Statistics Ph.D. Comprehensive, January 4, 2019

Instructions: *The exam has 5 (sometimes multi-part) problems. All of the problems will be graded. Write your code words on each of your answer sheets. Do not put your name or UNM ID on any of the sheets. Be clear, concise, and complete. All solutions should be rigorously explained.*

1. Let X_i, $i = 1, \ldots, n$ be i.i.d., where $X_i \in \{A_1, A_2, A_3\}$. Let $N_j = \sum_{i=1}^{n} \mathbb{1}_{\{X_j = A_j\}}$ where $\mathbb{1}_{\{\cdot\}} = 1$ if the condition in the subscript holds true and is otherwise 0. Let p_i be the probability that $X_j = A_j$, $j = 1, 2, 3$. We will consider two methods for testing $p_1 - p_2 = 0$.

   (a) What is the joint distribution of (N_1, N_2, N_3)?

   (b) Show that the $cov(N_j, N_k) = -np_jp_k$ for $j \neq k$.

   (c) In the first test, observe the number of observations that are either in category A_1 or A_2. This number is $M = N_1 + N_2$. Inferences are then done conditional on $M = m$. Then consider estimating the probability that $X_i = A_j$ given that $X_i \in \{A_1, A_2\}$. Let $\widetilde{p}_j = N_j/M$, $j = 1, 2$ estimate this conditional probability. Conditional on only observing $X_i \in \{A_1, A_2\}$, the null hypothesis can be tested by checking whether $\widetilde{p}$ is significantly different from 0.5. Give an approximate 95% confidence interval for the conditional probability that $X_1 = A_1$ given that $X_1 \in \{A_1, A_2\}$.

   (d) For the second test, let $\widehat{p}_j = N_j/n$, $j = 1, 2, 3$. Give a large sample confidence interval for $p_1 - p_2$ taking into account the covariance of $\widehat{p}_1$ and $\widehat{p}_2$.

   (e) Give an expression for the approximate (large-sample) power of testing $H_0 : p_1 - p_2 = 0$ versus an alternative $p_1 - p_2 > 0$ for the method in (c).

   (f) For large samples, which of the two methods above would you prefer for determining whether $p_1 > p_2$? Justify your answer.

   (g) Derive the likelihood ratio test for the null hypothesis $H_0 : p_1 - p_2 = 0$ versus the alternative $H_1 : p_1 - p_2 \neq 0$. For the likelihood ratio, do not condition on $M = m$. Use the full likelihood of the sample.

2. Let X be a nonnegative random variable with continuous density f where $f(0) = k$ and $0 < k < \infty$. Show that $E[1/X] = \infty$.

3. Figure 1 gives all trees (in a certain class) with 4 leaves. There are two models for generating trees, the uniform model that gives equal weight to all trees and the pure birth model. A statistic that is often used to describe trees is the number of leaf clusters of size 2. For example, in the trees in Figure 1, the first 12 trees have one cluster of size 2, and the last three have two clusters of size 2.

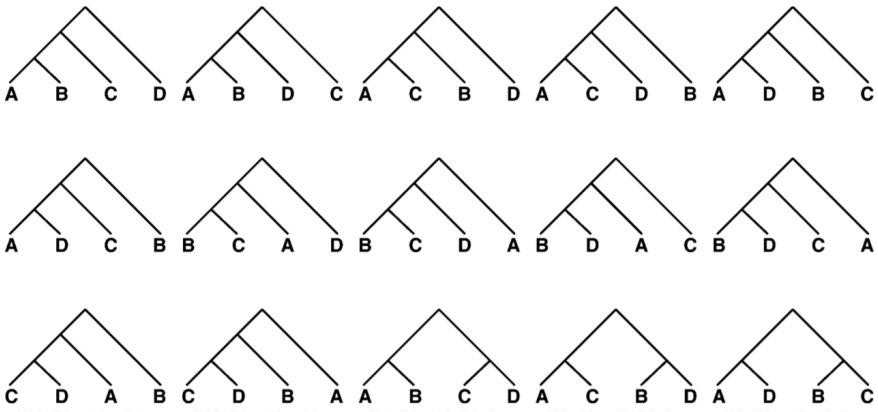

Figure 1: All leaf-labeled rooted binary trees with four leaves when leaves are uniquely labeled. The three trees at the bottom right have two clusters of size two. All other trees have only one cluster of size two. Under the uniform model, each tree has probability 1/15. Under the pure birth model, the three balanced trees in the third row have probability 2/18, and the other trees each have probability 1/18.

