Statistics: Numerical Methods for Describing Data

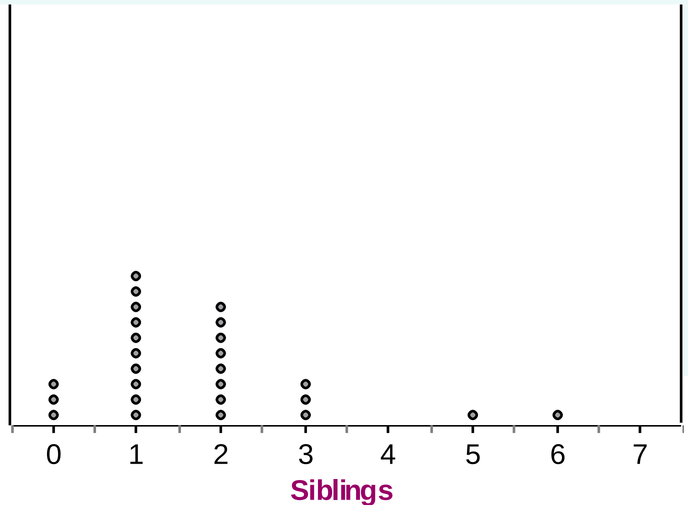
Describing Quantitative Data with Numbers

Data Distributions- Numerical Methods for Exploring Data

- **4.1** Describing the Center of a Data Set
- **4.2** Describing Variability of a Data Set
- **4.3** Summarizing a Data Set: Boxplots

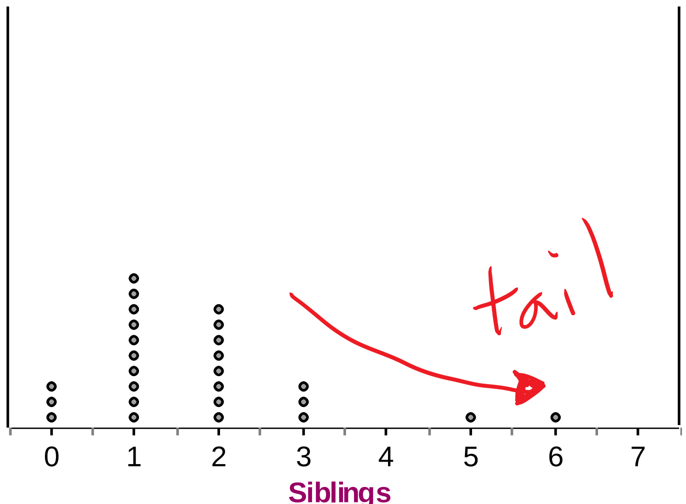
Sept 7, 2022 Warm-Up

1. If percents are referenced by ***percentiles***, then quarters must be referenced by ____________
2. What is an outlier?
3. How would you label the shape of this data?



Warm-Up

1. If percents are referenced for *percentiles*, then quarters must be referenced by ___*quartiles*___
2. What is an outlier? *Any data that is **unusually** large or **unusually** small compared to the data*
3. How would you label the shape of this data?



Skewed right or positively skewed

Dotplots

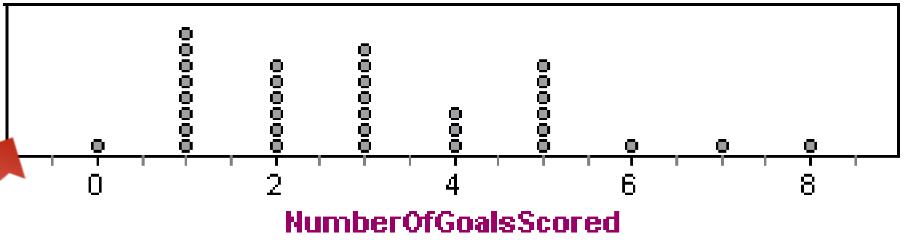
- One of the simplest graphs to construct and interpret is a **dotplot**. Each data value is shown as a dot above its location on a number line.

How to Make a Dotplot

1) Draw a horizontal axis (a number line) and label it with the variable name.
2) Scale the axis from the minimum to the maximum value.
3) Mark a dot above the location on the horizontal axis corresponding to each data value.

