STA 2023

Module 2B

Organizing Data and Comparing Distributions (Part II)

Learning Objectives

Upon completing this module, you should be able to

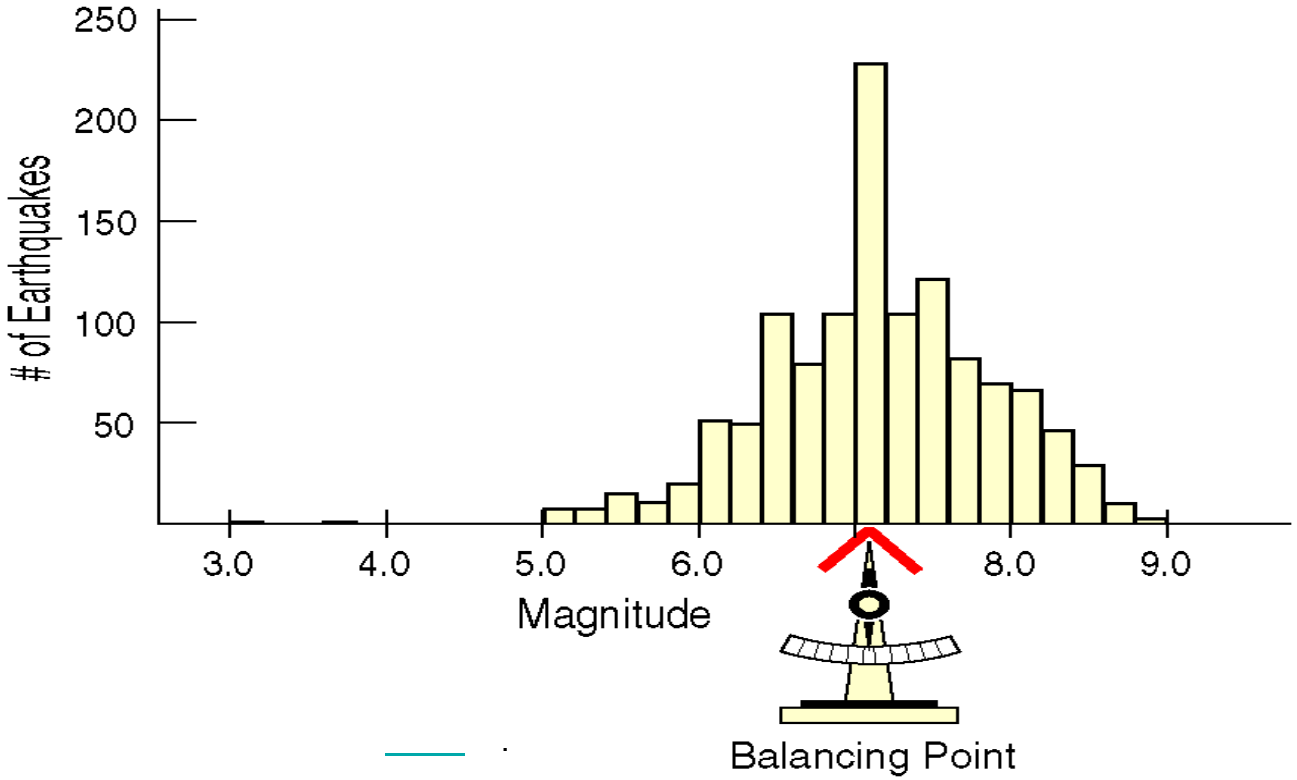
1. Explain the purpose of a measure of center.
2. Obtain and interpret the mean, median, and the mode(s) of a data set.
3. Choose an appropriate measure of center for a data set
4. Define, compute, and interpret a sample mean.
5. Explain the purpose of a measure of variation.
6. Define, compute, and interpret the range of a data set.

Learning Objectives (cont.)

7. Define, compute, and interpret a sample standard deviation.
8. Obtain and interpret the quartiles, IQR, and five-number summary of a data set.
9. Obtain the lower and upper limits of a data set and identify potential outliers
10. Construct and interpret a boxplot.
11. Use boxplots to compare two or more data sets.
12. Use a boxplot to identify distribution shape for large data sets.

What is the Mean of a Distribution?

The mean is like the center of the distribution because it is the point where the histogram balances



What is Mean of a Distribution? (cont.)

When a distribution is unimodal and symmetric, most people will point to the center of a distribution.

The center of a distribution is called mean. If we want to *calculate* a number, we can average the data. We use the Greek letter sigma to mean "sum" and write:

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

The formula says that to find the mean, we add up the numbers and divide by *n*.

What is the Mean of a Data Set?

