May 2014 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. Suppose that $F_0(x)$ is a cumulative distribution function (cdf). Suppose that X is a random variable with

$$F_X(x) = P_X(X \leq x) = (1+\delta)F_0(x) - \delta[F_0(x)]^2,$$

for $x \in \mathbb{R}$, where δ satisfies $-1 \leq \delta \leq 1$.

(a) Show that $F_X(x)$ is a valid cdf. You may assume that $F_0(x)$, the so-called "base cdf," is valid.

For parts (b) and (c) only, take the base cdf to be $F_0(x) = 1 - e^{-x/\lambda}$, for $x > 0$, where $\lambda > 0$. For $x \leq 0$, take $F_0(x) = 0$. Set $\boldsymbol{\theta} = (\delta, \lambda)'$.

(b) Show that the probability density function (pdf) of X is

$$f_X(x|\boldsymbol{\theta}) = \frac{1}{\lambda}e^{-x/\lambda}(1 - \delta + 2\delta e^{-x/\lambda})I(x > 0).$$

(c) For $k > 0$, show that

$$E(X^k) = \lambda^k\Gamma(k+1)(1 - \delta + \delta 2^{-k}),$$

where $\Gamma(\cdot)$ is the gamma function.

(d) Suppose that $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x|\boldsymbol{\theta})$ given in part (b). Show how you would find the method of moments (MOM) estimator of $\boldsymbol{\theta}$. You do not have to calculate the estimator in closed form (just set up how you would do it).

2. Consider data from Bell et al. (1994) on the result of spinal laminectomy, a corrective surgery performed on $n = 83$ children. The specific response of interest is the presence (1) or absence (0) of kyphosis, defined as a forward flexion of the spine of at least 40 degrees from vertical. The available (numerical) predictor variables are age in months at the time of operation (age) and the starting level of vertebrae involved in the operation (start). Build an appropriate statistical model for the outcome kyphosis. In particular, determine the age in months at which the risk for kyphosis is greatest, holding starting level constant. The data are available at `http://www.stat.sc.edu/∼hansont/kyphosis.sas`

3. Suppose that X_1, X_2, X_3, and X_4 are independent Poisson random variables with means θ_1, θ_2, $\theta_1 + \theta_2$, and $\theta_1\theta_2$, respectively.

(a) Derive the distribution of $T = X_1 + X_2$.

(b) Derive the conditional distribution of X_1 given $T = t$.

(c) Consider testing

$$\begin{gathered} H_0 : \theta_1 \leq \theta_2 \\ \text{versus} \\ H_1 : \theta_1 > \theta_2 \end{gathered}$$

on the basis of observing X_1 and X_2 only. Calculate $\lambda \equiv \lambda(x_1, x_2)$, the likelihood ratio test statistic, to test H_0 versus H_1. You do not need to do anything else other than calculate λ.

(d) Suppose that only X_3 and X_4 are observed. In this case, are the parameters θ_1 and θ_2 identifiable? Explain.

(e) You need to calculate the mean and variance of

$$U = \frac{X_1 + X_2}{1 + X_3 + X_4}$$

under the assumption that $\theta_1 = \theta_2 = 1$. Provide a strategy on how to do this, explain why your strategy works, and then provide your answers.

4. (a) Let $(X_1, \ldots, X_n)$ be a random sample from Bernoulli(p), where $0 < p < 1$. Define $S_n = \sum_{i=1}^n X_i$. Show that, as $n \to \infty$,

$$\frac{S_n - np}{\sqrt{npq}} \xrightarrow{d} N(0,1),$$

where $q = 1 - p$ and "$\xrightarrow{d}$" refers to "converges in distribution."

Note: The following extension of the above asymptotic result also holds.
If $(X_1, \ldots, X_n)$ is a random sample from Bernoulli(p_n), where $0 < p_n < 1$ for all n and $\lim_{n\to\infty} p_n = p \in (0,1)$, then, with $S_n = \sum_{i=1}^n X_i$ and $q_n = 1 - p_n$,

$$\frac{S_n - np_n}{\sqrt{np_n q_n}} \xrightarrow{d} N(0,1), \tag{1}$$

as $n \to \infty$. For example, one may have $p_n = 2^n/(2^{n+1}+1)$, so that, when $n = 2$, (X_1, X_2) is a random sample from Bernoulli(4/9); when $n = 4$, (X_1, X_2, X_3, X_4) is a random sample from Bernoulli(16/33). You do not need to prove (1) although you may use it for part (b).

