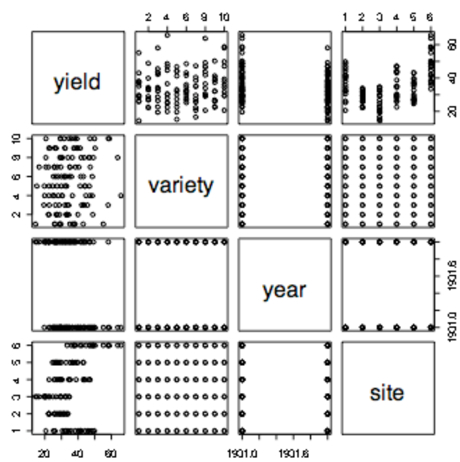
STAT 400 Data Analysis Sept 16, 2005

Today: Discussion of Barley Data Display

First, the problem of getting the data into a stat package like R.

Helps to use Excel's "Text to Columns". Also, use Word Edit to eliminate blanks.
Then save as a .txt file.

Once the four variables are entered via "read.table", a natural first step is
plot(barley.df)



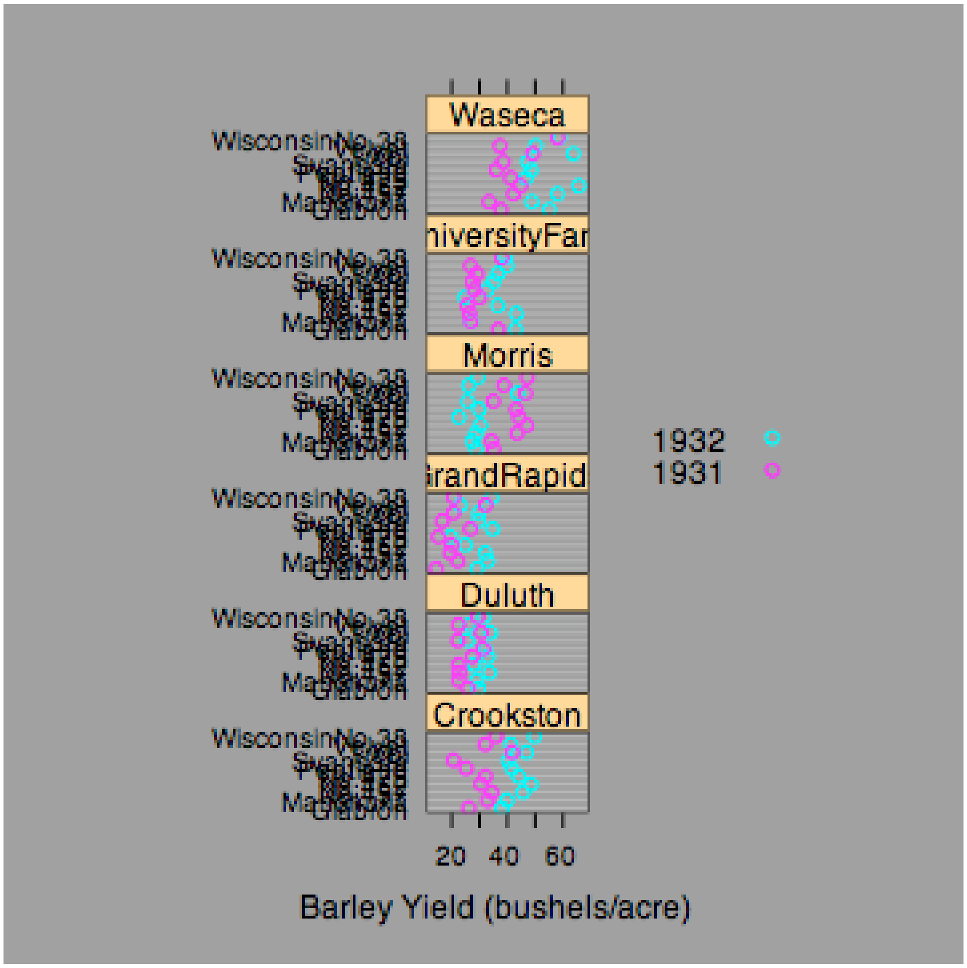

but not too revealing here.

To use the Lattice graphics, you need to add the packages lattice and grid (easy to do from within R).

