Baseball statistics

Wow Alex Rodriguez and Derek Jeter together in the same infield. I know most of you don't care, but how often are two hall of fame shortstops playing on the same team? I don't know if it ever happened before.

As we are all excited about this, lets look today at some data related to the game of baseball. First, we load the data sets. Type as shown:

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

This loads to data sets, `MLBattend` which has attendance records for Major League Baseball from 1969-2000, and `hall.fame` which has player statistics for over 1300 players.

The goal of this project is to use boxplots, and z-scores to look at data sets.

0.1 box plots

A box plot compactly displays the five-number summary of the min, Q_1, Median, Q_3 and the maximum. With these five numbers we can describe the center, the spread and symmetry of a distribution.

For a data set `x`, the five number summary is given by `fivenum(x)` and the boxplot drawn with `boxplot(x)`.

For example, after loading in the data above, if you type

```
> attach(MLBattend)
> attach(hall.fame)
> fivenum(OBP)
[1] 0.1940 0.3150 0.3350 0.3575 0.4830
> boxplot(OBP)
```

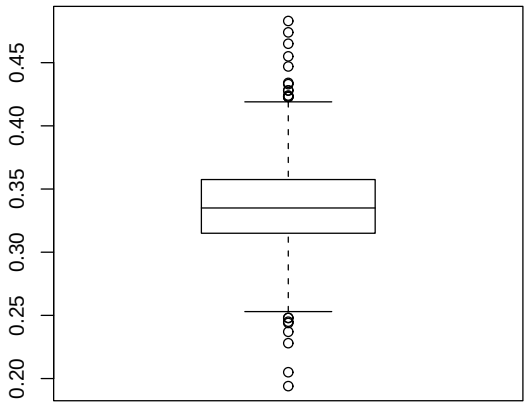

Figure 1: Box plot of On Base Percentage (`OBP`)

The extra dots on the end of the whiskers show points that are more than 1.5 IQR's beyond the first and 3rd quartiles. These are sometimes called outliers.

In this case they are the baseball players with better or worse On Base Percentages. To find out which players they are, you can use the following trick

```
> last[OBP > .45]
[1] "HAMILTON" "McGRAW"   "RUTH"     "WILLIAMS"
```

We see Babe Ruth and Ted Williams amongst the top four. This command found the last names (`last`) for those with `OBP` more than .45.

Question 1: Make a boxplot of batting average, `BA`. Describe the shape as symmetric or skewed. Find out the names of all the players with batting averages above .350.

Question 2: Make a boxplot of home runs, `HR`. Describe the shape as symmetric or skewed. Find out the names of all the players with more than 600 home runs.

Question 3: From the shape of the boxplot for `HR` do you expect the mean or median to be larger? Compare.

The commands

```
> hfm = Hall.Fame.Membership != "not a member"
> names(hfm)
```

will store a value of TRUE or FALSE for whether someone is in the Hall of Fame. We can check if the top players are in the hall of fame now with

```
> hfm[OBP > .450]
HAMILTON   McGRAW     RUTH WILLIAMS
    TRUE    FALSE     TRUE      TRUE
```

So a high on-base percentage is not guarantee of acceptance into the hall of fame.

Question 4: Repeat the above to see if hitting more than 550 home runs is enough to get into the hall of fame. What about a batting average over .350?

1 Multiple boxplots

Recall, the *z*-score of data point was the number of standard deviations away from the mean it is. The `scale()` command finds these for us. A boxplot of *z*-scores is identical in shape – but not scale – to a boxplot of the original data. However, if we use `scale()` we can compare boxplots for their shape.

For example to compare batting average, `BA`, to on-base percentage, `OBP` we do

```
> boxplot(scale(BA), scale(OBP))
```

Question 5: Do the box plot above. Are the shapes similar?

Question 6: Do a boxplot of the *z*-scores for both `HR` and `BA`. Do they have the same shape?

Question 7: DO the following commands

```
> boxplot(scale(hits),scale(hits/games))
```

This compares the distribution of hits and the number of hits per game. Are the shapes similar? Explain.

Multiple boxplots are great for comparing data sets. The `MLBattend` dataset contains attendance data several years. We can make boxplots of attendance data for each year quite easily if we use some special notation. Type the following exactly as shown. (That is use a tilde)

```
> boxplot(attendance ~ year)
```

This show boxplots for each year from 1969 to 2000. Annoyingly, the year 2000 shows up as 0 on the left. Let's not fuss about that. What do the box plots tell us?

Question 8: Are there any trends in attendance that you can describe? Are you comparing the medians, the spreads?

Question 9: There were three strikes during this period which shortened the season. Can you identify when these occurred by looking at your plot?

Question 10: To see if the number of wins changes the attendance, type these commands

```
> w = cut(wins,c(0,82,90,95,100,120))
> boxplot(attendance ~ w)
```

Describe any trend in the data. Does it match your intuition?

Question 11: The command

```
> boxplot(attendance ~ franchise)
```

will produces boxplots of attendance for each franchise. Which franchise has the best attendance?

Question 12: The command

```
> boxplot(attendance ~ league)
```

shows attendance for each league. Is it similar or different?

Question 13: The commands

```
> hr = cut(HR,c(0,200,300,400,500,600,800))
> boxplot(BA ~ hr)
```

create boxplots for different amounts of home runs hit. Are there any trends?