothesis that the mean stem length was equal for all the treatment-block combinations (all the cells of the design). Conduct this test, giving the rejection region (using $\alpha = 0.05$), test statistic value, and conclusion.

(d) Prior to the experiment, it was of interest to compare (1) the Compost type of soil to the non-Compost types and (2) the Clarion type to the Webster type, in terms of mean stem length. Note that the sample mean stem length for the Compost type was 29.67, for the non-Compost types was 33.04, for the Clarion type was 32.17, and for the Webster type was 31.1. Perform the desired tests, keeping the family significance level for the family of tests at no more than 0.05.

(e) Suppose the experimenter had set up the experimental conditions in exactly the same way, and the data were exactly the same as well, but suppose that the blocking factor "Region" had been ignored in the analysis. If this were the case, conduct the appropriate hypothesis test to determine whether the different types of soil produce the same mean stem length. Give the rejection region (using $\alpha = 0.05$), test statistic value, and conclusion.

5. In an experiment to determine the efficacy of a "quick acting" experimental steroid inhaler, one of two asthma medications was administered to $n = 138$ adult male volunteers. Each participant was assigned an asthma severity score on continuous scale from 0 (no asthma) to 20 (persistent, debilitating asthma), and an allergy severity index taking on values 1 (no allergies), 2 (mild, infrequent allergies), 3 (severe, frequent allergies), and 4 (constant, debilitating allergies). In the study, patients were cleared of asthmatic symptoms either by oral steroids or albuterol and given a standard preventative inhaler (treatment 1), or the experimental inhaler (treatment 2). The patient then recorded the first instance of moderate to severe wheezing and the time to first wheezing was computed in hours.

You are to analyze these data keeping in mind the implied goal of the experiment. In particular, you are to determine if there is a treatment difference in the presence of the other concomitant variables. Clearly state your final model, any treatment differences you find, and concisely summarize covariate effects, influential and/or outlying observations, and model suitability. Provide a predictive function for mean time to wheezing given the important variables.

Note: The raw data set can be found at the web page:
`http://www.stat.sc.edu/~hitchcock/wheezing.txt`
In addition, a SAS DATA step that will read in these data can be found at the web page:
`http://www.stat.sc.edu/~hitchcock/wheezingDATAstep.txt`

6. For regression data $(x_i, Y_i), i = 1, \ldots, n$, assume the model

$$Y_i \overset{ind.}{\sim} \text{Poisson}(x_i \beta), \quad i = 1, \ldots, n.$$

The $x_1, \ldots, x_n$ are univariate and strictly positive. Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

(a) Show that the MLE of β is $\hat{\beta} = \bar{Y}/\bar{x}$.

(b) Find the mean and variance of $\hat{\beta}$.

(c) Now assume that β has a gamma prior distribution $\beta \sim \Gamma(wb_0, w)$ where b_0 is our prior best guess and $w > 0$ is a weight attached to this guess. To be precise, β has the prior density

$$f(\beta) = \frac{w^{wb_0}}{\Gamma(wb_0)} \beta^{wb_0 - 1} \exp(-w\beta) I_{(0,\infty)}(\beta).$$

Find the posterior distribution of $\beta | \mathbf{Y}$.

(d) Show that the posterior mean is a weighted average of the prior mean and the MLE. What does the posterior mean converge to as $w \to 0+$?

(e) $n = 173$ female horseshoe crabs were sampled and the width of their carapice measured (cm) and number of satellites (nestmates besides their husbands) recorded. SAS PROC MEANS was used to get $\bar{x}$ and $\bar{y}$. The model above was fit in SAS PROC GENMOD using `model satell=width / noint link=identity dist=pois;` giving the following output:

```
                              The MEANS Procedure

Variable       N            Mean         Std Dev         Minimum         Maximum
--------------------------------------------------------------------------------
satell       173       2.9190751       3.1483357               0      15.0000000
width        173      26.2988439       2.1090610      21.0000000      33.5000000
--------------------------------------------------------------------------------

                             The GENMOD Procedure

                               Model Information
                        Distribution              Poisson
                        Link Function            Identity
                        Dependent Variable         satell

                  Analysis Of Maximum Likelihood Parameter Estimates

                                 Standard     Wald 95% Confidence          Wald
Parameter    DF    Estimate       Error           Limits             Chi-Square    Pr > ChiSq
Intercept     0      0.0000      0.0000      0.0000      0.0000               .             .
width         1      0.1110      0.0049      0.1013      0.1207          505.00        <.0001
```

Show how SAS obtains the estimate 0.1110, standard error 0.0049, and 95% confidence interval $(0.1013, 0.1207)$.