Mean of a Data Set

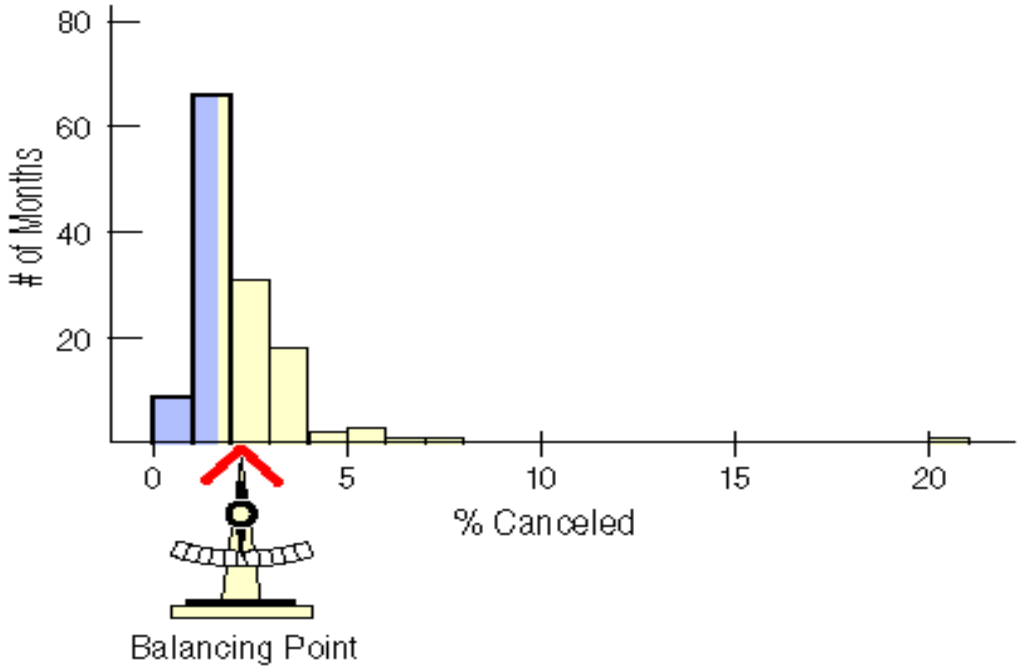
The **mean** of a data set is the sum of the observations divided by the number of observations.

$$\overline{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

The formula says that to find the mean, we add up the numbers and divide by *n*.

What is the Median?

- The median is the value with exactly half the data values below it and half above it.
  - It is the middle data value (once the data values have been ordered) that divides the histogram into two equal areas.
  - It has the same units as the data.



Mean or Median?

- In symmetric distributions, the mean and median are approximately the same in value, so either measure of center may be used.
- For skewed data, though, it's better to report the median than the mean as a measure of center.

How to Compute the Median?

Median of a Data Set

Arrange the data in increasing order.

- If the number of observations is odd, then the **median** is the observation exactly in the middle of the ordered list.

- If the number of observations is even, then the **median** is the mean of the two middle observations in the ordered list.

In both cases, if we let *n* denote the number of observations, then the median is at position (*n* + 1) / 2 in the ordered list.

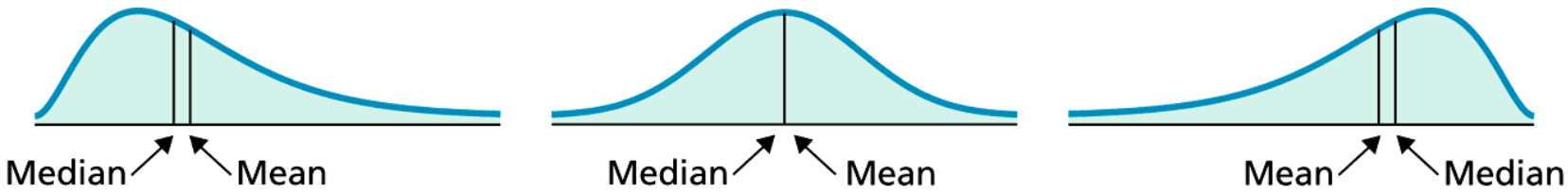
How to Find the Mode of a Data Set?

Mode of a Data Set

Find the frequency of each value in the data set.

- If no value occurs more than once, then the data set has *no mode.*
- Otherwise, any value that occurs with the greatest frequency is a **mode** of the data set.

Relative Positions of the Mean and Median



(a) Right skewed (b) Symmetric (c) Left skewed

Note that the mean is pulled in the direction of skewness, that is, in the direction of the extreme observations.