Number of Goals Scored Per Game by the 2004 US Women's Soccer Team

3	0	2	7	8	2	4	3	5	1	1	4	5	3	1	1	3
3	3	2	1	2	2	2	4	3	5	6	1	5	5	1	1	5



- **Examining the Distribution of a Quantitative Variable**
- The purpose of a graph is to help us understand the data. After you make a graph, always ask, "What do I see?"

How to Examine the Distribution of a Quantitative Variable

In any graph, look for the **overall pattern** and for striking **departures** from that pattern.

Describe the overall pattern of a distribution by its:

- **Shape**
- **Center**
- **Spread**

Don't forget your SOCS!

Note individual values that fall outside the overall pattern. These departures are called **outliers**.

Describing Shape

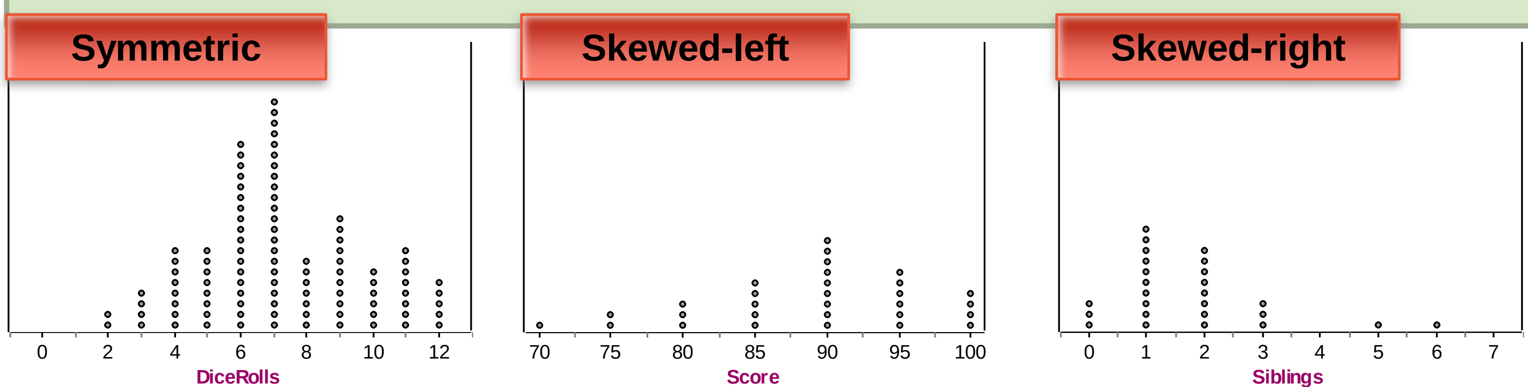
- When you describe a distribution's shape, concentrate on the main features. Look for rough **symmetry** or clear **skewness**.

Definitions:

A distribution is roughly **symmetric** if the right and left sides of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** (right-skewed or *positively skewed*) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

It is **skewed to the left** (left-skewed or *negatively skewed*) if the left side of the graph is much longer than the right side.



Distribution Shape

U.S. Income Distribution from 2005

Skewed Right or *positively skewed* →

Characteristics of Numerical Data
Describing Quantitative Data with Numbers

Learning Objectives

After this section, you should be able to…

- ✓ MEASURE center with the mean and median
- ✓ MEASURE spread with standard deviation and interquartile range
- ✓ IDENTIFY outliers
- ✓ CONSTRUCT a boxplot using the **five-number summary**
- ✓ CALCULATE numerical summaries with technology

Measures of Center and spread

- What are common measures of center for a numerical distribution of data?

mean & median

- What common measures of spread for a numerical distribution of data?

range, interquartile range (IQR),
& *standard deviation*

Measuring Center: The Mean

- The most common measure of center is the ordinary arithmetic average, or **mean**.

Definition:

To find the **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the *n* observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

In mathematics, the capital Greek letter Σis short for "add them all up." Therefore, the formula for the mean can be written in more compact notation:

$$\bar{x} = \frac{\sum x_i}{n}$$

Measuring Center: The Median

- Another common measure of center is the **median**. In section 1.2, we learned that the median describes the midpoint of a distribution.

Definition:

The **median M** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1) Arrange all observations from smallest to largest.

2) If the number of observations n is odd, the median M is the center observation in the ordered list.

3) If the number of observations ***n* is even**, the median M is the average of the two center observations in the ordered list.

