Stat 430 Homework 2

due: October 16, 2009 at 5pm

1. Access the reading scores data. These data are the reading scores of grade school students after they received a directed reading activities ("Treated") or not ("Control"). Construct 99% confidence interval for the difference in population means between treated and control. Report the p-value for testing the null hypothesis of equal population means. What is the power to detect a score difference of 5?

2. For independent samples

$$\begin{aligned} X_1, \ldots, X_n &\overset{\text{iid}}{\sim} N(\mu_X, \sigma_X^2) \\ Y_1, \ldots, Y_n &\overset{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2) \end{aligned}$$

   show that $\bar{X} - \bar{Y}$ is the maximum likelihood estimate for $\mu_X - \mu_Y$.

3. In this problem, you will explore the shape of the probability plot for various normal and non-normal distributions and additional tests for normality.

   (a) Simulate data sets of 100 standard normal random using `R`'s `rnorm()` function. Prepare probability plots (x-axis: standard normal quantiles; y-axis: observed and ordered random variables) for three such data sets.

   (b) Simulate 100 random variables from a Gamma(shape= 0.471, rate= 1) distribution using `R`'s `rgamma()` function. Prepare a probability plot (x-axis: standard normal quantiles; y-axis: ordered gamma-distributed random variables). Based on this plot, how does this distribution differ from the normal? Save this data for part d.

   (c) The double exponential distribution has pdf

$$f(x) = \frac{1}{2} e^{-|x|} \qquad\qquad -\infty < x < \infty.$$

   Compare this pdf with the normal pdf. Do the tails of this distribution decay slower or faster? Generate 100 random variables from the double exponential distribution by (1) finding its cdf $F(x)$, (2) deriving the inverse cdf $F^{-1}(z)$, (3) generating 100 standard uniform random variables $Z_1, \ldots, Z_{100}$ using `R`'s `runif()` function, and (4) mapping them to exponential random variable $X_1 = F^{-1}(Z_1), \ldots, X_{100} = F^{-1}(Z_{100})$ using the inverse cdf. Save this data for part d. Prepare a probability plot. Does it confirm your suspicions about the tails?

   (d) The coefficient of skewness (a measure of skew) is estimated as

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{ns^3}$$

   where s^2 is the sample variance. For normal random variables, this coefficient should be 0. Estimate this coefficient for the gamma-distributed random variables from part b and the double exponential-distributed random variables from part c. The question is whether either of these statistics is unusually different from 0.

A theoretical sampling distribution for b_1 is difficult to obtain, but the computer can help you by allowing you to sample coefficients of skewness under the assumption of $n = 100$ iid normal samples. To do this, generate $n = 100$ iid standard normal random variables, and compute the coefficient of skewness. Repeat $B = 1000$ times to obtain bootstrap skewness coefficients $b_1^{(1)}, \ldots, b_1^{(1000)}$. The fraction of bootstrap coefficients that are as extreme or more extreme than the coefficient computed on the original data is the p-value for rejecting the null hypothesis. Test for skewness of the gamma and double exponential random variables using this test.

Repeat the above procedure for the coefficient of kurtosis (a measure of peakedness or flatness), which is estimated as

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{ns^4}$$

4. Access the Titanic data to answer the following questions.

   (a) Use a t-test to test whether there is a difference in the mean age of people who survived and did not survive the disaster. Is there a difference in the mean age of people in first and second class (variable "PClass")?

   (b) Prepare probability plots to evaluate the appropriateness of the normal assumption for the two tests above. Which test results are most suspect? Can you think of a transformation that might fix this problem? (No need to implement, just discuss.)

   (c) Perform a test of whether the proportion of children two years and under among the survivors is different from among the non-survivors.

   (d) When data normality is suspect, the computer can help perform tests without distributional assumptions. For example, if there are no differences between the surviving and non-surviving populations, the combined age data can be viewed as a sample from a single population that is arbitrarily split into survivors and non-survivors. Use `R`'s `sample()` function to randomly sample 1000 of these assignments. For each one, compute a statistic sensitive to differences in population means. Count the number of times (out of 1000) where the statistic is as extreme or more extreme than the statistic computed on the original data. Divide this count by 1000 to obtain a p-value for testing the null hypothesis. How do the assumptions of this test differ from the assumptions of the two-sample t-test?

   (e) [added] Using the extended Titanic data, pair each survivor with a random non-survivor with the same embarkation location (consider two missing embarkation locations as matched). You will need to discard several non-survivors that have no match. Is the p-value for rejecting the null hypothesis of no age difference smaller or larger? Propose an explanation for what you find.

   (f) [added] Use one of the non-parametric tests we learned in class to test for a difference in mean age of survivors and non-survivors.

(g) [added] Questions a through f all address the question of whether age affected passenger survival. What is your overall conclusion? Why?

(h) [added] Use a Fisher's exact test to test whether survival is related to sex and, separately, class (combining 1st and 2nd class together).

5. Form teams of 1 to 4 members and prepare a first proposal of your class project. You should identify a dataset and questions you would like to address with this dataset. If you don't have good ideas for datasets, I may be able to provide you with one. In this case, let me know what type of statistical expertise you would like to develop in this project. Try to be at least a little specific.