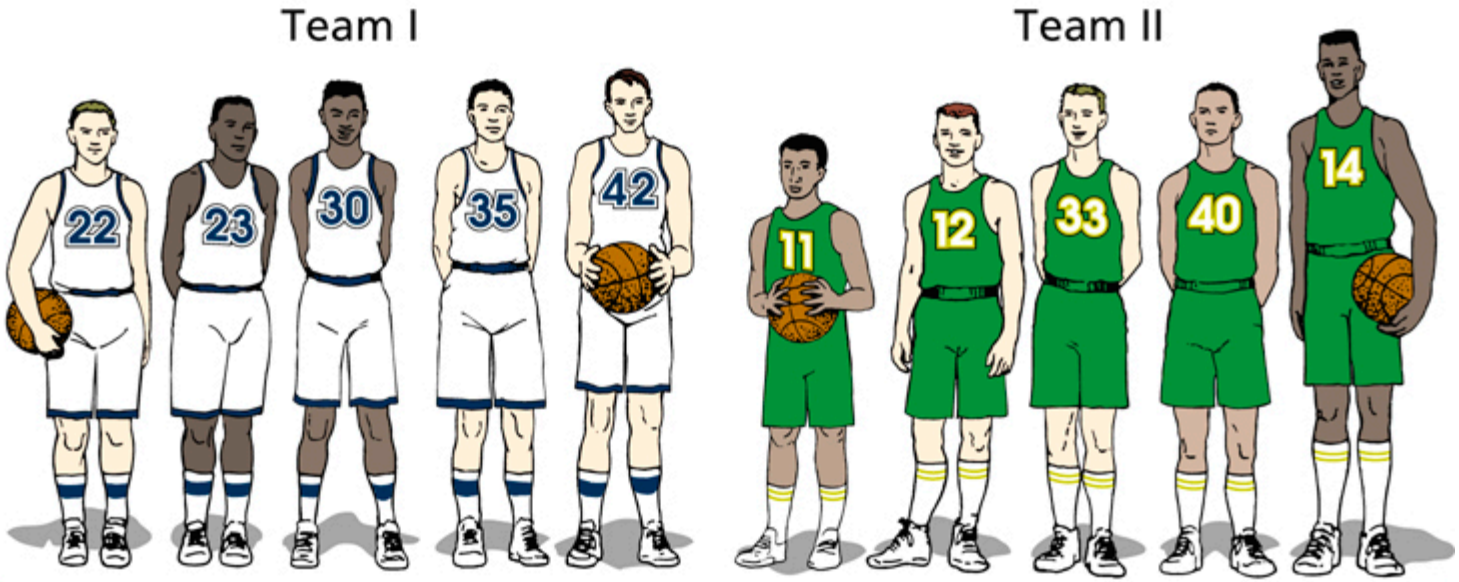
Measure of Center

$300	300	300	940	300
300	400	300	400	
450	800	450	1050	

$300	300	940	450	400
400	300	300	1050	300

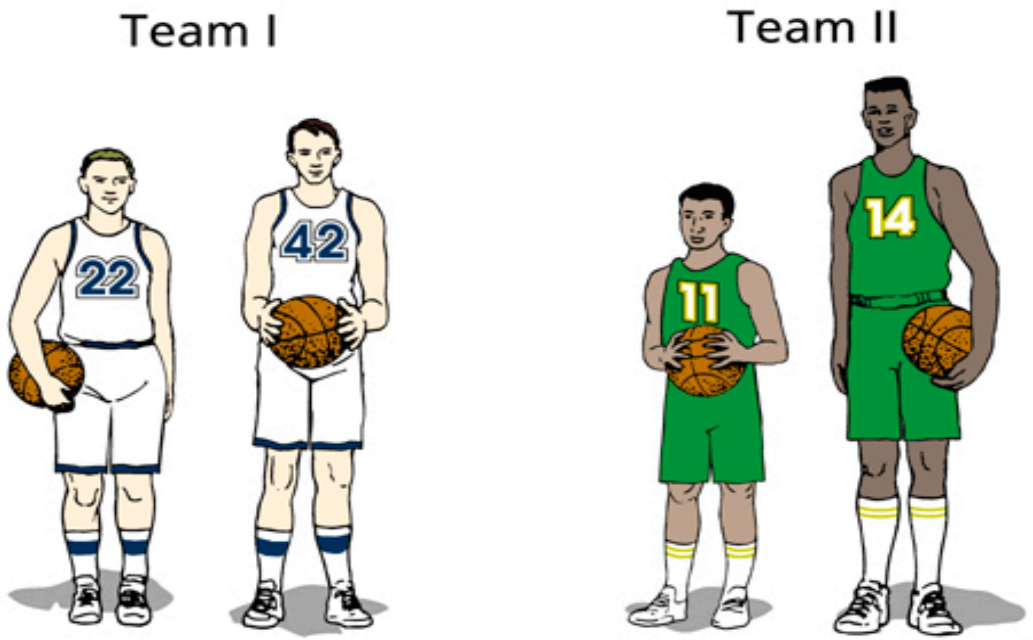
Measure of center	Definition	Data Set I	Data Set II
Mean	$\frac{\text{Sum of observations}}{\text{Number of observations}}$	$483.85	$474.00
Median	Middle value in ordered list	$400.00	$350.00
Mode	Most frequent value	$300.00	$300.00

Two Teams



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Shortest and Tallest (Min and Max)



	Team I		Team II	
Feet and inches	6'	6'6"	5'7 "	7 '
Inches	72	78	67	84

What is the Range of a Data Set?

Range of a Data Set

The **range** of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min},$$

where Max and Min denote the maximum and minimum observations, respectively.

The range of a data set is the difference between its largest and smallest values.

What is the Disadvantage of the Range?

- Always report a measure of spread along with a measure of center when describing a distribution numerically.
- The range of the data is the difference between the maximum and minimum values:

  Range = max – min

- A disadvantage of the range is that a single extreme value can make it very large and, thus, not representative of the data overall.

What is the Interquartile Range?

- The interquartile range (IQR) lets us ignore extreme data values and concentrate on the middle of the data.
- To find the IQR, we first need to know what quartiles are…

What are Quartiles?

Arrange the data in increasing order and determine the median.

- The **first quartile** is the median of the part of the entire data set that lies at or below the median of the entire data set.
- The **second quartile** is the median of the entire data set.
- The **third quartile** is the median of the part of the entire data set that lies at or above the median of the entire data set.

