MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART I

Tuesday, July 29, 2014: 9:00 AM - 3:00PM

Room: BCBS Auditorium

INSTRUCTIONS:

- This is a **CLOSED BOOK** examination.
- Submit answers to **exactly** 3 out of 4 questions. If you submit answers to more than 3 questions, then only questions 1-3 will be counted.
- Put the answers to different questions on **separate sets of paper**. Write on **one side** of the sheet only.
- Put your code letter, **not your name**, on each page, in the upper right corner.
- Return the examination with a **signed honor pledge form**, separate from your answers.
- You are required to answer **only what is asked** in the questions and not to tell all you know about the topics.

1. Suppose we conduct a study of heights of fathers and their sons in a particular population, letting X be the father's height in inches and Y the son's. Further, suppose that the random pair (X, Y) is distributed as bivariate normal with

$$\mathrm{E}[X] = \mathrm{E}[Y] = 68, \mathrm{Var}(X) = \mathrm{Var}(Y) = 4, \mathrm{Cov}(X, Y) = 0.6.$$

In what follows, give explicit expressions and simplify them as much as possible. Show your work, not just the final answer.

(a) What is the probability that the father is taller than the son?

(b) What is the probability that the father is at least 4 inches taller than the son?

(c) What is the distribution of the heights of sons whose fathers are 74 inches tall?

(d) Given that a father is 74 inches tall, find the probability that the son is taller than the father.

(e) One hundred father-son pairs are randomly sampled. Let $\overline{X}$ be the sample average for fathers and $\overline{Y}$ the sample average for sons. What is the joint distribution of $(\overline{X}, \overline{Y})$?

(f) What is the probability that the two sample averages are within 3 inches of each other?

Points: (a) 2.5, (b) 2.5, (c)-(f) 5 each.

2. A study collected data on the number of common colds encountered by individuals in a given population during a one-year period. Here we consider a simple model that might be used in the analysis.

   Let the random variable Y denote the number of common colds encountered by a given person during the study period. Suppose that the *expected number* of common colds encountered by that person is X, and conditional on X, the random variable Y has a Poisson distribution with mean X. Suppose further that, across all subjects, X is distributed as uniform on the interval $(0, 2)$. That is, the pdf of X is $f_X(x) = 0.5$ for $x \in (0, 2)$ and $f_X(x) = 0$ otherwise.

   In what follows, derive explicit expressions and simplify them as much as possible. Show *all* your derivations, not just the final answer. Hint: Conditioning.

   (a) Find $\mathrm{E}[Y]$ and $\mathrm{Var}(Y)$. Does Y have a Poisson distribution? Justify.

   (b) Find $\mathrm{Corr}(X, Y)$.

   (c) Define $W = 4X - Y + 4$. Compute $\mathrm{Cov}(W, Y)$. Are W and Y independent? Justify.

   (d) While Y is observable, X is not. Hence, we would like to use Y to say something about X. Is Y an unbiased predictor of X? Compute the prediction mean squared error. Note: We say that random variable U is an *unbiased predictor* of random variable V if $\mathrm{E}[U - V] = 0$. The *prediction mean squared error* is $\mathrm{E}[(U - V)^2]$.

   (e) Find constants a and b such that $a + bY$ is an unbiased predictor of X and such that the prediction variance is as small as possible; that is $\mathrm{E}[X - a - bY] = 0$ and $\mathrm{Var}(X - a - bY)$ is minimized.

   (f) What is the probability that a given subject gets no common colds within the study period? (Compute the numerical value).

   (g) Compute the conditional mean of X for a subject with no common colds during the study period. That is, compute the numerical value of $E[X|Y = 0]$.

   Points: (a) 2, (b) 2, (c) 3, (d) 3, (e) 6, (f) 3, (g) 6.

3. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from the distribution (pmf),

$$f_X(x|\lambda) = \lambda^x e^{-\lambda}/x!, \quad x = 0, 1, \ldots, \infty, \quad \text{and} \quad \lambda > 0.$$

   (a) Find the maximum likelihood estimator (MLE) of the parameter $\theta = P(X = 0)$.

   (b) Show that $\hat{\theta} = (1 - 1/n)^Y$ is an unbiased estimator of θ, where $Y = \sum_{i=1}^n X_i$.

   (c) Derive the variance of $\hat{\theta}$. Does the variance of $\hat{\theta}$ attain the Cramér-Rao lower bound on the variance of unbiased estimators of θ?

   (d) If one can only observe

$$Z_i = \begin{cases} 1, & X_i > 0, \\ 0, & X_i = 0, \end{cases}$$

   find the explicit expression for the MLE $\hat{\lambda}$ of λ, as a function of $Z_1, \ldots, Z_n$, and derive the limiting variance of $\sqrt{n}\hat{\lambda}$ as $n \to \infty$.

   (e) Based on $X_1, \ldots, X_n$, one can claim that the MLE of λ is $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, and $\sqrt{n}(\bar{X} - \lambda)$ converges in distribution to a normal distribution with mean 0 and variance λ as $n \to \infty$. Show that $\sqrt{n}\bar{X}$ has a smaller limiting variance than $\sqrt{n}\hat{\lambda}$, and give a heuristic explanation of why this makes sense.

Points: 5 for each part.

4. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with mean 0 and variance σ^2. To test the hypothesis $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$, it is suggested that one can use

$$\delta(X_1, \ldots, X_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^n X_i^2 < c_1 \text{ or } \sum_{i=1}^n X_i^2 > c_2 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find c_1 and c_2 such that the size of δ equals a predetermined $\alpha \in (0, 1)$.

(b) Show that, for a certain choice of c_1 and c_2, the power function of the test in (a) is

$$\beta(\sigma) = G_n \left(\frac{\sigma_0^2 \chi^2_{n,\alpha/2}}{\sigma^2} \right) + 1 - G_n \left(\frac{\sigma_0^2 \chi^2_{n,1-\alpha/2}}{\sigma^2} \right),$$

where for the chi-squared distribution with n degrees of freedom, $G_n(\cdot)$ denotes the cumulative distribution function and $\chi^2_{n,p}$ the pth quantile.

(c) Prove or disprove that the test δ is the uniformly most powerful (UMP) test of its size.

(d) For testing $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma > \sigma_0$, find the critical region of the UMP test with test size α.

(e) What is the power function of δ when used to test the hypothesis in (d)? Show that δ is less powerful than the UMP test you derived in (d).

Points: 5 for each part.