2008 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 answers. 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 any notes or textbooks.

You may use a calculator that does not do algebra or calculus.

Normal and t tables should be supplied with this exam.

1. Suppose that you ask a random sample of 36 male students at Swarthmore how many minutes they spend at the gym during a typical workout session. Using their data, you compute a 99% confidence interval for the population average and obtain 25.3 to 50.1 minutes. Assume that the data follow a normal distribution.

   (a) What is the standard deviation for these 36 men?

   (b) Below are five statements about the confidence interval. For each statement, if you believe that it is always true, simply write "true" on your answer sheet. Otherwise, write "false" and explain precisely why the statement could be false in <u>five or fewer</u> sentences.

      i. A 95% confidence interval made using the same data has a larger lower limit than 25.3 and a smaller upper limit than 50.1.

      ii. Approximately 99% of all male Swarthmore students work out between 25.3 and 50.1 minutes during a workout.

      iii. There is a 0.5% probability that the population average is greater than 50.1.

      iv. A 99% confidence interval obtained from a random sample of 100 male Swarthmore students has a better chance of containing the population average than a 99% confidence interval obtained from a random sample of 36 male Swarthmore students.

      v. If we took random samples of 36 male Swarthmore students over and over again, we would expect roughly 99% of the sample averages to fall between 25.3 and 50.1.

2. We want to determine if job training has a beneficial impact on earnings. To study this question, we randomly assign 800 people to get job training (the treated group) and 800 people not to get job training (the control group). The outcome variable is annual earnings one year after the training end date. The sample average for the treated group is \$23,500, and the sample average for the control group is \$22,500. The p-value for the one-sided hypothesis test for the difference of two means is 0.10. The earnings data are such that the central limit theorem applies.

   (a) Below are four statements about this p-value. For each statement, if you think that it is always true, simply write "true" on your answer sheet. Otherwise, write "false" and explain precisely why the statement could be false in <u>five or fewer</u> sentences.

      i. The alternative hypothesis is that the population average earnings for people who get job training exceeds \$22,500.

      ii. There is a 90% chance that the population average earnings for people who get job training exceeds the population average earnings for people who do not get job training.

      iii. There is a 10% chance that the population average earnings for people who get job training equals the population average earnings for people who do not get job training.

      iv. If the null hypothesis is true, there is a 10% chance that we'd observe a sample average earnings for the treated group that exceeds the sample average earnings for the control group.

   (b) There is a data entry error for one record in the control group. One person's earnings was entered as \$225,000 when it should have been \$22,500. If you re-do the hypothesis test with the corrected data, which statement is correct? The new p-value will be:

      - less than .10,
      - exactly the same as .10,
      - greater than .10,
      - possibly larger or possibly smaller than .10; one cannot tell from the information given.

      Justify your choice in <u>five or fewer</u> sentences.

3. A commonly used probability distribution for monetary random variables is the Pareto distribution. Assume all values of the random variable are greater than or equal to some baseline value k. The Pareto probability density function is

$$\begin{aligned} f(y) &= \theta k^{\theta} \left(\frac{1}{y}\right)^{\theta+1}, \quad y \geq k \\ f(y) &= 0, \qquad\qquad\qquad \text{otherwise} \end{aligned}$$

where $\theta \geq 1$. Suppose that you have five observations sampled from this distribution: $y_1 = 12, y_2 = 8, y_3 = 20, y_4 = 16, y_5 = 4$. Assume that $k = 3$.

(a) Determine the expected value of Y.

(b) Use the method of moments to estimate θ given the five data values.

(c) Determine the maximum likelihood estimate of θ given the five data values.

(d) Determine the probability density function of $Z = \log(Y)$, where log is the natural logarithm.

4. Suppose that you collect data $\{y_1, \ldots, y_n\}$ that are independent and identically distributed according to an exponential distribution with parameter λ,

$$\begin{aligned} f(y) &= \lambda e^{-\lambda y}, \qquad y \geq 0 \\ f(y) &= 0, \qquad\quad \text{otherwise.} \end{aligned}$$

One can show that $E(Y) = 1/\lambda$ and that $Var(Y) = 1/\lambda^2$.

(a) For large sample size n, what is an approximate distribution of $\bar{Y}$, where $\bar{Y}$ is the sample mean? Give an approximation for large n, not the exact gamma distribution.

(b) Using part (a), write an expression for a large sample, 95% confidence interval for $1/\lambda$.

