May 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 and write your Candidate Number on each sheet, but do not include your name.

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 datasets that are needed in some of the problems.

A formula sheet is included. Some SAS and R macros are available at

`http://www.stat.sc.edu/∼dryden/qualifier-day1`

Provide details in your solutions. You have **four hours** to complete this examination. Good luck.

1. Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) according to $N(\mu, 1)$ with an unknown μ. Suppose that one forgets to record $X_1, \ldots, X_n$ in a study and only records $Y_i = I(X_i < 0)$, for $i = 1, \ldots, n$.

   (a) Find the MLE of μ based on the observed data, $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

   (b) Is $\sum_{i=1}^n Y_i$ a sufficient statistic for μ? Is it a complete statistic? Explain.

   (c) Use the observed data $\mathbf{Y}$ to construct a size-α uniformly most powerful (UMP) test for testing $H_0 : \ \mu \leq \mu_0$ versus $H_0 : \ \mu > \mu_0$.

   (d) Propose a way to construct a $100(1 - \alpha)\%$ confidence interval for μ based on $\mathbf{Y}$.

2. Infectious diseases are sometimes modeled with a so called *SIR* model (the letters stand for Susceptible, Infected, and Recovered). People begin in class *S*, then possibly migrate to class *I* (i.e., become infected), and then to class *R* (i.e., recover); no other transitions are possible. In a simple version of the model, the i^{th} individual begins in class S, waits a random amount of time $T_i \sim Exp(1/\lambda)$ before migrating to class *I*, then waits another random amount of time $U_i \sim Exp(1/\mu)$ before migrating to class R, with all the exponentially-distributed random variables T_i and U_i independent. Here a random variable $X \sim Exp(1/\lambda)$ if X has the pdf $f(x|\lambda) = \lambda \exp(-\lambda x)$ for $x > 0$.

   (a) Derive the cdf $Pr(T_i \leq t)$.

   (b) Let N denote the number of Susceptibles at time 0 and let X_t be the number of these who become infected by time t. Find the probability distribution of X_t.

   (c) Let W_1 be the length of time until the *first* of the N Susceptibles becomes infected. Find the probability distribution for W_1.

   (d) Let W_N be the length of time until the *last* of the N Susceptibles becomes infected. Find the probability density function for W_N.

   (e) Let $Y_i = T_i + U_i$ be the total amount of time the i^{th} Susceptible waits before joining class R. Find the probability distribution of Y_i under the (simplifying) assumption $\lambda = \mu$, explaining your reasoning.

3. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f(y|\theta)$, where $\theta \in \Theta \subseteq \mathcal{R}^p$. Let $\widehat{\theta}$ denote the maximum likelihood estimator of θ and let $\widehat{\theta}_{(i)}$ denote the maximum likelihood estimator of θ when the ith observation Y_i is deleted from the sample. To test model misspecification using the data $Y_1, Y_2, ..., Y_n$, Presnell and Boos (*Journal of the American Statistical Association*, **99**, 216-227) propose the logarithm of the "in and out of sample" (IOS) likelihood ratio, given by

$$\text{IOS} = \log \left\{ \frac{\prod_{i=1}^{n} f(Y_i|\widehat{\theta})}{\prod_{i=1}^{n} f(Y_i|\widehat{\theta}_{(i)})} \right\}.$$

   (a) Show that the IOS statistic can be rewritten as

$$\text{IOS} = \sum_{i=1}^{n} \{ l(Y_i; \widehat{\theta}) - l(Y_i; \widehat{\theta}_{(i)}) \},$$

   where $l(y;\theta) = \log f(y|\theta)$.

