January 2015 PhD Qualifying Examination
Department of Statistics
University of South Carolina
9:00AM–3:00PM

Instructions: This exam consists of six problems. Answer all six problems. Use separate sheets of paper for each problem, and write on one side of each page (do **not** write on both sides). You are allowed to use the computers and the software in the examination room. However, you are **not** allowed to use the Internet, except to examine help files of the 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. Gastwirth, Bura, and Miao (2011) discuss the use of statistical evidence to identify potential employment discrimination. The authors cited a court case involving a company with locations in 19 different regions. In each region, Group A was said to be discriminated against (when compared to Group B) if its members were promoted less than 80 percent as often as members of Group B.

Let p_A (p_B) denote the population proportion of Group A (Group B) promotions. In each region, a hypothesis test is performed for

$$\begin{array}{c} H_0 : \theta = 0.8 \\ \text{versus} \\ H_1 : \theta < 0.8, \end{array}$$

where $\theta = p_A/p_B$ is the ratio of the population proportions. For each region, note that H_0 does not correspond to discrimination, but H_1 does.

An exact size $\alpha = 0.05$ test for H_0 versus H_1 was performed in each region. Here are the 19 observed p-values (in ascending order) from the tests, one for each region:

<0.001	0.006	0.007	0.078	0.086	0.171	
0.201	0.234	0.255	0.273	0.318	0.319	
0.435	0.597	0.664	0.828	0.850	0.862	0.921

These p-values are available at

`http://www.stat.sc.edu/~tebbs/p-value.htm`

Assume that promotion practices among the 19 different regions are independent (so that the 19 tests performed above are independent).

(a) Consider testing

$$\begin{array}{c} H_0^* : \text{none of the 19 regions were discriminatory} \\ \text{versus} \\ H_1^* : \text{at least one of the 19 regions was discriminatory.} \end{array}$$

Based only on the 19 reject/fail to reject decisions above, determine a p-value for testing H_0^* versus H_1^*. What is your decision at $\alpha = 0.05$?

(b) Each region has a small sample size resulting in small power for each of the 19 tests above. In each region, the power is only 20 percent when $\theta = 0.7$. Find the power of your test in part (a) under this assumption.

(c) If $\theta = 0.8$ in each region, describe the distribution that the 19 p-values would follow. Check this assumption graphically.

(d) Perform a goodness of fit test with the distribution you identified in part (c). State your conclusion in practical terms.

2. Suppose X and Y are independent random variables, each distributed as $\mathcal{N}(0,1)$.

(a) Show that $X - Y \overset{d}{=} X + Y$. Recall that "$\overset{d}{=}$" means "equal in distribution."

(b) Let $Z = \min(X, Y)$. Show that $Z^2 \sim \chi_1^2$. *Hint:* Derive the cumulative distribution function (cdf) of Z^2.

(c) Suppose $a > 0$. Show that

$$E[X \max(0, X, aY)] = \frac{1}{2} - \frac{1}{2\pi} \arctan(a).$$

Note: If you cannot show this result analytically, then for partial credit you could approximate $E[X \max(0, X, aY)]$ for a fixed value of a.

3. Suppose that $X_1, X_2, ..., X_n$ is an iid sample, each with probability p of being distributed as uniform over $(-1/2, 1/2)$ and with probability $1-p$ of being distributed as uniform over $(0, 1)$.

(a) Find the cumulative distribution function (cdf) and the probability density function (pdf) of X_1.

(b) Find the maximum likelihood estimator (MLE) of p and determine its asymptotic distribution.

(c) Find another estimator of p using the method of moments (MOM). Determine its asymptotic distribution.

(d) Which of the two estimators (MLE or MOM) is more efficient? Prove your answer.

4. Three separate wetlands sites were studied for density of three species of sedge (*Carex* species−a type of grassy plant). At each site, 36 1-m^2 plots were selected, twelve were randomly assigned to each sedge species and the number of stems for that sedge species were counted in each plot.

The data for this question are available at

`http://www.stat.sc.edu/∼tebbs/carex.htm`

The nominal factors “Site” and “Species” each have 3 levels. The response is “Stems.”

Consider this an observational study with 12 replications at each combination of the fixed effects Site and Species.

(a) Generate a graphical display so that interaction between the two factors can be evaluated, and the equal variances assumption for a two-way fixed effects model can be assessed. Do variances appear equal in each of the 9 cells? Does interaction appear to be present?

(b) Test for equality of variances.

(c) Analyze the data using a method appropriate for the data.

(d) Are interactions significant? If so, suggest and implement appropriate tests for multiple comparisons of factor means.

5. Let X_1 and X_2 be independent Poisson random variables with mean $\lambda > 0$. Based on these $n = 2$ observations, consider testing $H_0 : \lambda \leq 1$ versus $H_1 : \lambda > 1$.