Measuring Center

- Use the data below to calculate the mean and median of the commuting times (in minutes) of 20 randomly selected New York workers.

Example, page ??

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

$$\bar{x} = \frac{10 + 30 + 5 + 25 + \ldots + 40 + 45}{20} = 31.25 \text{ minutes}$$

0	5
1	005555
2	00**05**
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 represents a New York worker who reported a 45-minute travel time to work.

$$M = \frac{20 + 25}{2} = 22.5 \text{ minutes}$$

Comparing the Mean and the Median

- The mean and median measure center in different ways, and both are useful.
  - *Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.*

Comparing the Mean and the Median

The mean and median of a roughly symmetric distribution are close together.

If the distribution is exactly symmetric, the mean and median are exactly the same.

In a skewed distribution, the mean is usually farther out in the long tail than is the median.

Measures of spread

Range: the spread of all the data, calculated as the difference between the largest and smallest observations in the data.

Standard deviation: average or "typical" deviation from the mean for a set of data. Calculated by finding the average of the squared deviations from the mean.

Interquartile range (IQR) : the spread of the middle 50% of the data, calculated by difference in $Q_3 - Q_1 = IQR$

Measuring Spread: The Interquartile Range (*IQR*)

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles**:

1) Arrange the observations in increasing order and locate the median *M*.

2) The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.

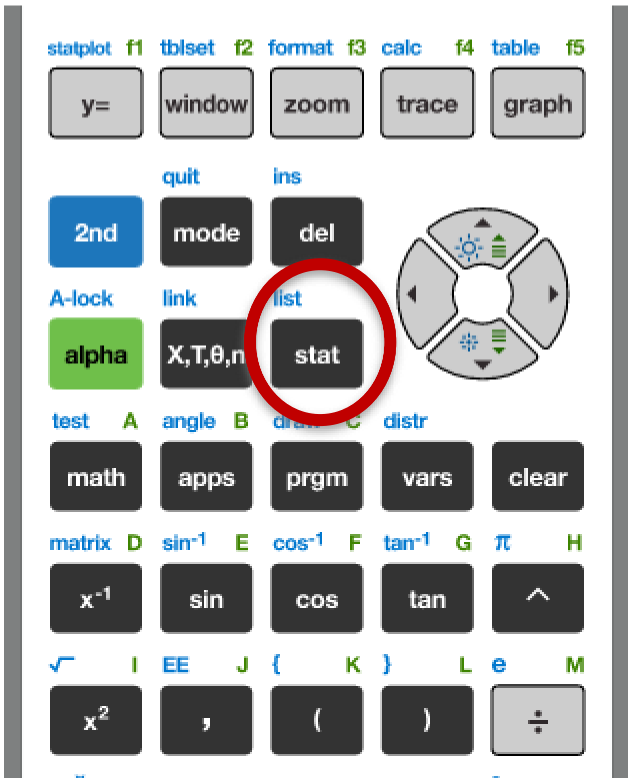
3) The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.

