STAT-S 470/670 Homework 3

Tentative Due Date: Wednesday 9/23, in class

1. For this first problem utilize the ”counties” data set within the ”noncensus” package of R.

```
install.packages("noncensus")
library(noncensus)
data(counties)
```

   (a) Construct a level versus spread plot for this data set by calculating the median county population M and fourth spread d_F for each state, then creating a scatter plot of $log(d_F)$ versus $log(M)$ for each state. Fit a least squares line to this data using lm($log(d_F) \sim log(M), data$) in R and then display the plot, the equation for the best fit line and the scatter plot.

   (b) Based upon the slope of the best fit line you found what should the recommended transform $T_p(x)$ be so that the boxplots have approximately uniform spread?

   (c) Display the boxplots for all 50 states without transform and under the recommended power or log transform $T_p(x)$ that you found in part b.

   (d) Now let us restrict our attention to the California subset. You would like to make this data approximately symmetric. For the California subset, construct a table just like you see in table 4-2 or your UREDA text book. You can use the letter value program for this and the table should look like this

Table 1: Transforms for symmetry

Letter Value	x_L	x_U	$\frac{(x_L+x_U)}{2} - M$	$\frac{(x_L-M)^2+(x_U-M)^2}{4M}$	Estimate of P
F					
E					
D					
C					
⋮	⋮	⋮	⋮	⋮	⋮

   (e) Using the table you constructed above plot the variable $\frac{(x_L+x_U)}{2} - M$ on the y-axis and the variable $\frac{(x_L-M)^2+(x_U-M)^2}{4M}$ on the x-axis. Based upon Tukey's theory and equation (8) on page 108 of UREDA, what would be the recommended power transform of the data for purposes of symmetry? How does the transform for symmetry compare to the transform for level spread?

   (f) Now based upon the transform you found to level the spread in part (b), find linear constants a and b so that a new transform $Z = a + bT_p(x)$ has the same expected value as the original California data. When doing this you might want to refer to the equations from 117 to 121 in the UREDA textbook.

2. In this problem you should analyze the CEO total compensation data set (ceo.txt). You can read in the data using data = read.table("ceo.txt").

   (a) How many CEO's total compensation are there in the dataset? What is the highest paid CEO paid? Are there any "unusual" values?

   (b) Present graphical display(s) for the data, and describe the distribution. Explain whether or not you want to transform the data.

   (c) What transformation(s) would you propose for this dataset, and why do you choose those?

   (d) Before transforming the data, do you need take care of the "unusual data", if you find any?

   (e) Present graphical display(s) for the transformed data (try at least two transformations). What have changed compared to the original data? What do you gain from the transformations?

   (f) If you can only choose one transformation, which one do you prefer?

3. Do problem 6 on Pages 127-128, UREDA.