(a) Let $\phi_1 = \phi_1(X_1, X_2)$ denote the test that rejects H_0 if and only if $X_1 \geq 2$. That is,

$$\phi_1(X_1, X_2) = \begin{cases} 1, & X_1 \geq 2 \\ 0, & \text{otherwise.} \end{cases}$$

Find the size of ϕ_1.

(b) Let $\phi_2(x_1, x_2) = E[\phi_1(X_1, X_2)|X_1 + X_2 = x_1 + x_2]$. Derive a simple formula for $\phi_2(x_1, x_2)$.

(c) Consider the test identified by $\phi_2 = \phi_2(X_1, X_2)$. Compare the power functions of ϕ_1 and ϕ_2 and graph them.

6. Figure 1 below shows two representations of data pertaining to weight measurements on 48 pigs over 9 successive weeks (one measurement per pig per week; weights are measured in lbs).

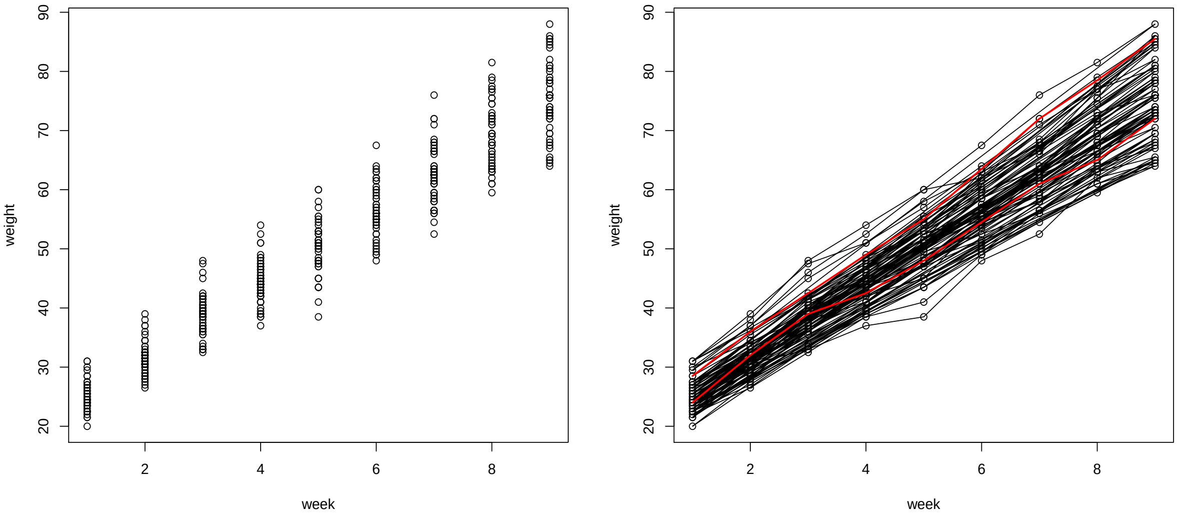

Figure 1: Pig weight data.

The left plot is a scatterplot of the weight responses against the corresponding week number. In the right plot, lines are drawn connecting measurements that belong to the same pig.

Let y_{ij} denote the weight of the ith pig during the jth week and let $x_j = j$ denote the corresponding week number, for $i = 1, 2, ..., 48$, and $j = 1, 2, ..., 9$. The data for this question are available at

`http://www.stat.sc.edu/∼tebbs/pig.htm`

The first column is the pig's ID number; the second column is the week number; the third column is the pig's weight.

Three different models are proposed to describe the pig weight gain over time:

Model 1: $y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}$, where $\epsilon_{ij} \sim \text{iid } \mathcal{N}(0, \sigma_\epsilon^2)$.

Model 2: $y_{ij} = \beta_0 + u_i + \beta_1 x_j + \epsilon_{ij}$, where $\epsilon_{ij} \sim \text{iid } \mathcal{N}(0, \sigma_\epsilon^2), u_i \sim \text{iid } \mathcal{N}(0, \sigma_u^2)$, and the u_i's and ϵ_{ij}'s are mutually independent.

Model 3: $y_{ij} = \beta_0 + u_i + (\beta_1 + v_i)x_j + \epsilon_{ij}$, where $\epsilon_{ij} \sim \text{iid } \mathcal{N}(0, \sigma_\epsilon^2), u_i \sim \text{iid } \mathcal{N}(0, \sigma_u^2), v_i \sim \text{iid } \mathcal{N}(0, \sigma_v^2)$, and the u_i's, ϵ_{ij}'s, and v_i's are mutually independent.

Four questions appear on the next page.

(a) Fit Model 1 and discuss model adequacy.

(b) For each model on the previous page, calculate $\text{corr}(y_{ij}, y_{ij'})$, for $j \neq j'$.

(c) An animal scientist has asked you to explain what each model means and when each model would potentially be useful for this type of data. What would you tell her? Your explanation should be appropriate for an animal scientist.

(d) Which model seems most natural to use for analyzing these data? Provide a brief discussion only. You do not have to fit the second and third models here.