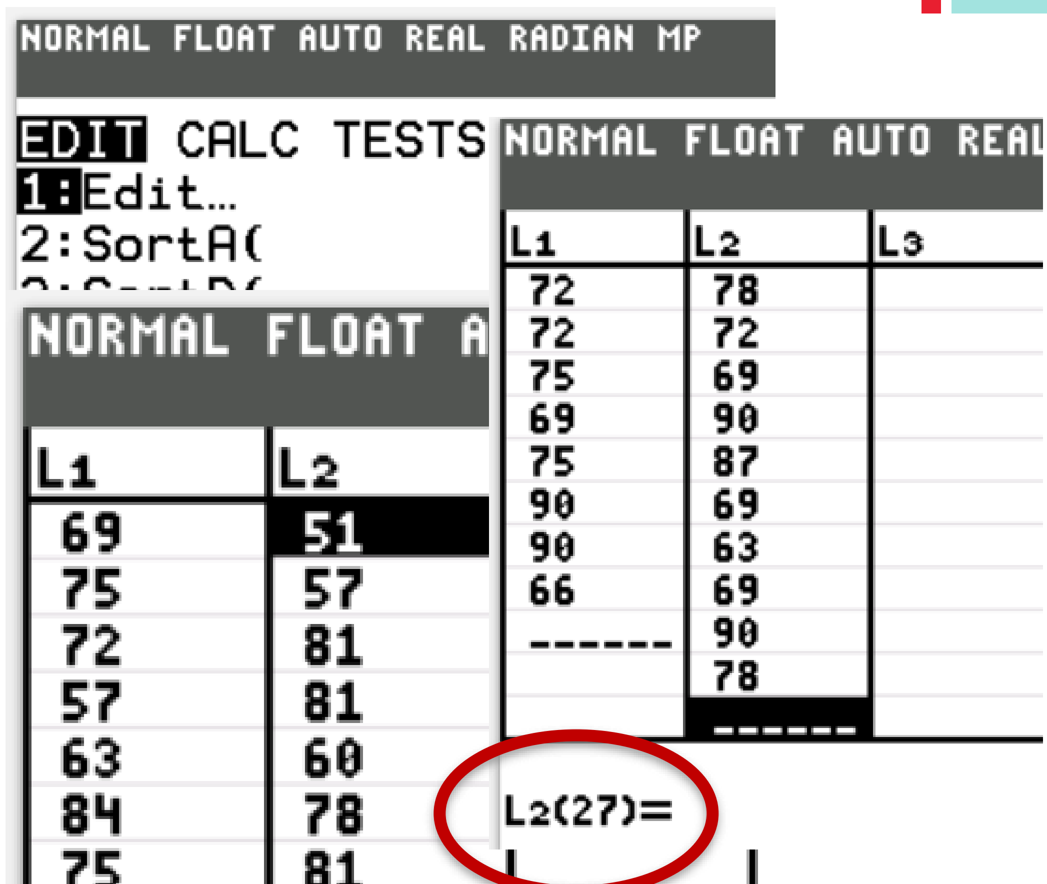
The **interquartile range (*IQR*)** is defined as:

$$IQR = Q_3 - Q_1$$

Entering Data in calculator (using TI-84)

- Choose the stat button, then enter





Travel times to work for 20 randomly selected New Yorkers

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

Find and Interpret the IQR

Example

Travel times to work for 20 randomly selected New Yorkers

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$

$M = 22.5$

$Q_3 = 42.5$

$$
\begin{aligned}
IQR &= Q_3 - Q_1 \\
&= 42.5 - 15 \\
&= 27.5 \text{ minutes}
\end{aligned}
$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

Quantitatviely Identifying Outliers

- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

Definition:

The 1.5 x IQR Rule for Outliers

Call an observation an outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.

Example

In the New York travel time data, we found Q_1=15 minutes, Q_3 = 42.5 minutes, and *IQR* = 27.5 minutes.

For these data, 1.5 x *IQR* = 1.5(27.5) = 41.25

Q_1 - 1.5 x *IQR* = 15 – 41.25 = **-26.25**

Q_3+ 1.5 x *IQR* = 42.5 + 41.25 = **83.75**

Any travel time shorter than -26.25 minutes or longer than 83.75 minutes is considered an outlier.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	**5**

The Five-Number Summary

- The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.
- To get a quick summary of both center and spread, combine all five numbers.

Definition:

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

Minimum $\quad Q_1 \quad M \quad Q_3 \quad$ ***Maximum***

Boxplots (Box-and-Whisker Plots)

- The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot**.

How to Make a Boxplot

•Draw and label a number line that includes the range of the distribution.

•Draw a central box from Q_1 to Q_3.

•Note the median *M* inside the box.

•Extend lines (whiskers) from the box out to the minimum and maximum values that are not outliers.

