January 2013 PhD Qualifying Examination
Department of Statistics
University of South Carolina
9:00AM–3:00PM

Instructions: This exam consists of six problems. You are to answer all six problems. 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 to examine help files of the statistical software and to examine data sets that are needed in some of the problems. Provide complete details in your solutions. You have **six hours** to complete this examination. Good luck.

1. In a quality control process for a production chain, we know the following:

- Any item in the production chain has probability p' of being inspected.
- Any item has probability p of being acceptable (not defective).

In other words, if I denotes the event that an item is inspected and A denotes the event that an item is not defective, then $P(I) = p'$ and $P(A) = p$. For notational purposes, let $q' = 1 - p'$ and $q = 1 - p$.

Assume that I and A are independent events. Thus, the quality control process can be described by the following 2×2 table:

	Not Defective	Defective	Total
Inspected	pp'	qp'	p'
Not Inspected	pq'	qq'	q'
Total	p	q	1

Each item belongs to exactly one of the four categories shown in the table above. Assume that all items are independent.

Let N be the number of items passing the production chain before the first defective item is detected. Let K be the number of undetected defective items among these N items.

(a) Justify (in words) the following:

(i) N follows a geometric distribution, specifically,

$$P(N = n) = (1 - qp')^n qp', \quad n = 0, 1, 2, ..., .$$

(ii) The conditional distribution of K, given $N = n$, is

$$P(K = k|N = n) = \binom{n}{k} \left(\frac{qq'}{1 - qp'}\right)^k \left(1 - \frac{qq'}{1 - qp'}\right)^{n-k}, \quad k = 0, 1, 2, ..., n.$$

(b) Find the joint distribution of N and K.
(c) Find the marginal distribution of K.
(d) Find $\text{cov}(N, K)$.

2. Consider data from a study with $n = 32$ adults. The study focused on alcohol metabolism, with the ultimate goal of answering questions related to female's lower tolerance for alcohol and greater propensity to develop accompanying alcohol-related liver disease, relative to males. The variables are:

- Metabol – First-pass metabolism of alcohol in the stomach (mmol/liter-hour); this is the response of interest.
- Gastric – The gastric alcohol dehydrogenase activity in the stomach (μmol/min/g of tissue).
- Sex – The subject's gender.
- Alcohol – Indicates whether subject is an alcoholic or not.

The data are available for download at `http://www.stat.sc.edu/~hansont/alcohol.txt`. You are to find a parsimonious, yet adequate, explanatory regression model for the metabolism response variable involving gender for sure, and including the remaining concomitant variables if necessary. Make sure that you carefully assess all assumptions for your final model and write a coherent and complete summary of your analysis, addressing the scientific question at hand.

3. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a uniform distribution over $(\theta,\ \theta + |\theta|)$, where $\theta \neq 0$.
(a) Find the method of moments estimator of θ.
(b) Find the maximum likelihood estimator (MLE) of θ.
(c) Is the MLE of θ a consistent estimator of θ? Explain.

4. Suppose $X_1, X_2, ..., X_n$ is an iid sample of $\mathcal{N}(\mu, \sigma^2)$ observations where σ^2 is known. Let M denote the sample median of $X_1, X_2, ..., X_n$. Our goal is to estimate $\sigma_M^2 = \text{var}(M)$. We will consider two different approaches to do this. These approaches are described in **bold** font.

Approach 1: Generate B samples of size n from a $\mathcal{N}(\mu, \sigma^2)$ distribution and compute the median for each sample resulting in $M_1, M_2, ..., M_B$. Compute the sample variance

$$S_M^2 = (B-1)^{-1} \sum_{b=1}^{B} (M_b - \overline{M})^2,$$

where M_b is the sample median of the bth data set, $b = 1, 2, ..., B$, and $\overline{M} = B^{-1} \sum_{b=1}^{B} M_b$.

(a) Argue that S_M^2 is a sensible estimator for σ_M^2. Under what conditions would you expect this to be a "good" estimator for σ_M^2?

(b) For any member of the $\mathcal{N}(\mu, \sigma^2)$ family, with σ^2 known, prove that

$$\text{var}(M) = \text{var}(M - \overline{X}) + \text{var}(\overline{X})$$

Hint: Write $M = M - \overline{X} + \overline{X}$. This result motivates the second approach.

Approach 2: Generate B samples of size n from a $\mathcal{N}(\mu, \sigma^2)$ distribution, define

$$T_b = M_b - \overline{X}_b,$$

where M_b and $\overline{X}_b$ are the sample median and sample mean, respectively, of the bth data set, $b = 1, 2, ..., B$, and compute

$$S_T^2 = (B-1)^{-1} \sum_{b=1}^{B} (T_b - \overline{T})^2,$$

where $\overline{T} = B^{-1} \sum_{b=1}^{B} T_b$. Finally, calculate $\widehat{\sigma}_M^2 = S_T^2 + \sigma^2/n$.

(c) Argue that $\widehat{\sigma}_M^2$ is a sensible estimator for σ_M^2. Under what conditions would you expect this to be a "good" estimator for σ_M^2?

(d) Which estimator do you prefer: S_M^2 or $\widehat{\sigma}_M^2$? Why?

5. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a population whose distribution is specified by the pdf $f(x)$. Let $f_0(x)$ and $f_1(x)$ be two known pdf's, and $f(x)$ is one of them. Suppose that $\delta(\mathbf{X})$ is a test function associated with the uniformly most powerful (UMP) test of size $\alpha \in (0,\ 1)$ for testing

$$H_0 : f = f_0$$
$$\text{versus}$$
$$H_1 : f = f_1,$$

and suppose that the power associated with $\delta(\mathbf{X})$ under H_1 is $\beta \in (0,\ 1)$. Derive the UMP test of size $1 - \beta$ for testing $H_0^* : f = f_1$ versus $H_1^* : f = f_0$. Express the corresponding test function in terms of $\delta(\mathbf{X})$.

6. A biologist designed an experiment to assess the weight gain in $n = 40$ rats fed diets comprised of four different combinations of two protein sources and two protein amounts. This is a completely randomized design with ten rats randomly allocated to each of the four treatments. The variables are:

- PreWt – The initial weight before the experiment (grams).
- PostWt – The weight after the experiment (grams).
- Protein – The protein source: either Beef or Cereal.
- Amount – The amount of protein: either High or Low.

The data are available for download at `http://www.stat.sc.edu/~hansont/rat_data.txt`. Build a model that best describes the relationship between the weight gain *as a proportion of initial weight* and the factors Protein and Amount. Make sure that you carefully assess all model assumptions and write a coherent and complete summary of your analysis.