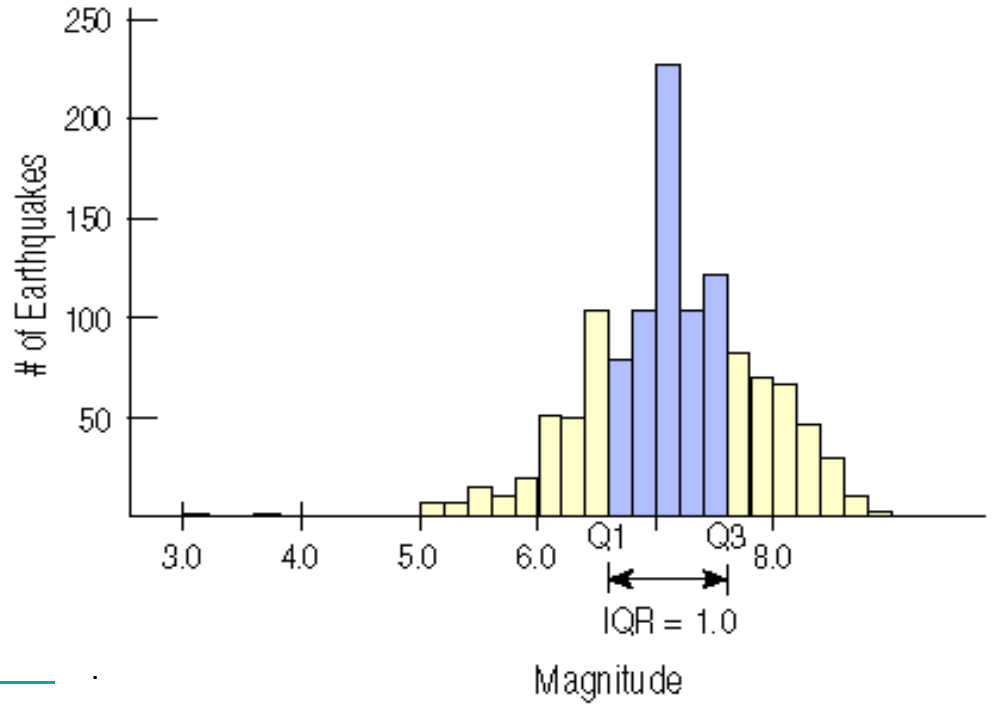
What are Quartiles? (Cont.)

- Quartiles divide the data into four equal sections.
  - One quarter of the data lies below the lower quartile, Q1
  - One quarter of the data lies above the upper quartile, Q3.
- The difference between the quartiles is the IQR, so

$$IQR = Q3 - Q1$$

The Interquartile Range (cont.)

- The lower and upper quartiles are the 25th and 75th percentiles of the data, so…
- The IQR contains the middle 50% of the values of the distribution, as shown in the following figure:



The Interquartile Range (Cont.)

Interquartile Range

The **interquartile range,** or **IQR,** is the difference between the first and third quartiles; that is, IQR = $Q_3 - Q_1$.

What does it Mean?

Roughly speaking, the IQR gives the range of the middle 50% of the observations.

Standard Deviation

- A more powerful measure of spread than the IQR is the standard deviation, which takes into account how far *each* data value is from the mean.
- A deviation is the distance that a data value is from the mean.
  - Since adding all deviations together would total zero, we square each deviation and find an average of sorts for the deviations.

What is Variance?

- The variance, notated by s^2, is found by summing the squared deviations and (almost) averaging them:

$$s^2 = \frac{\sum \left(y - \bar{y} \right)^2}{n - 1}$$

- The variance will play a role later in our study, but it is problematic as a measure of spread — it is measured in *squared* units!

Variance and Standard Deviation

The standard deviation, *s*, is just the square root of the variance and is measured in the same units as the original data.

$$s = \sqrt{\frac{\sum \left(y - \bar{y}\right)^2}{n-1}}$$

Thinking About Variation

- Since Statistics is about variation, spread is an important fundamental concept of Statistics.
- Measures of spread help us talk about what we *don't* know.
- When the data values are tightly clustered around the center of the distribution, the IQR and standard deviation will be small.
- When the data values are scattered far from the center, the IQR and standard deviation will be large.

Quantitative Variables

When telling about quantitative variables, start by making a histogram or stem-and-leaf display and discuss the shape of the distribution.

Shape, Center, and Spread

Next, always report the *shape* of its distribution, along with a *center* and a *spread*.

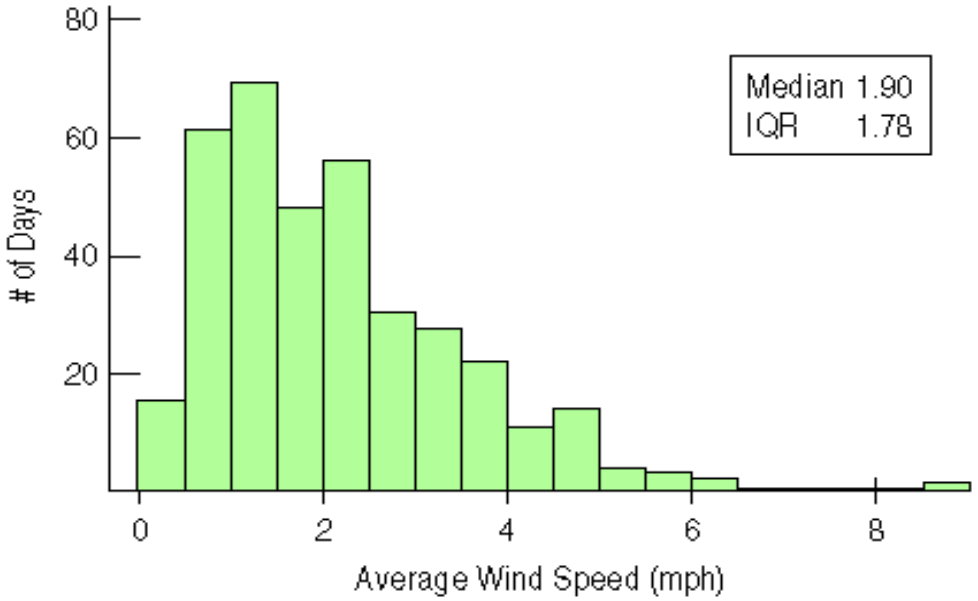
- If the shape is *skewed*, report the *median* and *IQR*.
- If the shape is *symmetric*, report the *mean* and *standard deviation* and possibly the median and IQR as well.