   (b) Take $p = 1$ and suppose that $Y_1, Y_2, ..., Y_n$ are iid Poisson with mean $\theta > 0$. Show that

$$\text{IOS} = \sum_{i=1}^{n} (\overline{Y}_{(i)} - \overline{Y}) + \sum_{i=1}^{n} Y_i \log(\overline{Y}/\overline{Y}_{(i)}) \approx \frac{S^2}{\overline{Y}},$$

   where S^2 is the usual sample variance. *Hint:* To show the approximate equality, do the following. First, show that

$$\overline{Y}_{(i)} - \overline{Y} = (\overline{Y} - Y_i)/(n-1).$$

   Second, use the first-order Taylor series expansion

$$\log(\overline{Y}_{(i)}) \approx \log(\overline{Y}) + (\overline{Y}_{(i)} - \overline{Y})/\overline{Y}.$$

   (c) In part (b), argue that IOS $\xrightarrow{p}$ 1, as $n \to \infty$.

   (d) Consider the data below:

```
6   10   6   8   8   7   10   3   11   4   10   6   8   7   8
```

   Discuss how you could use the results in parts (b) and (c) to formulate a procedure to test whether or not $Y_1, Y_2, ..., Y_{15}$ could be modeled using as iid observations from a Poisson distribution. Suggest a suitable test statistic. You are not being asked to carry out the test formally; just provide as many details as possible on how you would perform the test.

4. Consider a toxicology study with k groups of animals who are given a drug at distinct dose levels $d_1, d_2, ..., d_k$, respectively (these are fixed by the experimenter; not random). The animals are monitored for a reaction to the drug. In group i, let n_i (fixed) denote the total number of animals dosed, and let Y_i denote the number of animals that respond to the drug. The observations $Y_1, Y_2, ..., Y_k$ are treated as independent random variables, where $Y_i \sim$ Binomial(n_i, p_i); $i = 1, 2, ..., k$, where p_i is the probability that an individual animal responds to dose d_i. A standard assumption in such toxicology studies is that

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 d_i,$$

for $i = 1, 2, ..., k$, where β_0 and β_1 are real parameters (this is merely logistic regression using dose as a predictor).

(a) Find a two-dimensional sufficient statistic for $\boldsymbol{\beta} = (\beta_0, \beta_1)'$.

(b) Suppose that $k = 10$ and that $n_i = 3$, for $i = 1, 2, ..., 10$. The observed data from this study are below.

d_i	0.5	1.0	1.5	2.0	2.5	3.0	4.0	5.0	7.5	10.0
Y_i	1	0	0	1	2	1	1	2	3	2

Find (numerically) the maximum likelihood estimates of β_0 and β_1 based on these data. Treating p_i as a function of d_i, plot your estimated model.

(c) Based on the data in part (b), what can you say about the relationship between dosing and response to the drug?

(d) Based on the data in part (b), find a 90 percent confidence interval for the probability that an individual animal responds to the dose when $d = 6.0$.

5. *Monkey business*: A randomized block design was implemented to examine whether chimpanzees learn some words from American Sign Language (ASL) more quickly than others. Each of four chimpanzees, Booee, Cindy, Bruno, and Thelma, were shown 10 words in random order and the number of minutes it took a chimp to learn each word was recorded. The words are Listen, Drink, Shoe, Key, More, Food, Fruit, Hat, Look, and String.

   The data are available in the file `chimp.txt` at the website

   `http://www.stat.sc.edu/~dryden/qualifier-day1`

   Analyze these data keeping in mind the goal of this experiment. Be complete. There may be more than one satisfactory approach to modeling these data.

6. A furniture company has four factories (A,B,C,D) in different countries which make four particular types of chair (1,2,3,4). The variable measured is an overall customer satisfaction score for the quality of each product in a time period, where four distinct time periods are available (I, II, III, IV). The company is primarily interested in comparing the factories.

		Time Period			
		I	II	III	IV
Chair type	1	A 8.5	B 8.7	D 9.8	C 9.4
	2	B 9.2	A 7.4	C 9.2	D 9.2
	3	C 9.3	D 9.1	A 8.4	B 9.1
	4	D 9.1	C 8.9	B 9.3	A 9.1

(a) Explain what is meant by the term ‘blocking’ and state its advantages. What are the blocking factors in this study?