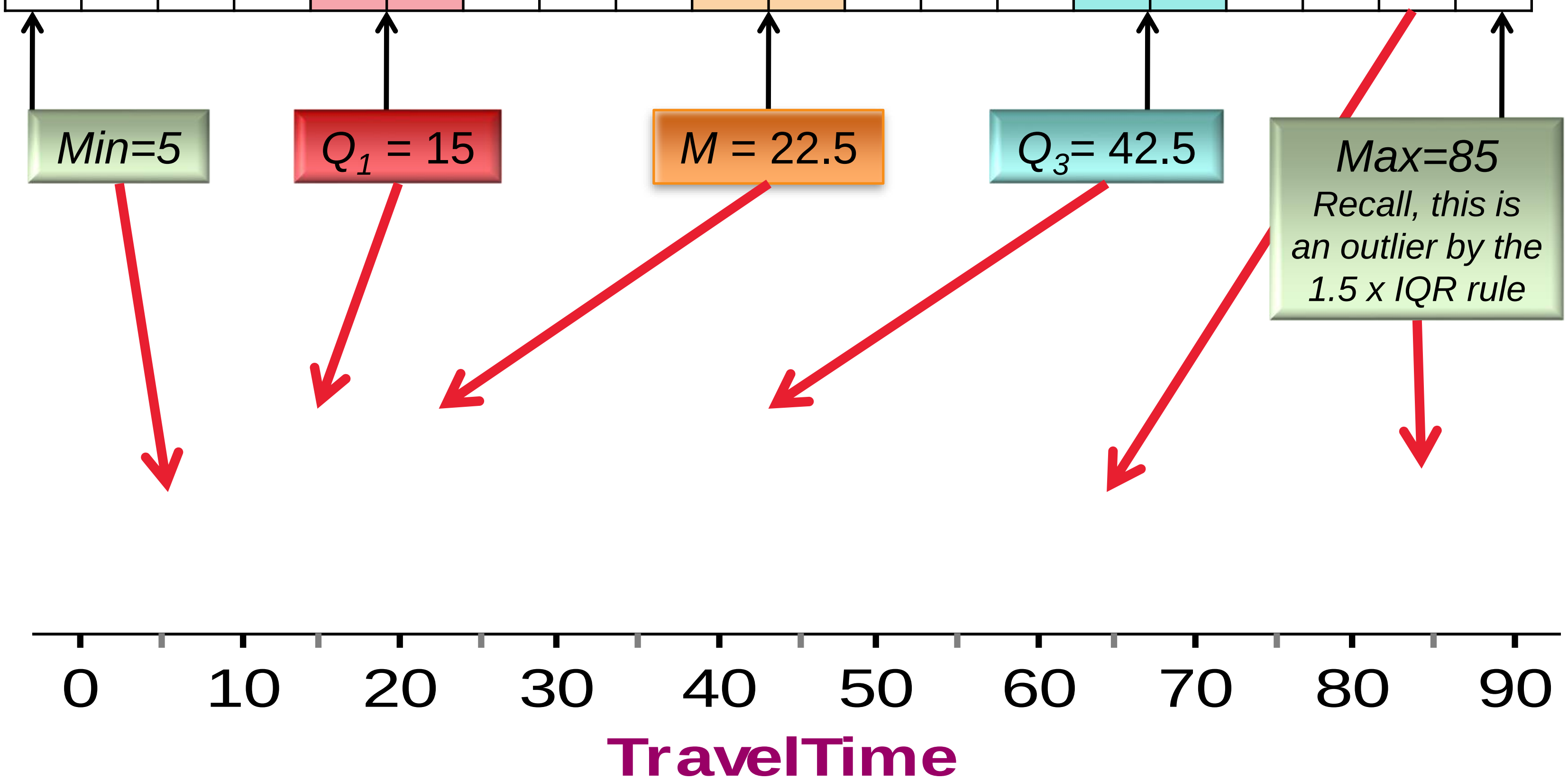
Construct a Boxplot

Example

- Consider our NY travel times data. Construct a boxplot.

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85



- **Boxplots (Box-and-Whisker Plots)**
- The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot**.

How to Make a Boxplot

•Draw and label a number line that includes the range of the distribution.

•Draw a central box from Q_1 to Q_3.

•Note the median *M* inside the box.

•Extend lines (whiskers) from the box out to the minimum and maximum values that are not outliers.