What About Unusual Features?

- If there are multiple modes, try to understand why. If you identify a reason for the separate modes, it may be good to split the data into two groups.
- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.

The Big Picture

- We can answer much more interesting questions about variables when we compare distributions for different groups.
- Below is a histogram of the *Average Wind Speed* for every day in 1989.



The Big Picture (cont.)

- The distribution is unimodal and skewed to the right.
- The high value may be an outlier
- Median daily wind speed is about 1.90 mph and the IQR is reported to be 1.78 mph.
- Can we say more?

What is the Five-Number Summary?

Five-Number Summary

The **five-number summary** of a data set is Min, Q_1, Q_2, Q_3, Max.

What does it mean?

The five-number summary of a data set consists of the minimum, maximum, median, first quartile and third quartile, written in ascending order.

What are the Lower Limit and Upper Limit?

Lower and Upper Limits

The **lower limit** and **upper limit** of a data set are

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$
$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

What do they mean?

The lower limit is the number that lies 1.5 IQRs below the first quartile; the upper limit is the number that lies 1.5 IQRs above the third quartile.

Example:
The Five-Number Summary

The five-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum).

- Example: The five-number summary for the daily wind speed is:

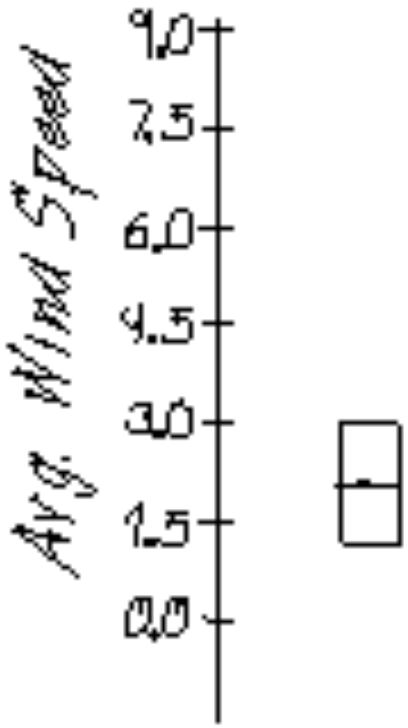
Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

Daily Wind Speed: Making Boxplots

- A boxplot is a graphical display of the five-number summary.
- Boxplots are particularly useful when comparing groups.

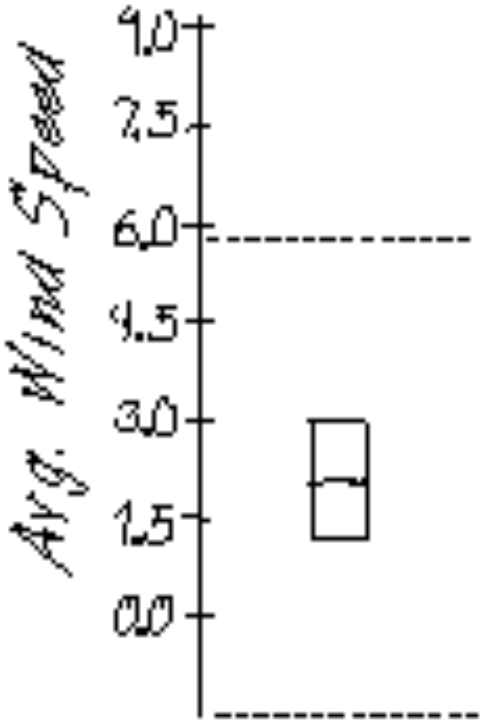
How to Construct a Boxplot?

1. Draw a single vertical axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.



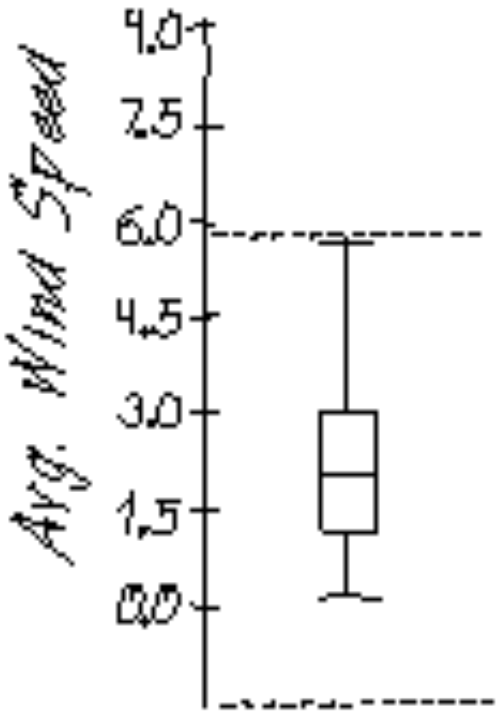
How to Construct a Boxplot?(cont.)

2. Erect "fences" around the main part of the data.
   - The upper fence is 1.5 IQRs above the upper quartile.
   - The lower fence is 1.5 IQRs below the lower quartile.
   - Note: the fences only help with constructing the boxplot and should not appear in the final display.



