

Chapter 5 – Understanding and Comparing Distributions

August 25, 2010

In Chapter 4 we learned about graphical representations and numerical summaries for Quantitative Variables. We also learned to describe the distribution of a data set in terms of the Shape, Center, and Spread.

If the distribution was symmetric we used the Mean for the Center and the Standard Deviation for the Spread

If the distribution was skewed or had outliers we used the Median for the Center and the IQR for the Spread.

Can we describe more about the data? We can use the **5-number summary** (p. 89)
Minimum Score, Q_1 , Median, Q_3 , Maximum Score.

The **Boxplot or Box and Whisker Plot** (p. 89) is a graphical representation of the 5-number summary.

Boxplots are useful when comparing different groups of data.

Creating a Boxplot:

1. Draw an axis and label appropriately.
2. Draw a box with the ends being the Q_1 and Q_3 and a line in the middle of the box for the median. The box shows where the data in the 25th percentile to the 75th percentile are located or the middle 50%.
3. Calculate **Upper and Lower “Fences” or Limits** to determine which points are outliers. Outliers will be any values outside interval formed by fences.

$$\text{Upper Fence} = Q_3 + 1.5 \times \text{I.Q.R.}$$

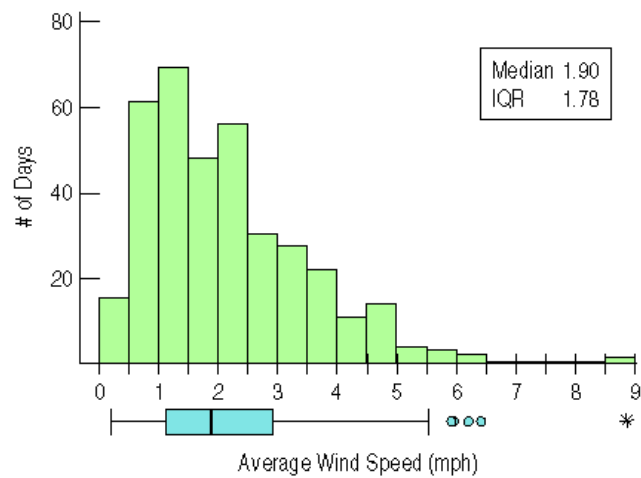
$$\text{Lower Fence} = Q_1 - 1.5 \times \text{I.Q.R.}$$

4. Mark outliers with a special symbol (*). Draw whiskers out to smallest and largest values of data that are not outliers. (Sometimes a different symbol is used for “Far Outliers” – data values farther than 3 IQRs from the quartiles.)

Example: The following data represent the weight (in pounds) of 15 five year old girls.

30 32 34 35 36 36 37 38 41 42 44 45 47 62 65

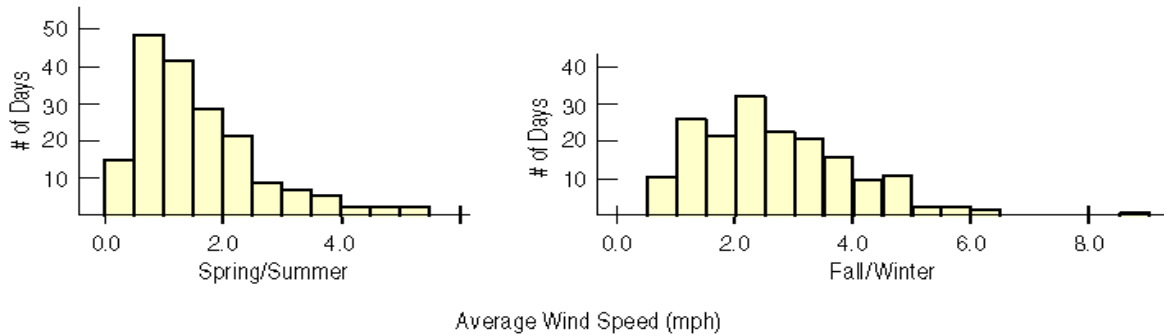
Histogram and Boxplot for daily wind speeds (Fig 5.2, p. 91).



How does each represent distribution?