(c) The maximum likelihood estimator for λ equals $1/\bar{Y}$. Suppose that we want to determine $Var(1/\bar{Y})$. We do so by means of a Taylor series approximation to the function $g(x) = 1/x$.

i. Write a first order Taylor series expansion of $1/\bar{Y}$ around $E(\bar{Y})$. A first order Taylor series of $g(x)$ around x_0 is of the form $g(x) \approx g(x_0) + g'(x_0)(x - x_0)$, where $g'(x_0)$ is the first derivative of $g(x)$ evaluated at x_0.

ii. Write an expression for the variance of the first order Taylor series expansion that you obtained in part (c.i).

(d) Suppose that you decide to use Bayesian inference to learn about λ. You use a prior distribution for λ that is also an exponential distribution with parameter equal to μ. Determine the posterior distribution of λ given the observed data and μ.

(e) Describe how you would use the posterior distribution in part (d) to simulate a 95% highest posterior density interval for $1/\lambda$. Assume that you have a software routine that can generate draws from exponential and gamma distributions.

5. Suppose the following linear model relates a response y and one predictor x:

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

for $i = 1, \ldots, n$. Let $\mathbf{y} = \{y_1, \ldots, y_n\}$ denote the $n \times 1$ vector of responses, and $\mathbf{x} = \{x_1, \ldots, x_n\}$ denote the $n \times 1$ vector containing the values of the predictor. You may assume that $\mathbf{x}$ and $\mathbf{1}_n$ (the $n \times 1$ vector of ones) are linearly independent, so that the $n \times 2$ matrix $\mathbf{X} = [\mathbf{1}_n\mathbf{x}]$ has rank two. All responses are independent.

(a) What is the distribution of the least-squares estimate of β_1?

(b) Construct centered response and predictor vectors by subtracting their means, i.e., set $\mathbf{y}_c = \mathbf{y} - \bar{y}\mathbf{1}_n$ and $\mathbf{x}_c = \mathbf{x} - \bar{x}\mathbf{1}_n$. True or false: There exist real numbers α_0 and α_1 such that

$$E(\mathbf{y}_c|\mathbf{X}) = \alpha_0\mathbf{1}_n + \alpha_1\mathbf{x}_c.$$

If true, find $\alpha = (\alpha_0, \alpha_1)$ in terms of $\beta = (\beta_0, \beta_1)$; if false, show why.

(c) Instead, transform the response by squaring it, so that $\mathbf{y}^2$ is the vector containing the squared elements of $\mathbf{y}$ (without centering). Transform the covariate vector in the same way, yielding $\mathbf{x}^2$. You wish to fit a linear model for $\mathbf{y}^2$ in terms of $\mathbf{x}^2$. True or false: There exist real numbers γ_0 and γ_1 such that

$$E(\mathbf{y}^2|\mathbf{X}) = \gamma_0\mathbf{1}_n + \gamma_1\mathbf{x}^2.$$

If true, find $\gamma = (\gamma_0, \gamma_1)$ in terms of β; if false, show why.

6. In World War II, U.S. Army doctors devised a new procedure to test whether groups of soldiers had a disease (it was syphilis). Here's how it worked.

   Suppose that there are n soldiers. The doctors take blood from all n soldiers, then pool the blood in one sample that is tested once for the disease. If the test on the pooled sample is negative, the doctors proclaim the group is free of disease and stop the procedure. If the test on the pooled sample is positive, the doctors take new blood from each soldier and test each individual's blood (i.e., they do n individual tests) to see which ones have the disease.

   Suppose that 20% of all soldiers in the Army had the disease. Assume the outcomes of tests are independent given disease status.

   (a) Assume that the test is perfect for the pooled and individual samples, i.e. it has 0% chance of false positives or false negatives. Find the expected number of tests that are done with this procedure for $n = 10$ randomly selected soldiers.

   (b) Suppose that the pooled test is not perfect. For the pooled sample, if at least one soldier has the disease, there is a 20% chance that the pooled test is negative. The false negative probability and its complement are the same regardless of how many have the disease (as long as at least one does). However, if no soldiers have the disease, there is a 15% chance that the test is positive. Find the expected number of tests that are done with the procedure for $n = 10$ randomly selected soldiers.

   (c) Assume the imperfect pooled test described in part (b). Given that the test on a pooled sample from $n = 3$ randomly selected soldiers is positive, what is the chance that none of them have the disease? Note that you are now working with $n = 3$, not $n = 10$.

   (d) Assume the imperfect pooled test described in part (b). Further assume that the test is not perfect for individuals. When testing any individual soldier who has the disease, there is a 5% chance that his test is negative. When testing any individual soldier who does not have the disease, there is a 25% chance that his test is positive.

   Given that the test on a pooled sample from $n = 3$ randomly selected soldiers is positive, and that the tests for each of the three individual soldiers are all negative, what is the chance that none of them has the disease?