(b) Let $(Y_1, \ldots, Y_n)$ be a random sample of size $n = 2m - 1$ from a distribution of which the cdf is $G(y)$, where m is a positive integer. Suppose $P(Y \le y) = G(y) = F(y - \theta)$, where $G(\theta) = F(0) = 0.5$ so that θ is a median of the distribution of Y. Denote by $\tilde{Y}_n = Y_{(m)}$ the sample median.

(i) Show that the cdf of $\tilde{Y}_n$ evaluated at y, i.e., $P(\tilde{Y}_n \le y)$, is equivalent to the cdf of S_n evaluated at $m - 1$, i.e., $P(S_n \le m - 1)$, where S_n follows a binomial distribution with n trials and probability of success equal to $1 - G(y)$.

(ii) Show that, as $n \to \infty$,

$$\sqrt{n}(\tilde{Y}_n - \theta) \xrightarrow{d} N(0, 1/\{2f(0)\}^2), \tag{2}$$

where $f(0) = (d/dt)F(t)|_{t=0}$.

(c) In the context of part (b), without knowing the specific form of $G(y)$, propose a strategy for constructing an asymptotically valid confidence interval for the population median, θ, based on a random sample $(Y_1, \ldots, Y_n)$ from this population.

5. During exercise, blood flow increases in some parts of the body in response to metabolic demand. Using radioactive microspheres, an experiment was conducted to determine in which of five parts of the body in rats this occurs: bone, brain, skin, muscle, and heart. The more microspheres that accumulate in a region, the greater the blood flow. Eight rats were injected with radioactive microspheres. After injection, four rats were exercised on a treadmill for 15 minutes and the other four rats were placed on a treadmill but it was not turned on. After the 15 minutes the rats were killed and the radioactivity in the five tissue locations was measured Y_{ijk}. Here $i = 1, \ldots, 8$ denotes the rat, $j = 1, 2$ denotes treatment (sedentary and exercise, respectively), and $k = 1, \ldots, 5$ denotes body part (bone, brain, skin, muscle, heart). Fit a model that accounts for repeated measures over time, with factorial treatment structure:

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}. \tag{3}$$

Here, $\rho_1, \ldots, \rho_8 \overset{iid}{\sim} N(0, \sigma_\rho^2)$ independent of $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$. These data are available at `http://www.stat.sc.edu/∼hansont/blood.sas`

(a) Obtain a profile (spaghetti) plot for each treatment group. Describe what you see in terms of differences in exercise vs. sedentary rats. Do you expect a body part by treatment interaction based on the plot? Explain.

(b) Let $1 \le k_1 < k_2 \le 5$. What is the *estimated* (not theoretical) correlation for two repeated measures within a rat for model (3), $\text{corr}(Y_{ijk_1}, Y_{ijk_2})$?

(c) Report the ANOVA table for the fixed effects (i.e. the Type III tests). Is there a significant body part by treatment interaction here? comment in light of the spaghetti plots.

(d) Report a test of $H_0 : \sigma_\rho = 0$ versus $H_A : \sigma_\rho > 0$; is blocking on rat effective?

(e) Look at pairwise differences in the two treatments *at each body part* and discuss significant differences adjusting for multiple comparisons.

(f) Examine plots of the conditional residuals versus fitted values, rat, treatment, and body part. Is constant variance and normality reasonable for the ϵ_{ijk}?

(g) Obtain the fitted $\hat{\rho}_i$. Is normality among the subject effects reasonable?

6. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from the probability density function (pdf)

$$f_X(x|\theta) = \frac{2x}{\theta} e^{-x^2/\theta} I(x > 0),$$

where $\theta > 0$.

(a) Show that $T = \sum_{i=1}^n X_i^2$ is a complete and sufficient statistic.

(b) Find the uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\theta) = \theta^2$.

(c) Find the uniformly most powerful (UMP) level α rejection region for testing

$$\begin{gathered} H_0 : \theta \geq \theta_0 \\ \text{versus} \\ H_1 : \theta < \theta_0, \end{gathered}$$

where θ_0 is known. State your rejection region in terms of T and a quantile from a named probability distribution.

(d) Derive the power function for your UMP test in part (c).