(b) What type of design has been used here?

(c) Explain why randomization is used when designing experiments, and explain how you would have randomized this particular design.

(d) Write down a model for analyzing the data from this design, explaining your notation.

(e) Calculate the Analysis of Variance (ANOVA) table for the data.

(f) Carry out an appropriate analysis of the data, using a 5% significance level and checking your model assumptions.

(g) If you were engaged as a consultant what would you advise the company if they wish to know the factory which makes the highest quality furniture?

(h) The company asks you to investigate what would have been the conclusion if the response for the first experimental unit (Factory A, Chair 1, Time I) had been x instead of 8.5, where $x \in \{4.5, 5.5, 6.5, 7.5, 8.5, 9.5\}$. Provide an answer for the company, again using a 5% significance level in your discussion.

May 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 and write your Candidate Number on each sheet, but do not include your name.

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 datasets that are needed in some of the problems.

A formula sheet is included. Some SAS and R macros are available at

`http://www.stat.sc.edu/∼dryden/qualifier-day2`

Provide details in your solutions. You have **four hours** to complete this examination. Good luck.

1. Suppose that X is a χ^2_1 distributed random variable with probability density function

$$f_X(x) = \frac{e^{-x/2}}{\sqrt{2\pi x}} \quad , \qquad x > 0 \ ,$$

and Y is also independently distributed as χ^2_1. Let $U = X + Y$ and $V = \frac{X}{Y}$.

(a) Derive the joint probability density function of U and V.

(b) Derive the marginal probability density function of V.

(c) Derive the marginal probability density function of U.

(d) Are U and V independent? Justify your answer.

(e) Consider k independent random variables $U_1, \ldots, U_k$ from the same distribution as U. Given that the mean and standard deviation of U are both 2, what is the approximate distribution of

$$Z = \frac{1}{k} \sum_{i=1}^{k} U_i \ ,$$

for large k ?

(f) A random variable X has p.d.f. $f(x)$, the functional form of which is unknown. A random sample of size n $(X_1, \ldots, X_n)$ is drawn to test the null hypothesis

$$H_0 : f(x) = f_0(x)$$

against the alternative

$$H_1 : f(x) = f_1(x).$$

Hence find the form of the most powerful critical region for the test of H_0 against H_1 in the case where

$$f_0(x) = \frac{e^{-x/2}}{\sqrt{2\pi x}} \quad , \qquad x > 0 \ ,$$

and

$$f_1(x) = \left(\frac{2}{\pi}\right)^{1/2} e^{-x^2/2} \quad , \quad x > 0.$$

2. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample from $N(\mu,\, \sigma^2)$ where μ is unknown and $\sigma^2 > 0$ is known.

   (a) Find the Cramer-Rao lower bound (CRLB) for the unbiased estimators for $\tau(\mu) = e^{t\mu}$, where $t \neq 0$ is a fixed constant (not a parameter to be estimated).

   (b) Find the uniform minimum variance unbiased estimator (UMVUE) for $\tau(\mu) = e^{t\mu}$, denoted by $\hat{\tau}$.

   (c) Show that Var$(\hat{\tau})$ is larger than the CRLB in part (a), but the ratio, Var$(\hat{\tau})$/CRLB, tends to one as $n \to \infty$.

3. *The shifted Exponential*: Let $X_1, \ldots, X_n$ be a random sample from the density

$$f(x|\theta, \alpha) = \frac{1}{\theta} \exp\left(-\frac{x-\alpha}{\theta}\right) I_{[\alpha,\infty)}(x),$$

for $\theta > 0$ and real α. Suppose θ and α are unknown.

(a) Let the first order statistic be denoted $X_{(1)} = \min\{X_1, \ldots, X_n\}$. Find the distribution of

$$W = \frac{n(X_{(1)} - \alpha)}{\theta}.$$

(b) Derive the MLE $(\hat{\theta}, \hat{\alpha})$ for (θ, α). Hint: First find the MLE $\hat{\alpha}$, then use $\hat{\alpha}$ to find the MLE $\hat{\theta}$.