Let C_n be the number of clusters of size 2 for a tree with n leaves. Under the pure birth model,

$$\frac{C_n - n/3}{\sqrt{2n/45}} \xrightarrow{L} N(0,1).$$

Under the uniform model

$$\frac{C_n - n/4}{\sqrt{n/16}} \xrightarrow{L} N(0,1).$$

Suppose you observe a single random tree, T_n, where n is large.

(a) If $n = 100$ and $C_n = 30$, which model is your best guess? Justify your answer.

(b) Construct an approximate large-sample level α test of the pure birth model against the uniform model.

(c) Now consider the two hypothesis testing procedures. Procedure 1: the null hypothesis is the pure birth model and is based on rejecting H_0 if $C_n \geq k_1$ for some k_1. Find k_1 for an approximately $\alpha = .05$ level test. For procedure 2 the null hypothesis is the uniform model and rejects H_0 if $C_n \leq k_2$ for some k_2. Find k_2 for an approximately $\alpha = .05$ level test. Which test is more powerful? Justify your answer.

(d) Let the leaves on a tree T_n be $A_1, \ldots, A_n$. Let $X_{ij} = 1$ if A_i and A_j form a cluster of size 2 on the tree (and otherwise $X_{ij} = 0$). Under the uniform model, $P(X_{ij} = 1) = 1/(2n-3)$ for $i \neq j$. Write C_n as a function of the X_{ij}s. Show that $C_n/n \xrightarrow{p} 1/4$.

(e) Find an appropriate asymptotic distribution for $\log(C_n)$ under each model.

4. Consider two q-vectors y_i with $\mathrm{E}(y_i) = \mu_i$ and $\mathrm{Cov}(y_i) = \Sigma$, and consider z independent of the y_is with $z \sim \mathrm{Bern}(p)$. Define the mixture random vector $y \equiv zy_1 + (1-z)y_2$. Assume that $a_1, \ldots, a_q$ are eigenvectors of Σ associated with eigenvalues $\phi_1 > \cdots > \phi_q > 0$.

(a) Show that $\mathrm{Cov}(y) = \Sigma + p(1-p)(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$.

(b) Show that if $\mu_1 - \mu_2$ is an eigenvector for ϕ_k, then the a_is are all eigenvectors of $\mathrm{Cov}(y)$ with corresponding eigenvalues ϕ_i for $i \neq k$ and eigenvalue $\phi_k + p(1-p)\|\mu_1 - \mu_2\|^2$ corresponding to a_k.

(c) Recalling that the first r principal components of y are determined by the eigen analysis of $\mathrm{Cov}(y)$, if $\mu_1 - \mu_2$ is an eigenvector for ϕ_k, what does it take for the first r principal components of y to agree with the first r principal components of y_i?

(d) Show that if $(\mu_1 - \mu_2) \in C(a_{r+1}, \ldots, a_q)$ and $\phi_{r-1} + p(1-p)\|\mu_1 - \mu_2\|^2 < \phi_r$, then the first r principal components of y_i agree with the first r principal components of y.

5. In discussing Alley (1987), George Casella pointed out that for stagewise estimators to have uniformly smaller t^2 statistics than least squares estimators, the estimated variance from the stagewise model must be no smaller than from the usual model. This problem is to prove that fact about the estimated variances.

(a) Consider a linear model

$$Y = \gamma_0 J + \gamma_1 X_1 + X_2\gamma_2 + e, \qquad E(e) = 0$$

where J is an $n \times 1$ vector of 1s and X_1 and X_2 are column vectors. Use ACOVA to find the SSE in terms of M_1, Y and X_2 where M_1 is the perpendicular projection operator onto $C(J, X_1)$.

(b) The stagewise regression model is

$$(I - M_1)Y = \beta_0 J + \beta_2 X_2 + e, \qquad E(e) = 0.$$

Show that the SSE for this model can be written as

$$\{(I - M_1)Y\}' \left[W - WX_2(X_2'WX_2)^{-1}X_2'W\right] \{(I - M_1)Y\}$$

where $W = I - \frac{1}{n}J_n^n$ and J_n^n is an $n \times n$ matrix of 1s. Alley recognizes that the appropriate degrees of freedom for this SSE is $n - 3$,

(c) Show that the stagewise SSE can be rewritten as

$$Y' \left\{ (I - M_1) - (I - M_1) X_2 \left[X_2' W X_2 \right]^{-1} X_2' (I - M_1) \right\} Y.$$

(d) Show that

$$X_2' W X_2 \geq X_2' (I - M_1) X_2.$$

(e) Show that the stagewise SSE must be no smaller than the least squares SSE from part (a). Does this establish that the MSE for stagewise is no smaller than the MSE for least squares?