PSTAT 105: Lab D Due November 30, 2017

Please type up a report showing your analysis and submit it through GauchoSpace.

1. The `Drug Treatment` data set followed up on 628 people who had participated in a drug treatment program. The first column of the data is the number of days until the first relapse into drug use or the number of days until that subject was no longer followed. The second column indicates whether the patient relapsed (code =1) or the data was censored (code =0). A censored observation is to be interpreted as the subject went longer than X days before a relapse. We are going to use the `survival` library in analyzing this data set.

   (a) Use the `Surv` function to create a survival analysis data set. Looking at the data, describe the relationship between the number of days and whether or not the data is censored.

   (b) Use the `survfit` to calculate the Kaplan-Meier estimator of the survival function and plot it.

   (c) Calculate the ECDF for this data. Use it to estimate the median and quartiles of the data.

   (d) The third column of the data categorized each subject as to whether they have Never used IV drugs, they had Previous IV drug use some time in the past, or they have Recent use of IV drugs. Some subjects were not categorized. They are labeled "Dont Know" and we are going to ignore them in this question. Calculate separate median and quartile estimates for each of three categories.

   (e) A variance formula due to Greenwood is used by `survfit` to give a confidence interval around each point in the estimated survival function. Use these confidence intervals to fashion confidence intervals for the upper quartiles in each of the three categories as calculated in part D). Do these intervals suggest that there is a statistically significant difference between the three categories of subjects?

2. We are going to use a data set `lung` which is available from the `survival` library. This data set contains the `lung$time` vector of measurements of survival time in days and `lung$status` which is 1 for censored observations and 2 for uncensored observations.

   (a) Use `survfit` to plot the Kaplan–Meier estimator of the data.

   (b) Calculate an estimate and a 95% confidence interval for the survivor function at 150 days.

   (c) Calculate an estimate and a 95% confidence interval for the median survival time.

   (d) The data set also includes the gender of each patient in the vector `lung$sex` which are coded as 1 for male subjects and 2 for female subjects. Plot separate estimators of the survival function for men and women. Generally speaking, do men or women have better survival rates? Is this consistent throughout the time of the study?

   (e) Calculate separate estimates and 95% confidence intervals for the median survival time for men and women. What would you conclude about the difference between male and female survival rates from these intervals? Do you think this tells the whole story?

3. Use a simulation of one million random trials to approximate the probability for an exponential random variable X with $\mathbb{E}(X) = 1$ of the event $A = \{\cos(X) > 0.7\}$. Calculate a margin of error for this approximation by using 95% confidence interval from a binomial experiment.

4. We want to calculate a critical value for a Goodness of Fit test based on the Kolmogorov–Smirnov test for an exponential distribution.

   (a) Generate 1000 samples where each consists of 50 independent exponential random variables with mean 1. Estimate the mean of each sample. Draw a histogram of the means.

   (b) Perform a KS test on each sample against the null hypothesis that they are from an exponential random variable with a mean that matches the mean of the data set. Draw a histogram of the 1000 values of D.

(c) From the simulated values of D, find an critical value c such that only 5% of the time will the test statistic D exceed that critical value.

(d) Write up `R` code that will perform multiple batches of 1000 samples each consisting of 50 independent draws. Decide how many batches your computer can run in two hours, and run the program. Use the results to generate a more accurate estimate of c.

(e) Calculate the critical value that is suggested in Table 1.4 of Stephens(1974) paper. Compare this to the results from your simulations. Do you think this value is significantly different from your simulation results?

5. Using our Old Faithful data set, we want to determine the accuracy of an estimate of the spread of the data.

   (a) Calculate the sample standard deviation of the times between eruptions from the geyser data set.

   (b) Use the `sample` function to draw 1000 bootstrap samples (with replacement) and use these to estimate the variance of our standard deviation estimator.

   (c) The IQR measures the distance between the first and third quartile of the data and rescales it to be an estimate of the standard deviation. Use the `IQR` function to estimate the standard deviation of the geyser data, and then use the bootstrap to estimate the variance as we did above.

   (d) The median absolute deviation is a third way to estimate the standard deviation of a data set. Use the `R` function `mad` to find the estimate of the standard deviation. Use the bootstrap again to estimate the variance of this estimator.

   (e) Which estimator do you think is the best among these three for estimating the standard deviation of our geyser eruption data?