Cleveland's plot in R is given by

```
dotplot(variety ~ yield | site, data = barley, groups =
year,
        key = simpleKey(levels(barley$year), space =
"right"),
        xlab = "Barley Yield (bushels/acre) ",
        aspect=0.5, layout = c(1,6), ylab=NULL)
```

and this produces the chart like p 4 (it looks better when full sized.)



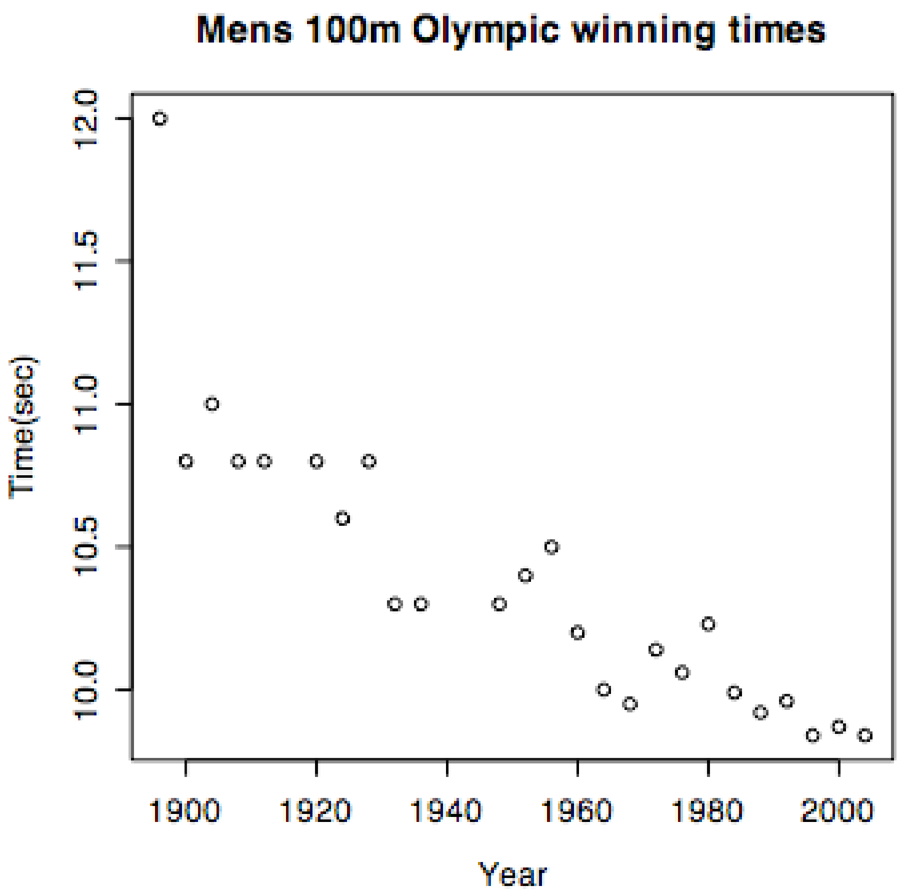

Monotone Spread: (p 47 ff)

This is a property of univariate data that may be exhibited when there are two or more groups. The usual version of this property is that the larger-mean-groups have larger variability. For example, the price of a \$100 stock is likely to vary more than the price of a \$10 stock (in absolute terms). Sometimes the percentage variability is about the same however. If one looked at the logarithm of the daily prices of the stocks, they might well have comparable variability. See pp 47-53.