+ FUN Friday! Sept 2, 2022

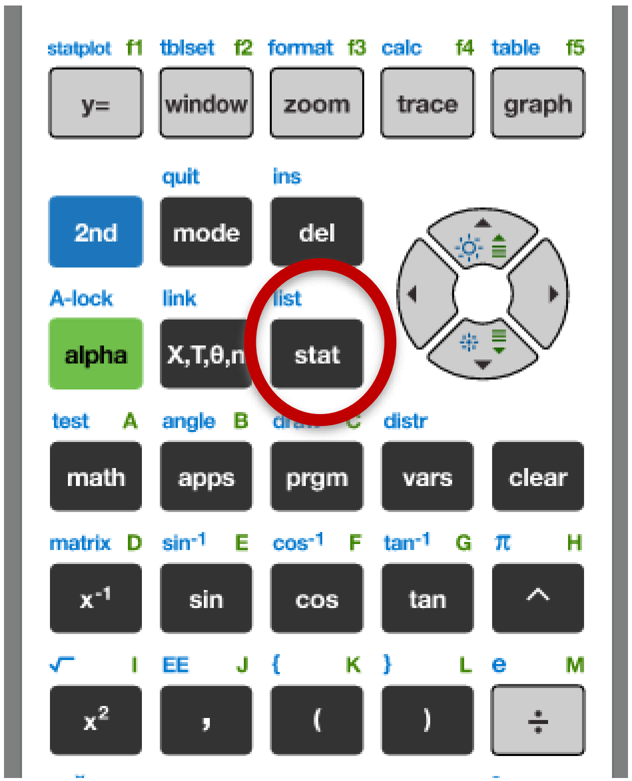
- **<u>Warm-Up</u>:** Letter to Future **ME**
- **Review measures of Center & Spread**
- **SU-DO-KU? Game of Skunk?**
- **HW Time, Video time**
- **TEST review DUE Next week!**
- **Questions?**

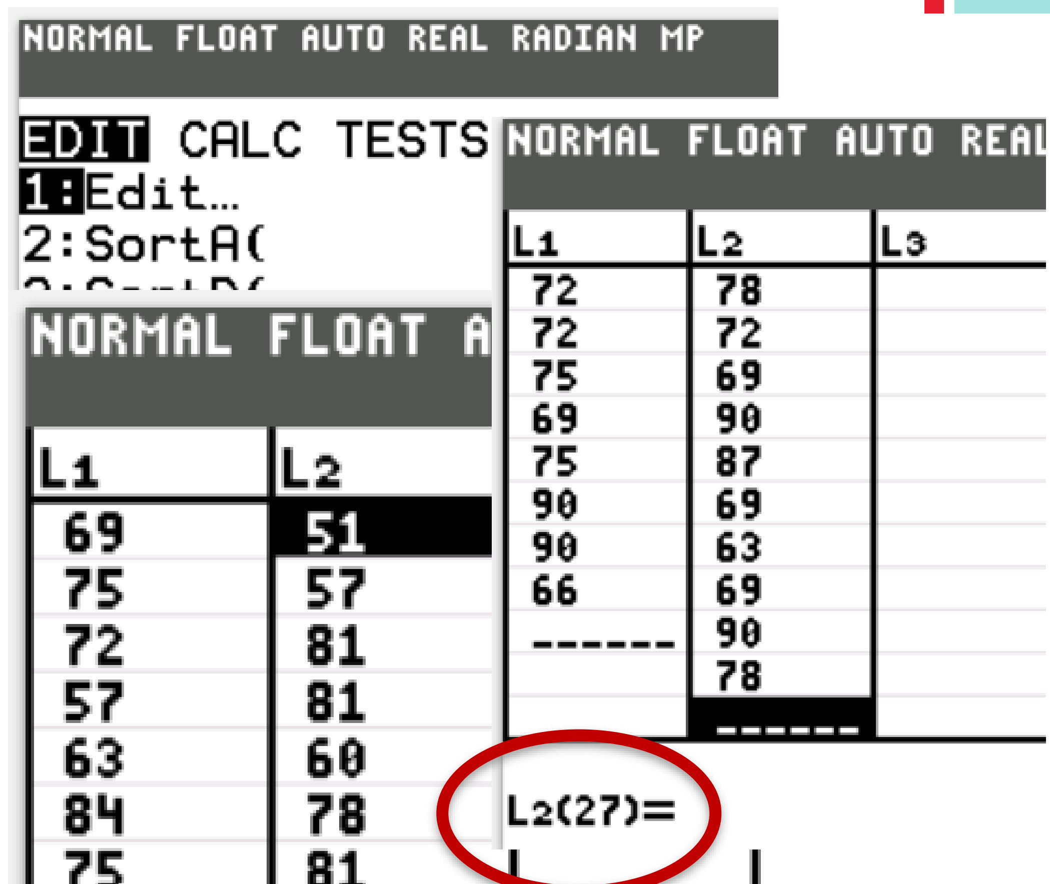
Box Plot Practice

- Use the data provided to make two lists
- Use the STAT PLOT menu to create box plots
- Be prepared to discuss the distributions