(c) Use (a) and (b) to develop a 95% confidence interval for α, by plugging in $\hat{\theta}$ for θ. The times it took for $n = 10$ preschoolers to complete a task in minutes are

$$2.3 \;\; 1.6 \;\; 1.1 \;\; 1.7 \;\; 1.1 \;\; 1.7 \;\; 1.2 \;\; 1.6 \;\; 4.4 \;\; 2.4.$$

Find, *and interpret*, a 95% confidence interval for α for these data.

(d) Derive the method of moment estimators $(\tilde{\theta}, \tilde{\alpha})$ for (θ, α).

4. *More Bang!*

Investors are concerned about the return of their money. Suppose you invested $\$1000$ in the US stock market last year. After one year, your investment is worth $\$1080$. The return is then $1080/1000 = 1.08$, and the rate of the return is $\log(1080/1000) = 0.077$.

Now, we consider a model for investment strategy. Label the initial value of your investment as W_0, (e.g. $W_0 = \$1000$) and the annual return of the investment as R_t (e.g. $R_1 = 1.08$) during year t. The value at the end of the first year is $W_1 = W_0 R_1$, and by the end of year T the value is

$$W_T = W_{T-1} R_T = \cdots = W_0 R_1 R_2 \cdots R_T.$$

We suppose $\{R_t\}_{t=0}^T$ are i.i.d. random variables.

Based upon the historical data, we have summary statistics of the return R_t for different assets as in Table 1.

	Stocks	T-bills
Mean	1.10	1.05
Std Dev	0.20	0.04

Table 1: *Mean, standard deviations of annual returns, R_t, on US stocks and Treasury Bills*

If we start with $\$1000$ in each of the stock and the T-bills, we would expect to have $\$1,100$ in stock and $\$1,050$ in T-bills after one year. Because the expected value of a product of independent random variables is the product of expectations, we can find the expectations for each investment over a longer horizon given this assumption. Over 20 years, the initial investment of $\$1000$ in stock grows in expectation to $\$1000 \times (1.1)^{20} = \$6727$. By comparison, the initial investment in T-bills grows to $\$2653$.

At first glance, the above calculation of expected values seems quite reasonable. However, it uses only the mean of the returns and has no appreciation of the standard deviation - the risks!

To have a deeper understanding of impact of the variance on the long term returns, we convert the product to a sum:

$$\log(W_T) = \log(W_0) + \sum_{t=1}^{T} \log(R_t) = \log(W_0) + \sum_{t=1}^{T} r_t$$

where $r_t = \log(R_t)$ is the *continuously compounded rate of return* for $t = 1, \ldots, T$. Now for large T and by law of large numbers, we have

$$\log(W_T) \approx \log(W_0) + T\mathrm{E}(r_t)$$

Please answer the following questions on the next page.

(a) Let μ_r be the expectation of the rate of return (i.e. $\mu_r \equiv \mathrm{E}(r_t)$). Approximate μ_r for stocks and T-bills, respectively, using the mean and variance for R_t in Table 1 and the second-order Taylor series expansion $\log(1+x) \approx x - x^2/2$.

(b) Let σ_r^2 be the variance of the rate of return (i.e. $\sigma_r^2 \equiv \mathrm{Var}(r_t)$). Approximate σ_r^2 for stocks and T-bills, respectively, using the mean and variance for R_t in Table 1 and the first order Taylor series expansion $\log(1+x) \approx x$.

(c) Now suppose r_t are iid $N(\mu_r, \sigma_r^2)$ so that R_t follows a *log-normal distribution*. Use the calculated μ_r and σ_r from (a) and (b). Simulate the path of $\{W_t\}_{t=1}^{40}$ 10,000 times for $W_0 = \$1000$ and for stock. What is the largest W_{40} among these 10,000 simulations?

