Computer review for final exam

Load the data sets:

```
> source("http://www.math.csi.cuny.edu/st/R/baseball.R")
```

This gives two data sets: `hall.fame` and `MLBattend`. Attach them both with the commands:

```
> attach(hall.fame)
> attach(MLBattend)
```

To see the variable names that were loaded you can type `names(hall.fame)` or `names(MLBattend)`.

1 summary statistics

We learned how to summarize a data set using various numeric and graphical summaries: the mean, standard deviation, median, stem-and-leaf, boxplot, histogram etc.

Make histogram of `AB` the number of atbats. Is it a symmetric data set? Skewed?

Make a boxplot of `HR` (home runs). Is it symmetric? Skewed? From a normally distributed population?

What is the mean number of `doubles`. Compare to the median number. Which is more. Look at the histogram and see if you could have guessed which was more.

Find the standard deviation of the number of triples. Then find the percent of values less than 2 standard deviations from the mean. This can be found with a command like

```
sum(abs(triples - mean(triples)) < sd(triples))/ length(triples)
```

Would you guess that the distribution of the number of triples is normally distributed? Why not? Can you check?

Make a stem and leaf diagram of attendance, wins and losses for the New York Yankees. To get the values do this:

```
> nyy.attend = attendance[franchise == "NYA"]
> nyy.wins = wins[franchise == "NYA"]
> nyy.losses = losses[franchise == "NYA"]
```

Describe the distribution of values.

2 statistical inference

Perform a significance test to see if the population mean number for carreer RBIs is 550 against a two sided alternative. What is the p-value? Do you accept or reject at the 0.05level? Assume the data is a random sample of all major league baseball players.

In the data set are 254 catchers out of 1340 players. Is this proportion consistent with a population proportion of 1/9th? Perform a significance test of proportion to decide. What is the p-value? Do you accept or reject at the 0.05 level?

We want to compare statistics for both the national and american leagues. First figure out who is who:

```
> AL = league == "AL"
> NL = league == "NL"
```

Now perform a two-sample t-test for `attendance` and `runs.scored`. Do you accept or reject at the 0.05 level?

To get the attendance for each league is done with

```
> att.nl = attendance[NL]
> att.al = attendance[AL]
```

3 correlation, regression

Make a scatterplot of caught stealing `CS` and stolen bases `SB`. Guess the correlation.

Make a scatterplot of `doubles` and `triples` Guess the correlation, then confirm your guess with `cor()`.

Make a scatterplot of `losses` versus `attendance`. Guess the correlation, then confirm you guess with `cor()`.

Plot `SO` (number of strikeouts) against `HR`. Add the regression line using `abline(lm(HR ~ SO))`. Dies this say to hit more home runs you should strike out more?

Plot `BA` (batting average) versus `OBP` (on-base percentage). Is there a linear relationship? What is the estimated regression line? Use it to predict the `OBP` for a "300" hitter.

Come up with some questions on your own. Can you find other linear relationships? Look at the variable names with `names(hall.fame)` and guess. What do you find?