Entering Data in calculator (using TI-84)

- Choose the stat button, then enter





Sample Data Sets

Data Set A

Values 1-10	Values 11-20	Values 21-25
69	78	75
75	78	90
72	75	90
57	57	66
63	75	**84**
84	63	
75	72	
75	72	
66	75	
75	69	

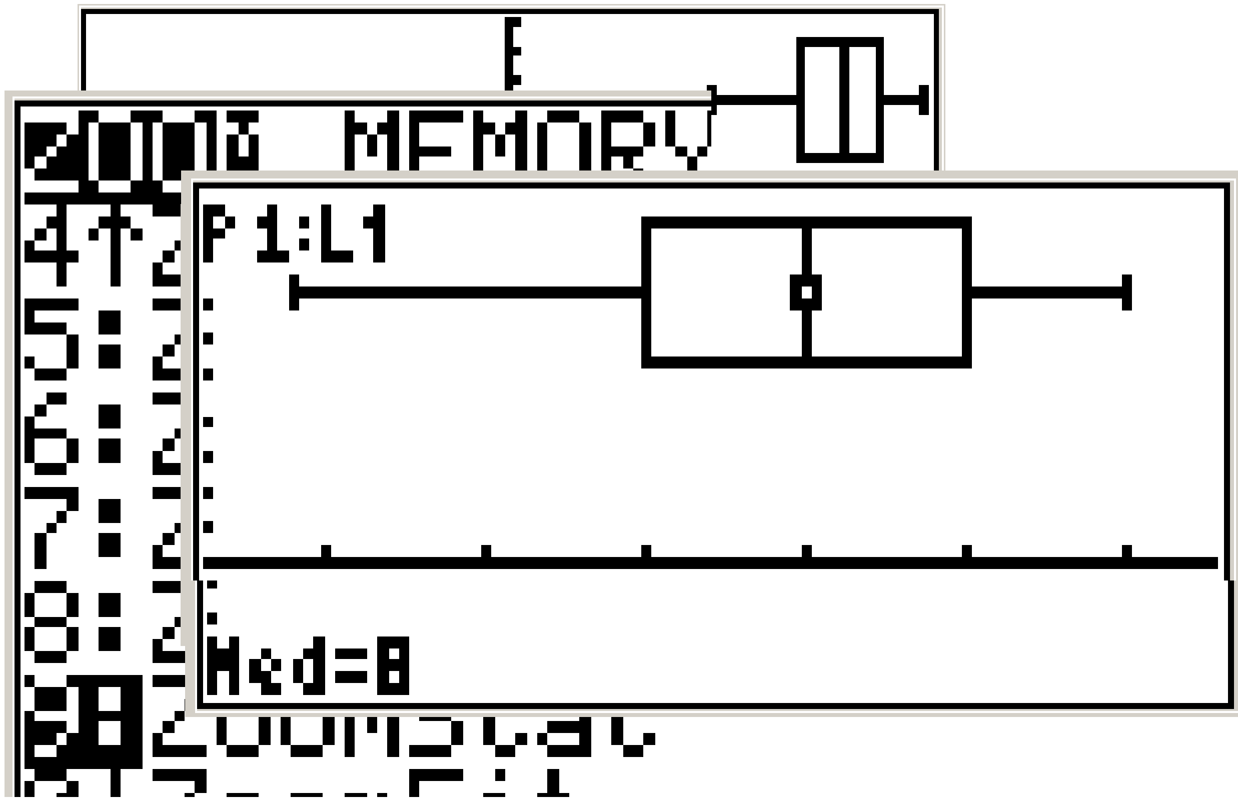
Data Set B

Values 1-10	Values 11-20	Values 21-26
51	81	87
57	57	69
81	72	63
81	75	69
60	87	90
78	75	78
81	78	
75	72	
84	69	
78	90	

