PSTAT 105: χ^2 Goodness of Fit Tests Due October 19, 2017

Please submit your report on these two data sets through the GauchoSpace page by the end of the day on Thursday.

1. A well-known analysis in Malcolm Gladwell's book *Outlier* argues that the best hockey players are more likely to be born earlier in the year presumably because this gives them advantages in the youth hockey leagues. We are interested in checking whether there is a similar effect in basketball.

   (a) The data set `BballBirths.txt` contains the names and the month of birth for all the professional basketball players listed on the `http://www.basketball-reference.com` web site. Use the `table` function to calculate how many players were born in each month. Draw an appropriate plot.

   (b) Perform a χ^2 test to see if the players are equally likely to be born in any month.

   (c) In order to focus our attention on modern players, repeat this analysis with only those players that were born after 1/1/1955. (also use this smaller data set for the following questions.)

   (d) To be more careful, we should realize that more people are probably born in January than February just because there are more days in January. Perform a χ^2 test where the null hypothesis is that the probability of each month is proportional to the average number of days in that month.

   (e) Going even further, it seems that some months generally are favored over others for having babies (summer births are more likely). We should probably compare our basketball player data to the following probabilities from the CDC.

Month	Jan	Feb	Mar	Apr	May	Jun
Prob.	0.0815	0.0752	0.0837	0.0816	0.0859	0.0813
Month	Jul	Aug	Sep	Oct	Nov	Dec
Prob.	0.0883	0.0892	0.0866	0.0849	0.0787	0.0830

   Perform a χ^2 test to see if the basketball player data has the same distribution.

   (f) Interpret your results. Is there significant evidence at an $\alpha = 0.05$ level that professional basketball players are born earlier in the year than the normal population?

2. The data set `Selltimes.txt` consists of the time that elapses between when sell orders for CISCO stock were placed during April 5, 2010. My hypothesis is that these times have an exponential distribution with CDF

$$F(t) = 1 - \mathrm{e}^{-\lambda t}$$

   for some unknown rate λ.

   (a) Use the `hist` function to plot an informative histogram of the data.

   (b) Calculate the MLE, $\hat{\lambda} = \bar{x}^{-1}$, from the data.

   (c) Use this estimate of λ to divide the sample space into 10 intervals that will be big enough that the χ^2 approximation will be appropriate.

   (d) Use the `hist` function to count the number of observations in each of those intervals.

   (e) Perform the appropriate χ^2 test

   (f) Inspect the counts and the expected values and give some description of how the data looks different from an exponential distribution.

   (g) What difference does it make if we used 25 or 100 intervals instead of 10? Experiment a little with different sets of intervals and report the results and whether they demonstrate anything different from the original 10-interval analysis.