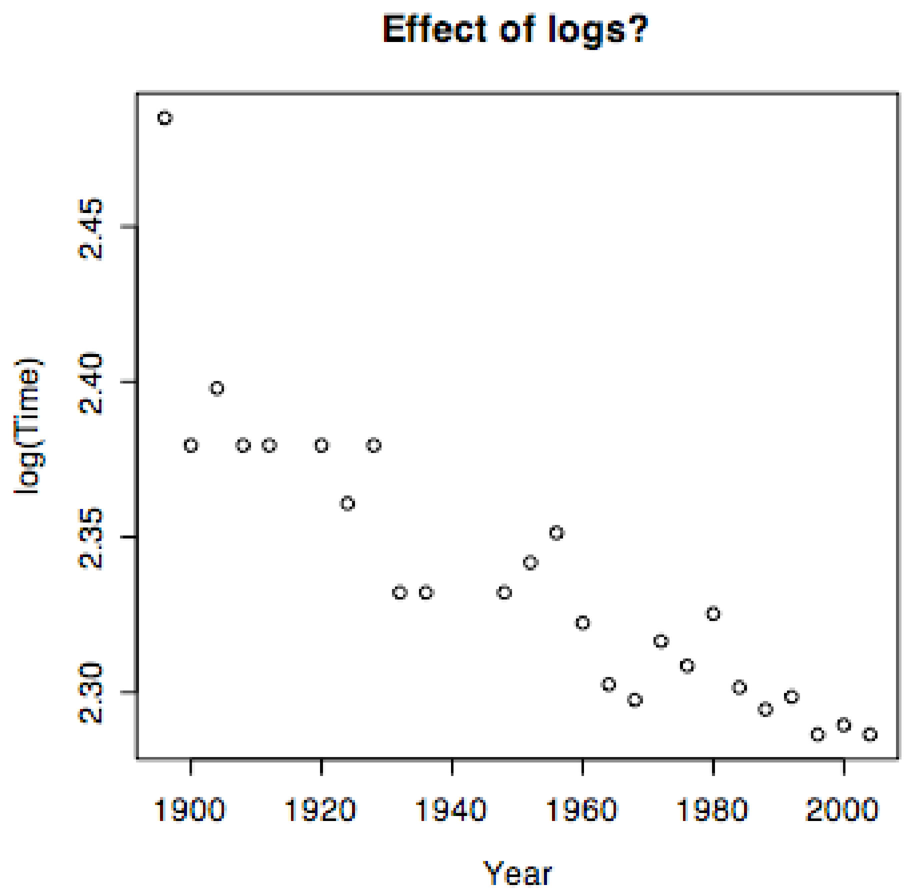
**Power Transformations**: Data $x_1, x_2, \ldots, x_n$ (pp 56-67)

Let $y_i = x_i^t$ . For these transformations with $t<1$, we need $x>0$.
If {x} are skewed right (long tail right) then for $0 \le t < 1$, {y} will be less skewed to the right. Note t -> 0 gives the logarithm transformation.
If {x} are skewed right (long tail right) then for $t<0$, it is not clear what skewness y will have – it depends on how close to zero the smallest values of x are.
The 1/x transformation is often useful when the data is recorded in un-natural units.
Note the force of a power transformation depends on the ratio of largest to smallest value.
(p 67) Here is some data from the 100 m Olympics 1896-2004.

| | Year | Time(sec) |
|---|---|---|
| 1 | 1896 | 12.00 |
| 2 | 1900 | 10.80 |
| 3 | 1904 | 11.00 |
| 4 | 1908 | 10.80 |
| 5 | 1912 | 10.80 |
| 6 | 1920 | 10.80 |
| 7 | 1924 | 10.60 |
| 8 | 1928 | 10.80 |
| 9 | 1932 | 10.30 |
| 10 | 1936 | 10.30 |
| 11 | 1948 | 10.30 |
| 12 | 1952 | 10.40 |
| 13 | 1956 | 10.50 |
| 14 | 1960 | 10.20 |
| 15 | 1964 | 10.00 |
| 16 | 1968 | 9.95 |
| 17 | 1972 | 10.14 |
| 18 | 1976 | 10.06 |
| 19 | 1980 | 10.23 |
| 20 | 1984 | 9.99 |
| 21 | 1988 | 9.92 |
| 22 | 1992 | 9.96 |
| 23 | 1996 | 9.84 |
| 24 | 2000 | 9.87 |
| 25 | 2004 | 9.84 |