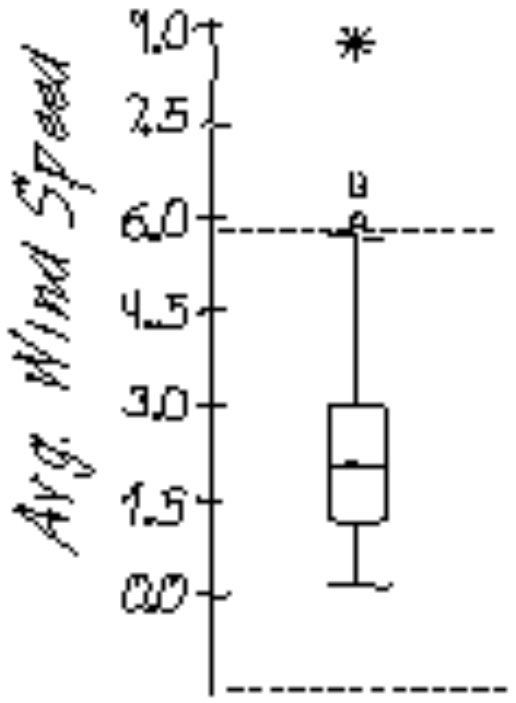
How to Construct a Boxplot? (cont.)

3. Use the fences to grow "whiskers."
   - Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*.
   - If a data value falls outside one of the fences, we do *not* connect it with a whisker.



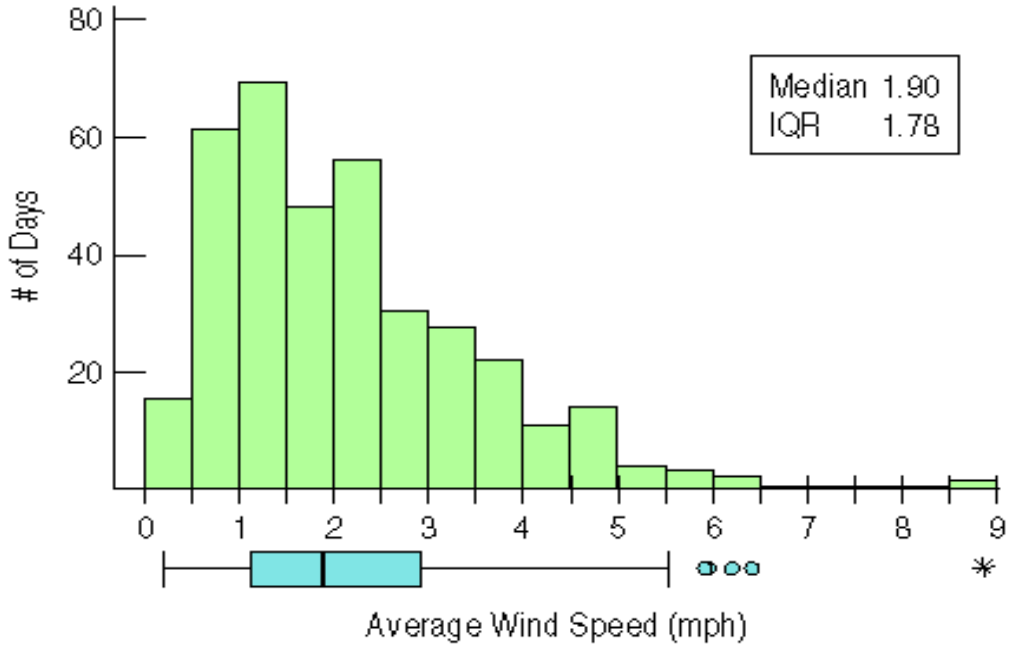
How to Construct a Boxplot (cont.)

4. Add the outliers by displaying any data values beyond the fences with special symbols.
   - We often use a different symbol for "far outliers" that are farther than 3 IQRs from the quartiles.



Histogram and Boxplot

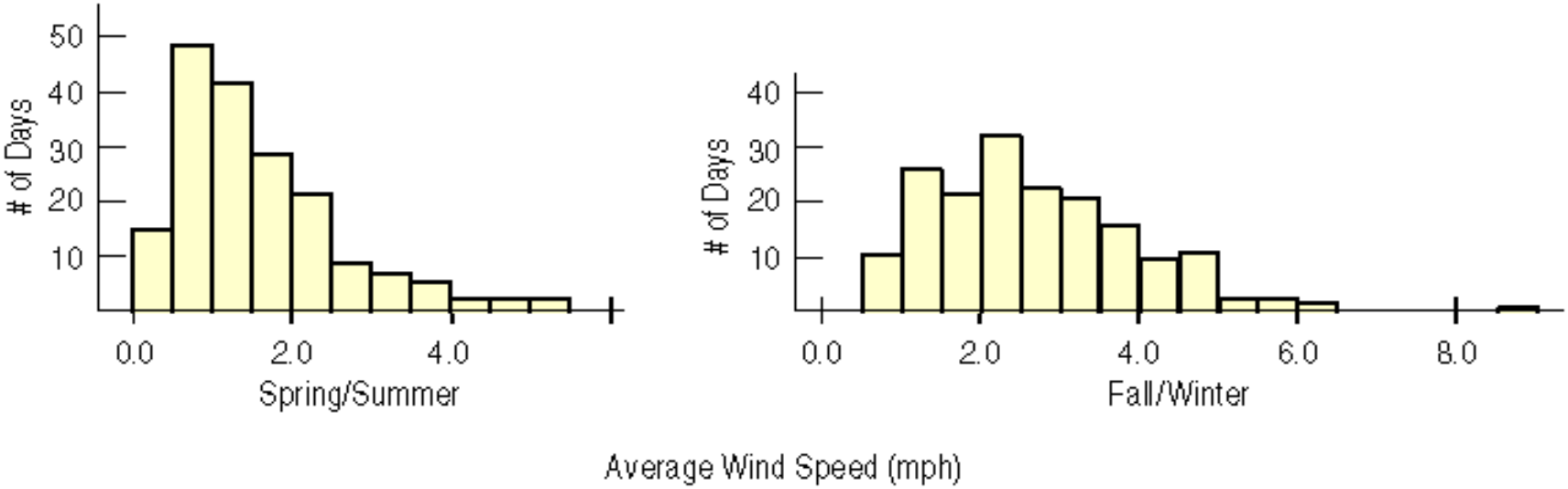
Compare the histogram and boxplot for daily wind speeds:



How does each display represent the distribution?

