2003 Honors Examination in Statistics

Department of Mathematics and Statistics
Swarthmore College

Name: ____________________

Instructions: This examination consists of six questions. Number the questions clearly in your work and start each question on a new page. You must make it clear how you arrived at your answer. Answers without any work may lose credit even if they are correct.

This is a closed-book three-hour examination. You may not refer to notes or textbooks.

1. In 2001, Woodmen Insurance Company paid an average of $85 per day when paying supplemental hospital benefits on claims sent to their Daily Hospital Supplement Benefit policy. In 2003, Woodmen sampled 40 records of payments through this policy, and calculated an average payment of $97 with a standard deviation of $42.

   (a) The company is interested in testing whether or not the average payment under this policy has increased. State hypotheses in symbols and words. Conduct the test and report your conclusion.

   (b) What is the approximate power achieved in this test? That is, find the rejection region (for what values would you reject the null hypothesis) and find the probability of rejecting, assuming that the alternative mean of $97 is true. You may draw pictures if it is helpful to illustrate what you want to calculate.

   (c) What are two things that affect power in this example? How do they affect it? You may describe their impact using a combination of words and pictures.

   (d) Suppose that Woodmmen Insurance can identify individuals who had claims for supplemental hospital benefits in both 2001 and 2003. Rather than collecting one sample in 2003, a proposal is made to sample 40 records from individuals who had claims in both 2001 and 2003 and use these data to evaluate whehter or not the average payment under this policy has increased.

      i. Describe how the change in the structure of the data changes the nature of the test. You can state hypotheses and a test statistic if it is helpful, but do not have to conduct a test.

      ii. The second sampling scenario likely will result in a more precise estimate of the average change, but it might not be more desirable. Comment briefly on the variability and bias of the estimate in the second case versus that in the original plan.

2. Two processes are being compared. The time until completion of each process is modeled as an exponential distribution. Independent samples are taken from each process. Let $x_1, \ldots, x_m \sim$ i.i.d. Exponential (β_1) and $y_1, \ldots, y_n \sim$ i.i.d. Exponential (β_2). The question of interest is whether or not the processes actually have different times until completion. That is, the manufacturer wants to know if the two processes have the same exponential distribution or not.

   The joint densities of the observations from the two separate samples are

$$f(\vec{x}|\beta_1) = \left(\frac{1}{\beta_1}\right)^m e^{-\sum_{i=1}^m x_i/\beta_1} \quad \text{and} \quad f(\vec{y}|\beta_2) = \left(\frac{1}{\beta_2}\right)^n e^{-\sum_{i=1}^n y_i/\beta_2},$$

   where $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$.

   The joint distribution of all observations if they come from one exponential distribution is

$$f(\vec{x}, \vec{y}|\beta) = \left(\frac{1}{\beta}\right)^{m+n} e^{-(\sum_{i=1}^m x_i + \sum_{i=1}^n y_i)/\beta}.$$

   (a) What are the maximum likelihood estimates of β_1, β_2, and β?

   (b) Consider just the first m observations, $\vec{x}$. Does the maximum likelihood estimate of β_1 achieve the Cramer-Rao Lowerbound (CRLB)? Explain briefly the importance of the CRLB in this example.

   (c) Form the likelihood ratio for testing whether there is one exponential distribution or there are two. Simplify to the degree possible.

   (d) How small does the ratio in the previous part have to be before you reject the hypothesis of one process at the $\alpha = .05$ significance level? Use a large sample approximation to determine the cutoff value for the ratio.

3. A closed population of apes is being studied by researchers. Each year the number of births in the population is related to the number of adult females. Assume the birth rates in a year follow Poisson distributions with a rates that depend on the number of adult females. The Poisson distribution in year t is given by

$$P(B_t = b) = (\lambda n_t)^b e^{-\lambda n_t}/b!,$$

where B_t is the number of births in year t, n_t is the number of adult females at the start of year t, and λ is a rate parameter. The colony is observed for T years; n_t, b_t is observed for $t = 1, \ldots, T$.

A prior distribution for λ is taken to be a Gamma distribution with parameters α and β:

$$g(\lambda) \propto \lambda^{\alpha-1} e^{-\lambda\beta}.$$

(a) What is the expected number of births in a year when the number of adult females is n_t? What is the variance?

(b) What is the posterior distribution of λ given data observed for T years?

(c) Suppose you have generated 10,000 independent values $(\lambda_1, \lambda_2, \ldots, \lambda_{10000})$ from the posterior distribution of λ given the data. How can you summarize the posterior distribution given these values? Be specific.

4. Let X have an exponential distribution with parameter β. Then $f_X(x) = \frac{1}{\beta} e^{-x/\beta}$. Let $Y = (\frac{2X}{\beta})^{1/2}$. The random variable Y has a Ralyeigh distribution. $X > 0$, $Y > 0$, and $\beta > 0$.

(a) What is the density of Y? Either the distribution method or the transformation method can be done in this case.

(b) The mean is Y^2 is 2 and its variance is 4.

i. Suppose you have 4 independent random variables (Y_1, Y_2, Y_3, Y_4) with the same distribution as Y. What are the mean and variance of $W = .1\ Y_1^2 + .2\ Y_2^2 + .3\ Y_3^2 + .4\ Y_4^2$?

ii. Suppose you have 2 random variables Y_5 and Y_6 with the same distribution as Y that have correlation 0.7. What is the standard deviation of $V = Y_5^2 - Y_6^2$? Explain intuitively why the standard deviation is less than $2\sqrt{2}$.

5. Suppose the density of random variable X is given by

$$f(x) = k(1 - x^3)$$

for $0 < x < 1$ and $f(x) = 0$ otherwise.

(a) Find k and sketch the density function.

(b) The mean of X is 2/5. What is the standard deviation of X?

(c) What is the probability that X is within 1.5 standard deviations of its mean? Either calculate this or report what Chebyshev's (Tchebysheff's) Theorem says and comment on how the statement relates to this density.

6. Let Y be a Poisson(λ) random variable. Let $I = 1$ if $Y = 0$ or $Y = 1$, and $I = 0$ otherwise.

(a) What is the expected value of I?

(b) What is the moment generating function (MGF) of I?

(c) Use the moment generating function to find $E(I^2)$.
(If you do not have (b), then use the MGF of a Poisson to find $E(Y^2)$.)