(d) There are millions of investors seeking profits in the US stock market. A few of them, such as Warren Buffet or Peter Lynch, are famous for consistently generating huge positive returns over time. Their performances are often attributed to their knowledgeable investment strategy. Criticize these statements based on your statistical thinking.

5. An ornithologist wants to relate the amount of energy (in calories) to temperature using birds of two similar species. Energy use was measured for a different bird from each species at each of 16 different temperatures (in deg C). The data from the study are given below and are provided in the file called `birds.txt` at the website

   `http://www.stat.sc.edu/∼dryden/qualifier-day2`

Bird	Species	Temperature	Calories	Bird	Species	Temperature	Calories
1	A	0	37.4	17	B	0	41.1
2	A	2	34.9	18	B	2	40.9
3	A	4	34.6	19	B	4	38.9
4	A	6	35.3	20	B	6	37.3
5	A	8	32.8	21	B	8	37.0
6	A	10	31.7	22	B	10	36.1
7	A	12	31.0	23	B	12	36.3
8	A	14	29.2	24	B	14	34.2
9	A	16	29.1	25	B	16	33.4
10	A	18	28.2	26	B	18	32.8
11	A	20	27.4	27	B	20	32.0
12	A	22	27.8	28	B	22	31.9
13	A	24	25.5	29	B	24	30.7
14	A	26	24.9	30	B	26	29.5
15	A	28	23.7	31	B	28	28.5
16	A	30	23.1	32	B	30	27.7

   (a) Come up with a suitable statistical model for the ornithologist. Fit the statistical model you choose and explain to the ornithologist why you have decided on it.

   (b) Provide a point prediction and 95 percent prediction interval for the energy use of a single new bird from species A held at 15 deg C. Interpret the prediction interval in a way that the ornithologist can understand.

   (c) The ornithologist is interested in a theory which indicates that birds from species B will burn more calories than birds from species A if both are held at the same temperature. Do the data support this theory? Perform an analysis and explain your findings to the ornithologist.

6. The table below presents measurements of a particular chemical in tablets which are expected to contain 4mg of this chemical. The measurements were taken in six laboratories, and each laboratory had eleven measurements taken by different inspectors who visited each lab in turn. It is of primary interest to examine whether there is a difference in the mean measurement in the laboratories, taking into account that the inspectors may carry out the measurements in a slightly different manner.

inspector	Lab1	Lab2	Lab3	Lab4	Lab5	Lab6
1	3.99	3.86	4.00	3.88	4.02	4.02
2	4.07	3.85	4.02	3.88	3.95	3.86
3	4.04	4.08	4.01	3.91	4.02	3.96
4	4.07	4.11	4.01	4.02	3.89	3.97
5	4.05	4.08	4.04	3.92	3.91	4.00
6	4.04	4.01	3.99	3.97	4.01	3.82
7	4.02	4.02	4.03	4.02	3.89	3.95
8	4.06	4.04	3.97	3.90	3.89	3.99
9	4.10	3.97	3.98	3.97	3.99	4.02
10	4.04	3.95	3.98	3.90	4.00	3.93
11	4.03	4.02	4.02	4.01	4.11	4.01

The data are available in the file `chemical.txt` at the website

`http://www.stat.sc.edu/~dryden/qualifier-day2`

(a) Carry out an analysis of the data. In particular, suggest a suitable model for the data and state the assumptions. Fit the model in a computer package, taking care to check the model assumptions. Assuming the assumptions are reasonable carry out suitable tests, and report your conclusions.

(b) Consider now the situation where the inspectors return a month later to each laboratory and take a second set of measurements. Suggest a suitable model for modeling the full dataset of 132 experimental units, carefully describing the assumptions. Provide details of the tests that you would carry out as part of your analysis, giving the test statistics as functions of the mean square error terms MSA, MSB, $MSAB$, MSE, which have sources given by the laboratories, inspectors, interactions and errors respectively.