Comparing Groups

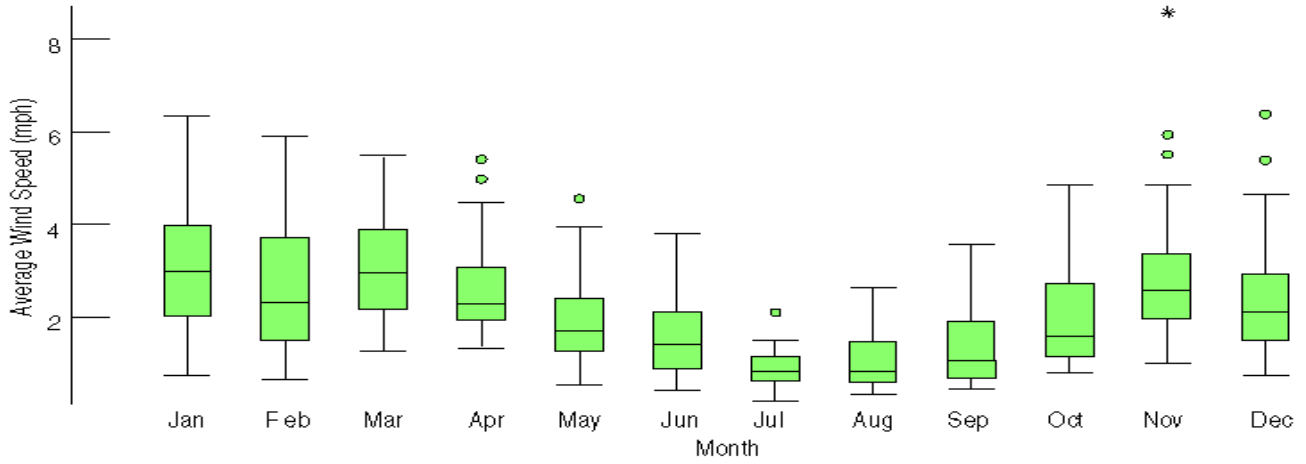
- It is always more interesting to compare groups.
- With histograms, note the shapes, centers, and spreads of the two distributions.



- What does this graphical display tell you?

Comparing Groups (cont.)

- Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information.
- We often plot them side by side for groups or categories we wish to compare.



- What do these boxplots tell you?

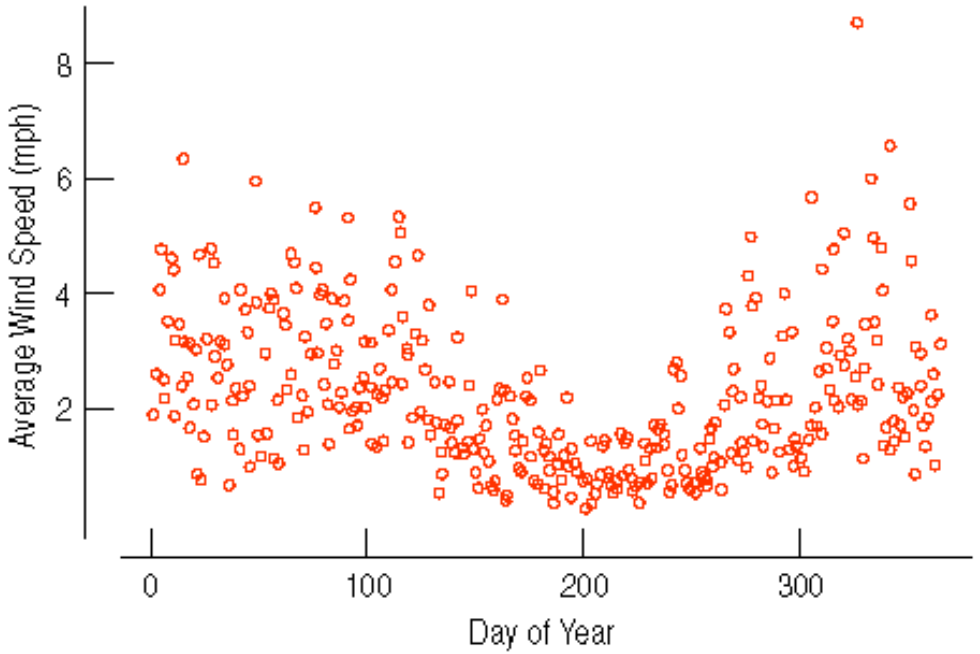
What About Outliers?

If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.