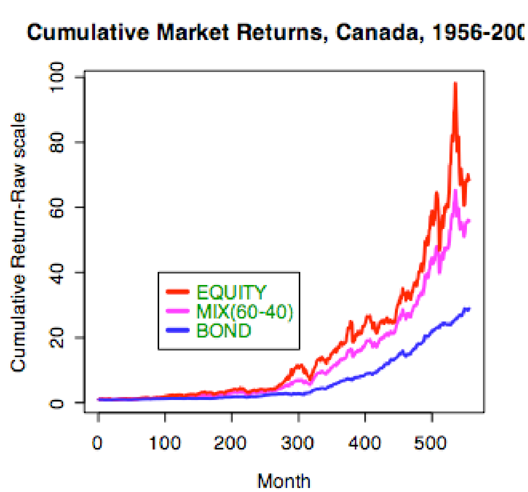


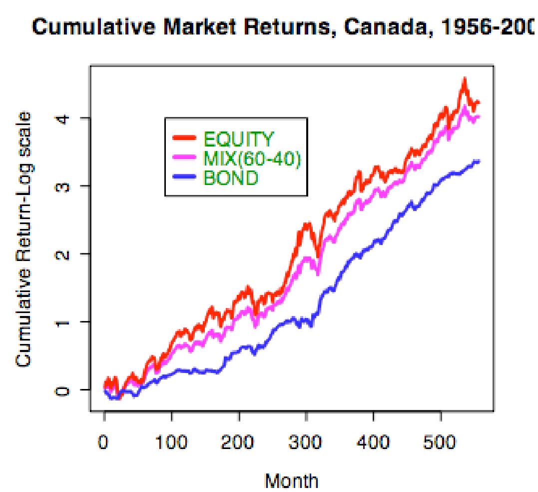

What happens when we take logs of the Times?



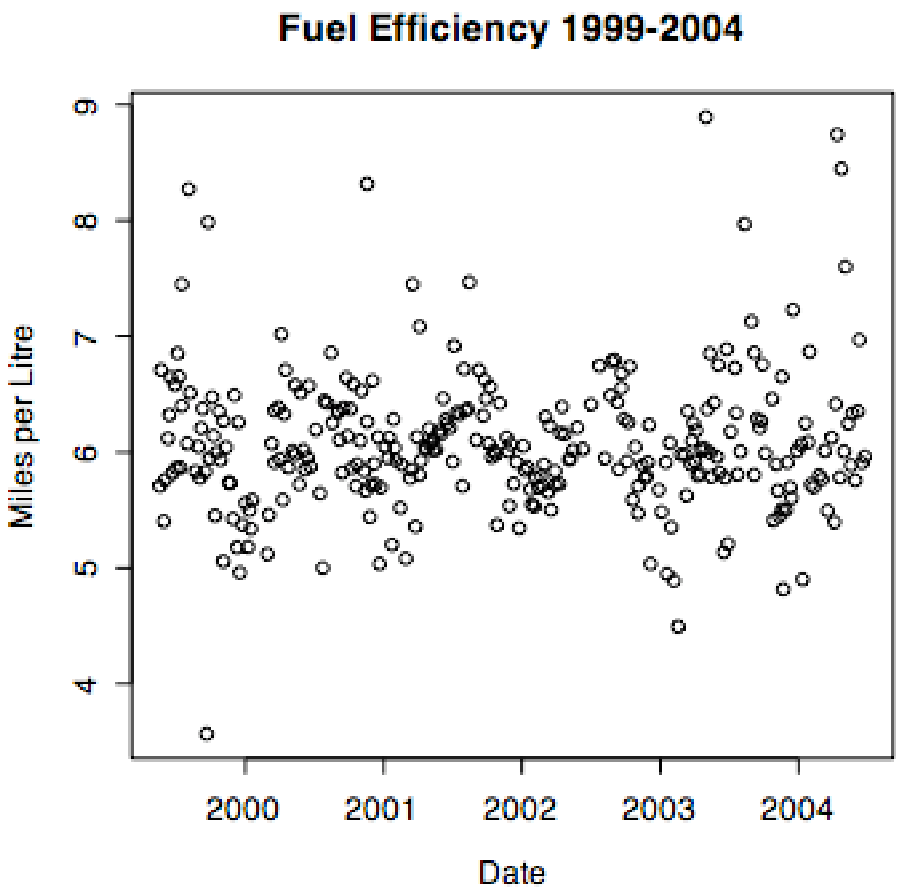

A better example of the utility of the log transformation is with stock price indexes.

TSE 300 (Canadian Equity Index)



I have sent you a data set by e-mail called "mercedes.txt".  Here is a graph of it:




Here is an exercise to hand in Friday Sept 23, at class.

Re-graph the data including a smoothed version.  Write a short description of what the graph shows.  Comment on any anomalous features of the data.

Read up to p 100. Look at "loess".