January 2011 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 *official* documentation (official help files) of the statistical software. You may also view the particular web pages specified within the exam, in order to use the data sets that are needed in some of the problems. Provide details in your solutions. You have **four hours** to complete this examination. Good luck.

1. Suppose that $X_1, \ldots, X_n$ is an iid sample with size n from the Poisson distribution with mean λ. We are interested in estimating $\theta = P(X_1 = 0) = \exp(-\lambda)$. Consider the following two estimators:

$$T_{n,1} = e^{-\bar{X}_n} \quad \text{and} \quad T_{n,2} = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i=0)},$$

where $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ and I is the indicator function.

(a) Find the asymptotic distribution of $T_{n,1}$.

(b) Find the asymptotic distribution of $T_{n,2}$.

(c) Which estimator is more efficient in estimating θ when a large sample size is available? Show your argument.

2. Let $X_1, X_2, \ldots, X_n$ be IID from an exponential distribution with mean $1/\lambda$ so that their common density function is

$$f(x|\lambda) = \lambda \exp\{-\lambda x\}, x \geq 0.$$

Denote by $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ the order statistics of $X_1, X_2, \ldots, X_n$. Define, for $i = 1, 2, \ldots, n$,

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)})$$

with $X_{(0)} = 0$.

(a) Prove that $D_1, D_2, \ldots, D_n$ are IID from an exponential distribution with mean $1/\lambda$.

(b) Use the result in (a) to find a 'nice' expression of $E(X_{(n)})$.

(c) For a fixed $K \in \{3, 4, \ldots, n\}$, suppose that you are only able to observe the exact values of $X_{(1)}, X_{(2)}, \ldots, X_{(K)}$ and you only know that each of the $X_{(j)}$'s for $j > K$ are at least equal to $X_{(K)}$. Obtain an *unbiased* estimator of λ based on the total-time-on-test (TTOT) statistic

$$T = \sum_{i=1}^{K} X_{(i)} + (n - K)X_{(K)}.$$

(d) Find an expression for the variance of your estimator in (c).

3. Let $X \sim POI(\mu)$ and $Y \sim POI(\nu)$ with X and Y independent, where $POI(\omega)$ means the Poisson probability mass function

$$p(x|\omega) = \frac{\exp(-\omega)\omega^x}{x!}, x = 0, 1, 2, \ldots.$$

The parameter vector (μ, ν) takes values in $\Re_+^2$.

The ultimate goal is to test the pair of composite hypotheses $H_0 : \mu = \nu, \nu \in \Re_+$ versus $H_1 : \mu = 2\nu, \nu \in \Re_+$ based on the observed value of (X, Y).

(a) Find the distribution of X given $S = X + Y = s$. Express this conditional distribution in terms of s and $\rho \equiv \mu/(\mu + \nu)$.

(b) What is this conditional distribution under H_0? How about under H_1?

(c) Use the distributional results in (a) and (b) and the Neyman-Pearson Lemma to construct a (conditional on $S = s$) most powerful size α test of H_0 versus H_1.

(d) Find an expression of the power function of your test in (c). [**Remark:** Note that your power function should *not* depend on any random entity (e.g., S or X). The expression need not be in a really closed or compact form (e.g., it could for instance be an infinite sum). It should also depend on (μ, ν).]