Comparing Groups

Histograms (p. 91) - Axis should be similar. Note the shapes, centers, and spreads of the distributions. (Fig 5.3, p.91)



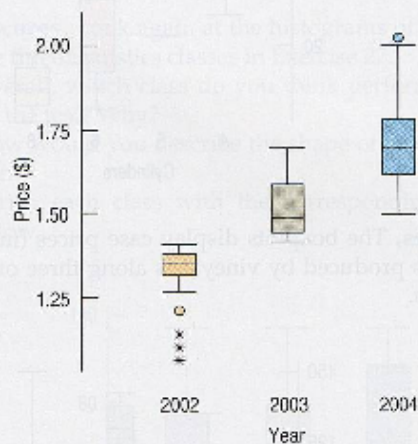
Back to Back Stem and Leaf Plot - Stem plot used to compare two different data sets by putting one set of leaves to the left of the stems and the other set to the right of the stems.

(Example: see p. 111, #15)

Boxplots (p.92) - Use same axis and plot boxplots for different groups side by side or above each other.

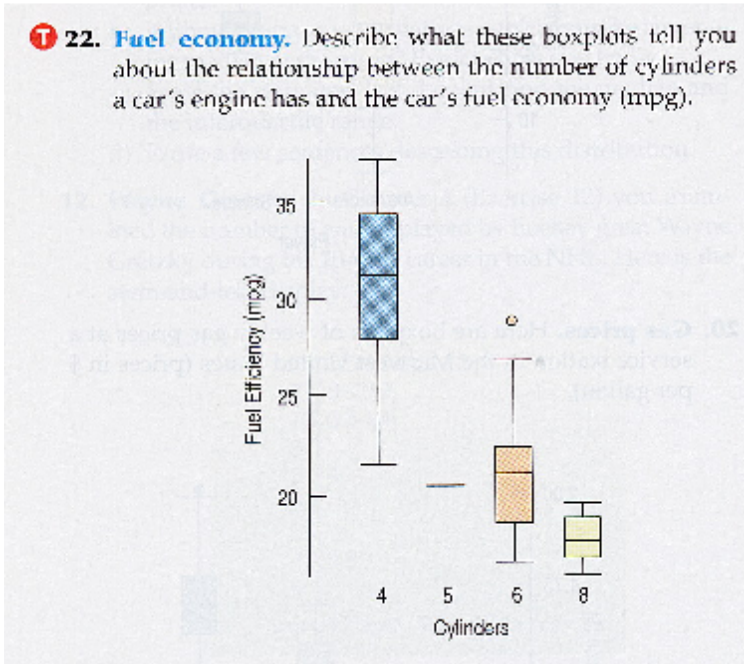
Examples, p. 113:

20. Gas prices. Here are boxplots of weekly gas prices at a service station in the Midwest United States (prices in \$ per gallon).



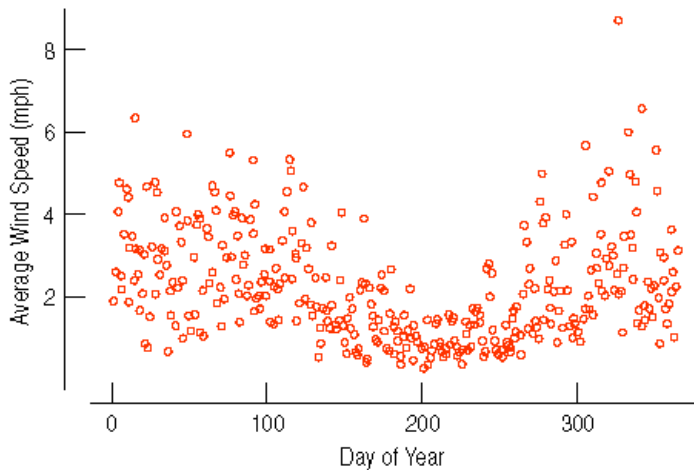
- Compare the distribution of prices over the three years.
- In which year were the prices least stable? Explain.

22. Fuel economy. Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg).



Timeplots (p. 98) are graphs that plot data values versus time. Used in Stock Market values, Stock prices, Unemployment Rates, Temperature, Global Temperatures, etc.

Figure 5.8, p. 98



One should resist using timeplots to predict for the future unless there is strong reasons for doing otherwise. (Path a ball follows when thrown from a certain height at a given speed, Seasonal Temperatures, Interest Rates?)