Box Plot Practice

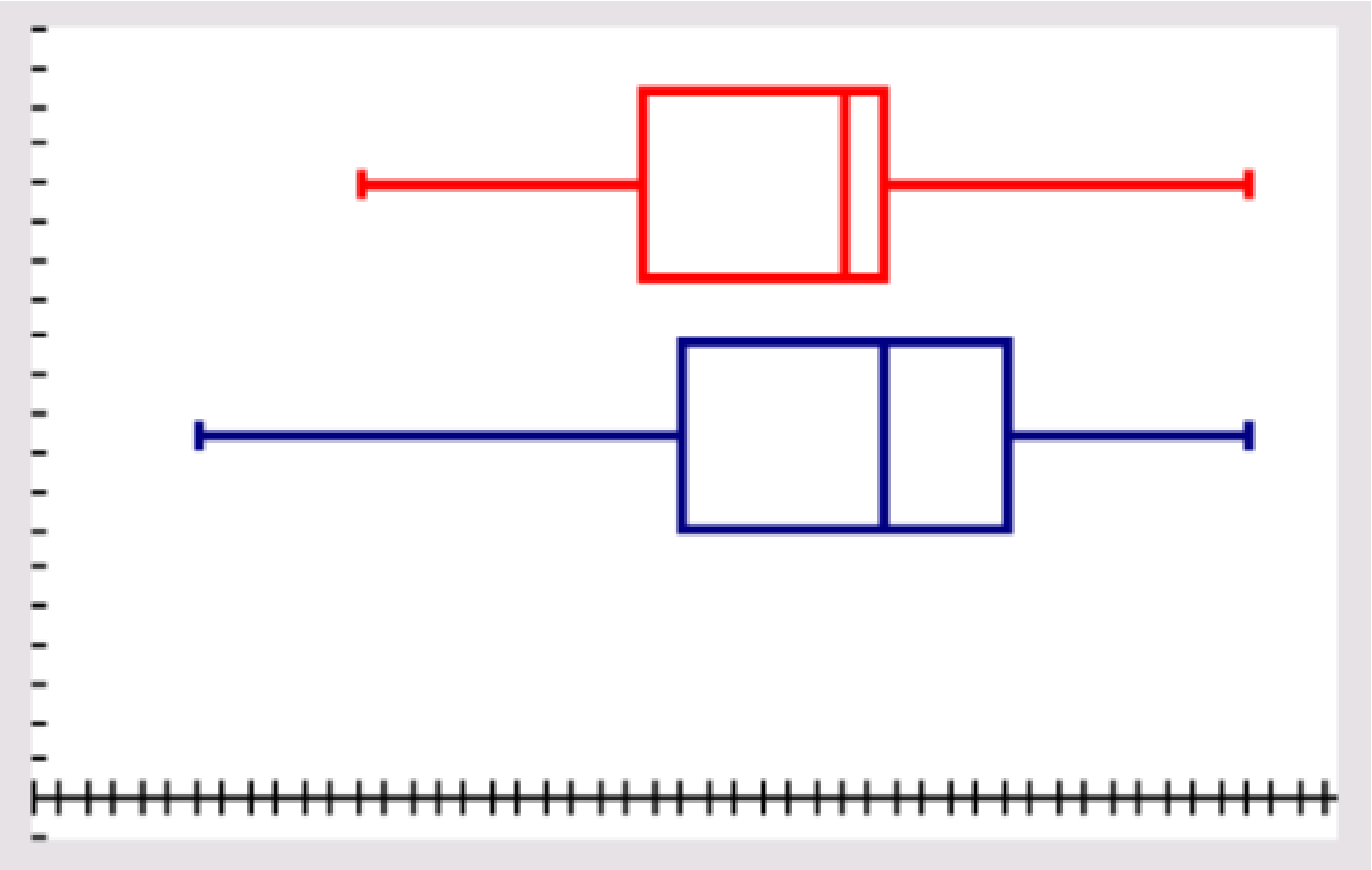
Data Set A

	Data A
Obser #1	8
Obser #2	5.6
Obser #3	4.8
Obser #4	9.2
Obser #5	6.8
Obser #6	9.6
Obser #7	8
Obser #8	8.8
Obser #9	7.6
	etc. …



Box Plot Practice Comparing Data

How do the data sets compare?



Looking Ahead…

<u>In the next part of Chapter 4…</u>

We'll learn how to model distributions of data…

- **Calculating the Standard deviation of a distribution**
- **Describing Location in a Distribution**
- **Introduction to Normal Distributions**