4. A Latin Square design was implemented to determine which of four movies, *Skyline* (A), *Hereafter* (B), *Due Date* (C), or *Megamind* (D), had the greatest appeal to moviegoers. It is conjectured that both the time during the day and the day of the week may affect a movie's appeal. Each movie was shown in one of four time slots over a four-day period to 50 volunteers who were asked upon exiting the theater if they would recommend the movie to a friend. The response is the number out of the 50 moviegoers who *would* recommend the movie to a friend. Note: At each showing a *different* group of 50 people watched the movie. The data are

	Monday	Tuesday	Wednesday	Thursday
11:00 AM	*C* 32	*D* 23	*B* 36	*A* 40
1:00 PM	*B* 33	*A* 36	*C* 31	*D* 22
3:00 PM	*D* 17	*C* 37	*A* 34	*B* 41
5:00 PM	*A* 35	*B* 37	*D* 18	*C* 31

Analyze the data. Be complete. There may be more that one satisfactory approach to modeling these data.

Note: The raw data set can be found at the web page:
`http://www.stat.sc.edu/~hitchcock/movies.txt`
In addition, a SAS DATA step that will read in these data can be found at the web page:
`http://www.stat.sc.edu/~hitchcock/moviesDATAstep.txt`

5. Suppose, in a factory experiment, we fit the usual regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

via least squares for a set of 15 independent observations. Here, the regression model includes an intercept and two (inherently continuous) predictor variables, X_1 and X_2. During this experiment, over the 15 trials, the first predictor variable is set to take values $1, 2, 3, \ldots, 15$. The second predictor is set to take value 0 for the first 5 trials, 1 for the next 5 trials, and 2 for the last 5 trials. However, because of *possible* machine malfunction, we learn that the first response value (i.e., for trial 1) may be contaminated. The true model equation, therefore, may be

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Delta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\Delta}$ is an 15×1 vector containing zeroes except the first component, which is some unknown real number δ. Assume the random error terms follow a normal distribution with mean zero and constant variance σ^2.

(a) Show that the least squares estimator of $\boldsymbol{\beta}$ is in general biased in this situation.

(b) Given the fixed predictor values in this experiment, derive a simplified expression, in terms of δ, for the vector containing the bias of each component of the least squares estimator.

(c) Outline a formal test for whether the first observation is, in fact, contaminated.

(d) If the response values for this experiment are (in time order)

```
32.10 28.47 24.12 27.13 26.75 23.59 28.43 25.90 28.37 32.91 27.34 24.72 27.93 28.27 28.36
```

use your test from part (c) to determine (using $\alpha = 0.10$) whether the first observation is, in fact, contaminated. Also give a point estimate for δ.

Note: This data set can be found at the web page:
`http://www.stat.sc.edu/~hitchcock/factory.txt`

6. Suppose the weights of a set of specimens follow a normal distribution with mean 26.1 ounces and standard deviation 0.05. However, the scale used to weigh the specimens only shows the weight to one decimal place. So the observed weight measurements $X_1, \ldots, X_n$ are each rounded to the nearest tenth.

(a) If X represents a randomly selected weight (rounded to the nearest tenth) from this population, write down a table that gives the probability distribution of X (you may ignore values of X that have probability of zero to three decimal places).

(b) We will take a random sample of $n = 2$ specimens. Let X_1 and X_2 be the weights (rounded to the nearest tenth) of the two specimens, and let $\bar{X} = (X_1 + X_2)/2$. Give the sampling distribution of $\bar{X}$.

(c) Suppose the biologist studying the specimens is aware that their weights follow a normal distribution with population standard deviation 0.05, but she is not sure what the true population mean weight is. Based on $n = 2$ randomly selected observations, she decides to perform an ordinary Z-test (with nominal $\alpha = 0.05$) of $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$, where μ_0 is a constant that is of interest to the biologist. (Thus she will use the standard rejection region $Z > 1.645$, where Z is the usual Z-statistic.) Assuming the true $\mu = 26.1$, calculate the power of this Z-test in terms of μ_0. Give a plot of the power function (as a function of μ_0). Based on this plot, what would you say the actual (not nominal) significance level of the test is?

(d) Write a small simulation study to check the empirical Type I error rate of the Z-test in part (c). Turn in your code and your calculated Type I error rate. How well does the empirical Type I error rate agree with your answer(s) to part (c)?