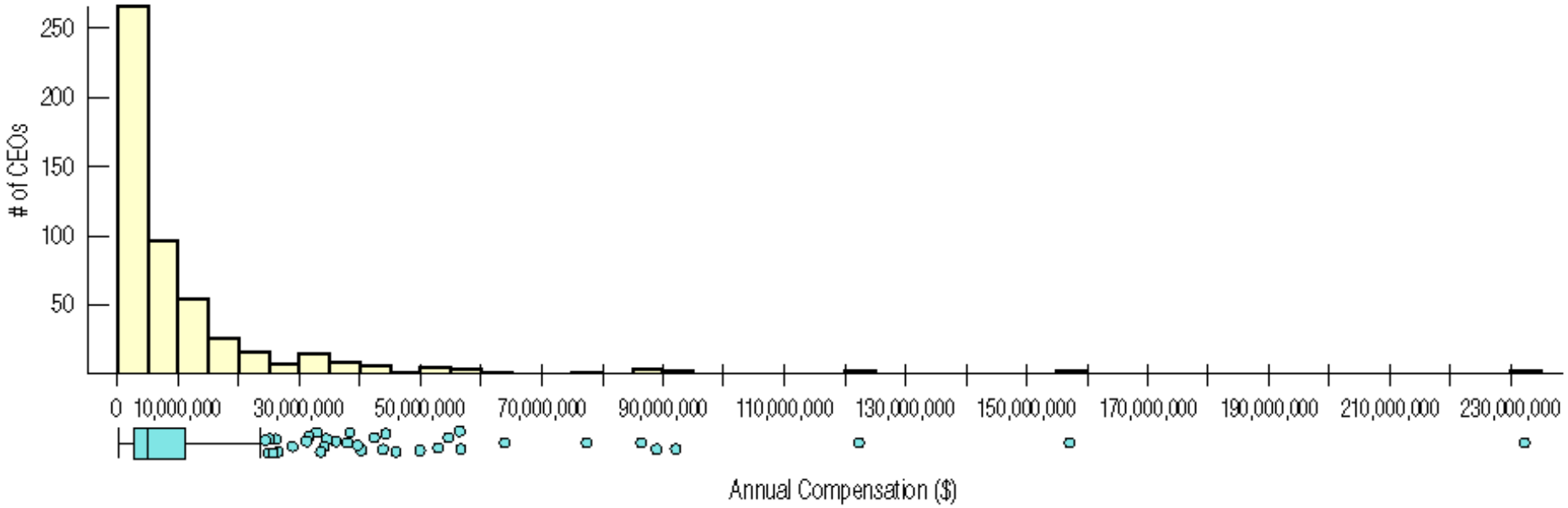
Note: The median and IQR are not likely to be affected by the outliers.

Timeplots

For some data sets, we are interested in how the data behave over time. In these cases, we construct timeplots of the data.



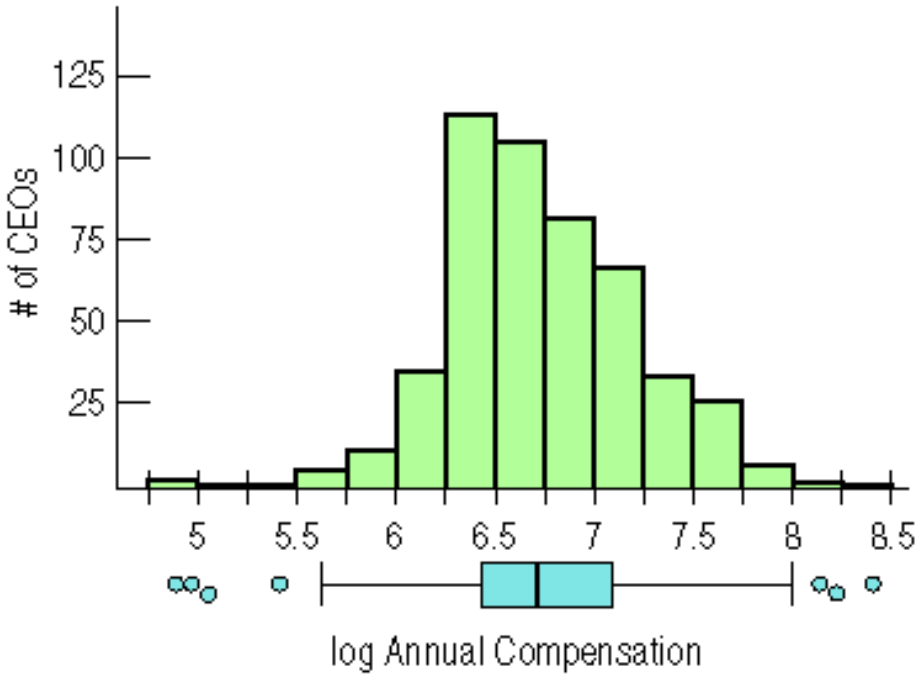
Re-expressing Skewed Data to Improve Symmetry



Re-expressing Skewed Data to Improve Symmetry (cont.)

One way to make a skewed distribution more symmetric is to re-express or transform the data by applying a simple function (e.g., logarithmic function).

Note the change in skewness from the raw data (previous slide) to the transformed data (right):



What have we learned?

We have learned to:

1. Explain the purpose of a measure of center.
2. Obtain and interpret the mean, median, and the mode(s) of a data set.
3. Choose an appropriate measure of center for a data set
4. Define, compute, and interpret a sample mean.
5. Explain the purpose of a measure of variation.
6. Define, compute, and interpret the range of a data set.

What have we learned? (cont.)

7. Define, compute, and interpret a sample standard deviation.
8. Obtain and interpret the quartiles, IQR, and five-number summary of a data set.
9. Obtain the lower and upper limits of a data set and identify potential outliers
10. Construct and interpret a boxplot.
11. Use boxplots to compare two or more data sets.
12. Use a boxplot to identify distribution shape for large data sets.

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition