Chapter 1: Exploring Data

Section 1.2

Displaying Quantitative Data with Graphs

The Practice of Statistics, 4th edition - For AP*

STARNES, YATES, MOORE

Chapter 1
Exploring Data

- **Introduction**: Data Analysis: Making Sense of Data
- **1.1** Analyzing Categorical Data
- **1.2** **Displaying Quantitative Data with Graphs**
- **1.3** Describing Quantitative Data with Numbers

Section 1.2
Displaying Quantitative Data with Graphs

Learning Objectives

After this section, you should be able to…

- ✓ CONSTRUCT and INTERPRET dotplots, stemplots, and histograms
- ✓ DESCRIBE the shape of a distribution
- ✓ COMPARE distributions
- ✓ USE histograms wisely

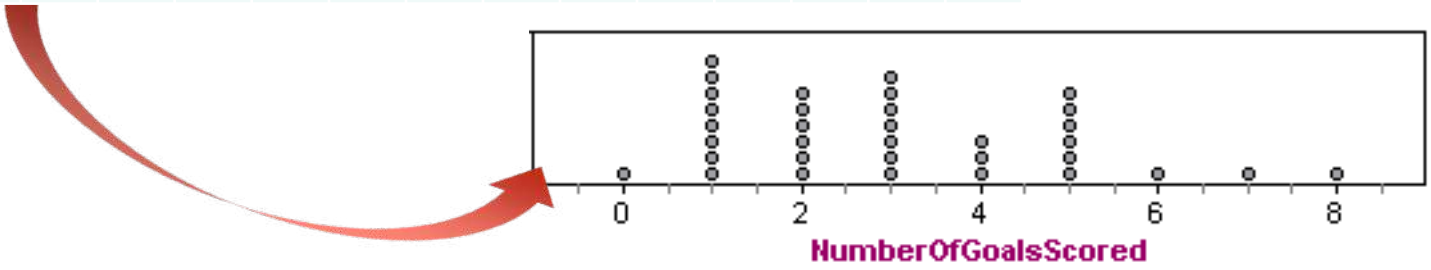
Dotplots One of the simplest graphs to construct and interpret is a **dotplot.** Each data value is shown as a dot above its location on a number line.

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.

Displaying Quantitative Data

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



amining the tribution of uantitative iable

purpose of raph is to us erstand the a. After you ke a graph, ays ask, at do I ?"

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!

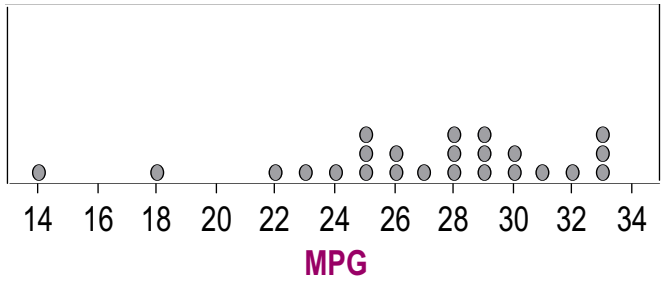
Note individual values that fall outside the overall pattern. These departures are called **outliers**.

Example, page 28

Exam… this d…

The tab… and dotplot …ow displays the Environmental Protection Agency's estimates of highway gas mileage in miles per gallon (MPG) for a sample of 24 model year 2009 midsize cars.

MODEL	MPG	MODEL	MPG	MODEL	MPG
Acura RL	22	Dodge Avenger	30	Mercedes-Benz E350	24
Audi A6 Quattro	23	Hyundai Elantra	33	Mercury Milan	29
Bentley Arnage	14	Jaguar XF	25	Mitsubishi Galant	27
BMW 5281	28	Kia Optima	32	Nissan Maxima	26
Buick Lacrosse	28	Lexus GS 350	26	Rolls Royce Phantom	18
Cadillac CTS	25	Lincolon MKZ	28	Saturn Aura	33
Chevrolet Malibu	33	Mazda 6	29	Toyota Camry	31
Chrysler Sebring	30	Mercedes-Benz E350	24	Volkswagen Passat	29



Describe the shape, center, and spread of the distribution. Are there any outliers?

escribing

hape

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

Dis
pla
yin
g
Qu
anti
ati
ve
Dat
a

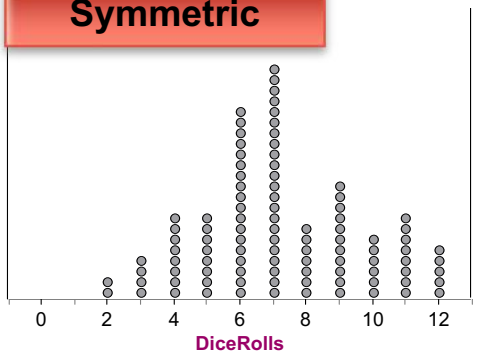
Definitions:

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) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

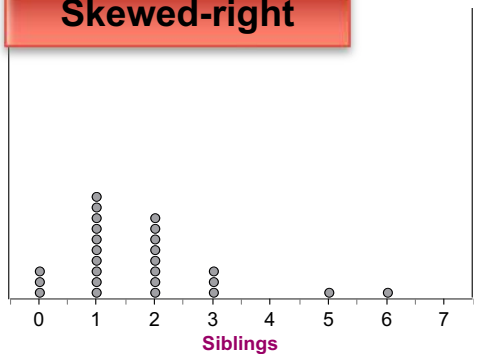
It is **skewed to the left** (left-skewed) if the left side of the graph is much longer than the right side.

Symmetric



Skewed-left

Skewed-right



Comparing Distributions

- Some of the most interesting statistics questions involve comparing two or more groups.
- Always discuss shape, center, spread, and possible outliers whenever you compare distributions of a quantitative variable.

Example, page 32

Compare the distributions of household size for these two countries. Don't forget your SOCS!

emplots
tem-and-
eaf Plots)

Another simple graphical display for small data sets is a stemplot. Stemplots give us a quick picture of the distribution while including the actual numerical values.

Displaying Quantitative Data

How to Make a Stemplot

1)Separate each observation into a **stem** (all but the final digit) and a **leaf** (the final digit).

2)Write all possible stems from the smallest to the largest in a vertical column and draw a vertical line to the right of the column.

3)Write each leaf in the row to the right of its stem.

4)Arrange the leaves in increasing order out from the stem.

5)Provide a key that explains in context what the stems and leaves represent.

Displaying Quantitative Data

Stemplots (Stem-and-Leaf Plots)

These data represent the responses of 20 female AP Statistics students to the question, "How many pairs of shoes do you have?" Construct a stemplot.

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

1	
2	
3	
4	
5	

Stems

1	93335
2	664233
3	1840
4	9
5	0701

Add leaves

1	33359
2	233466
3	0148
4	9
5	0017

Order leaves

Key: 4|9 represents a female student who reported having 49 pairs of shoes.

Add a key

Displaying Quantitative Data

Splitting Stems and Back-to-Back Stemplots

- When data values are "bunched up", we can get a better picture of the distribution by **splitting stems**.
- Two distributions of the same quantitative variable can be compared using a **back-to-back stemplot** with common stems.

Females

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Males

14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

"split stems"

	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Females | **Males**

Key: 4|9 represents a student who reported having 49 pairs of shoes.

Histograms

- Quantitative variables often take many values. A graph of the distribution may be clearer if nearby values are grouped together.
- The most common graph of the distribution of one quantitative variable is a **histogram**.

Displaying Quantitative Data

How to Make a Histogram

1)Divide the range of data into classes of equal width.

2)Find the count (*frequency*) or percent (*relative frequency*) of individuals in each class.

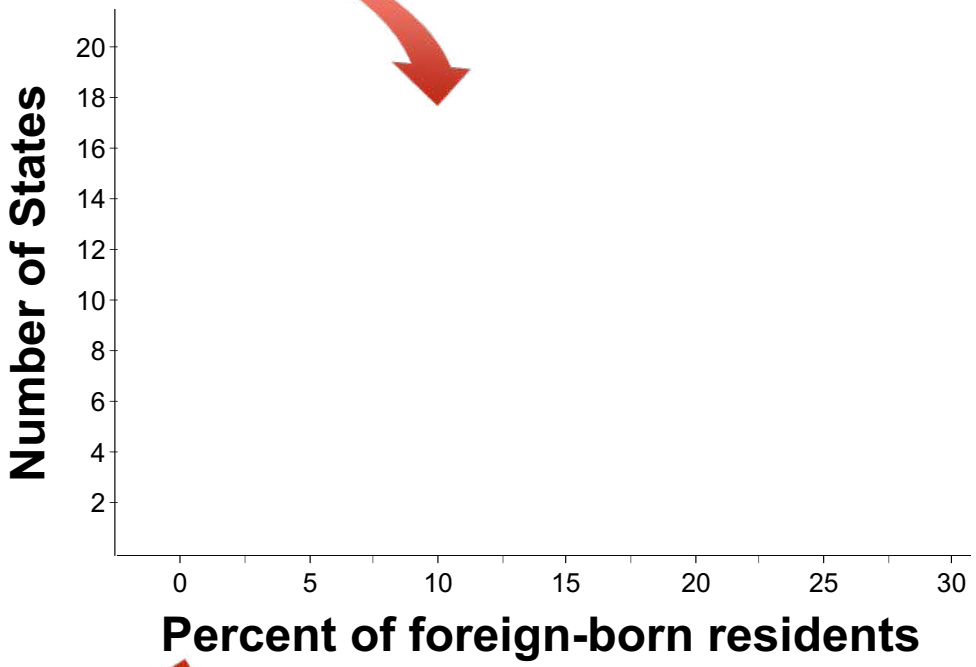
3)Label and scale your axes and draw the histogram. The height of the bar equals its frequency. Adjacent bars should touch, unless a class contains no individuals.

Example, page 35

Making a Histogram

The table on page 35 presents data on the percent of residents from each state who were born outside of the U.S.

Frequency Table	
Class	**Count**
0 to <5	20
5 to <10	13
10 to <15	9
15 to <20	5
20 to <25	2
25 to <30	1
Total	50



Using Histograms Wisely

Here are several cautions based on common mistakes students make when using histograms.

Cautions

1)Don't confuse *histograms* and *bar graphs.*

2)Don't use counts (in a frequency table) or percents (in a relative frequency table) as data.

3)Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations.

4)Just because a graph looks nice, it's not necessarily a meaningful display of data.

Section 1.2
Displaying Quantitative Data with Graphs

Summary

In this section, we learned that…

- ✓ You can use a **dotplot, stemplot, or histogram** to show the distribution of a quantitative variable.
- ✓ When examining any graph, look for an **overall pattern** and for notable **departures** from that pattern. Describe the **shape, center, spread,** and any **outliers**. Don't forget your SOCS!
- ✓ Some distributions have simple shapes, such as **symmetric** or **skewed**. The number of **modes** (major peaks) is another aspect of overall shape.
- ✓ When comparing distributions, be sure to discuss shape, center, spread, and possible outliers.
- ✓ Histograms are for quantitative data, bar graphs are for categorical data. Use relative frequency histograms when comparing data sets of different sizes.

Looking Ahead…

In the next Section…

We'll learn how to describe quantitative data with numbers.

- ü**Mean and Standard Deviation**
- ü**Median and Interquartile Range**
- ü**Five-number Summary and Boxplots**
- ü**Identifying Outliers**

We'll also learn how to calculate numerical summaries with technology and how to choose appropriate measures of center and spread.

Chapter 1: Exploring Data

Section 1.3

Describing Quantitative Data with Numbers

The Practice of Statistics, 4th edition - For AP*

STARNES, YATES, MOORE

Chapter 1
Exploring Data

- **Introduction**: Data Analysis: Making Sense of Data
- **1.1** Analyzing Categorical Data
- **1.2** Displaying Quantitative Data with Graphs
- **1.3** **Describing Quantitative Data with Numbers**

Section 1.3
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

asuring
nter: The
an

e most
mmon
easure of
nter is the
dinary
thmetic
erage, or
ean.

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:

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

De
scri
bin
g
Qu
anti
ati
ve
Dat
a

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.

De
scri
bin
g
Qu
anti
ati
ve
Dat
a

Describing Quantitative Data

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 53

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.

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:

$$IQR = Q_3 - Q_1$$

Example, page 57

Find a Interpret the IQR

Travel times to work for 20 randomly selected New Yorkers

10	30	45	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

$Q_1 = 15$ $M = 22.5$ $Q_3 = 42.5$

$$IQR = Q_3 - Q_1$$
$$= 42.5 - 15$$
$$= 27.5 \text{ minutes}$$

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

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, page 57

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**

De
scri
bin
g
Qu
anti
ati
ve
Dat
a

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 Q_1 ***M*** Q_3 ***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.

Example

Construct a Boxplot.

Consider your NY trip 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

Min=5

$Q_1 = 15$

$M = 22.5$

$Q_3 = 42.5$

Max=85

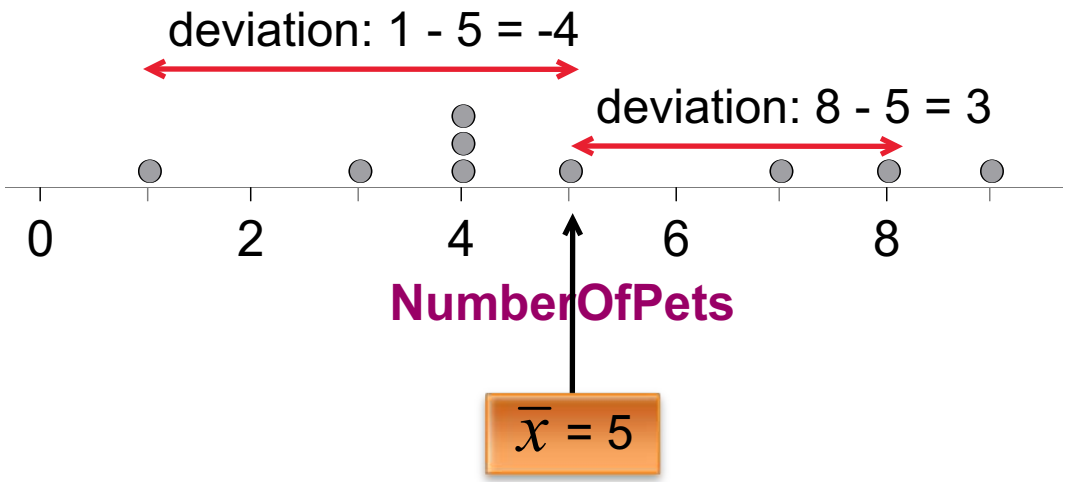
Recall, this is an outlier by the 1.5 x IQR rule

Describing Quantitative Data

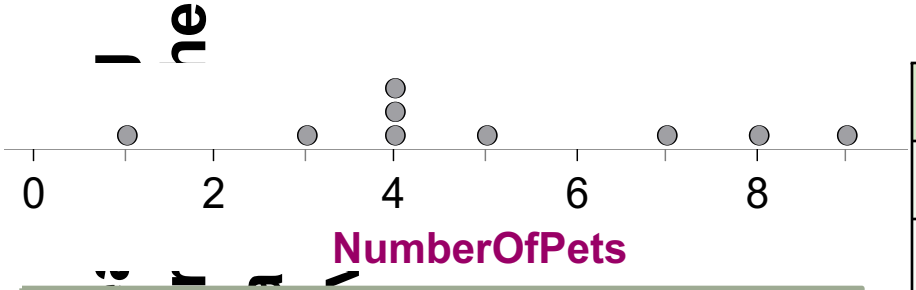
Measuring Spread: The Standard Deviation

- The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation**. Let's explore it!
- Consider the following data on the number of pets owned by a group of 9 children.

1) Calculate the mean.
2) Calculate each *deviation.*
   deviation = observation – mean



Describing Quantitative Data



x_i
1
3
4
4
4
5
7
8
9

3) Square each deviation.

4) Find the "average" squared deviation. Calculate the sum of the squared deviations divided by (n-1)…this is called the **variance.**

5) Calculate the square root of the variance…this is the **standard deviation.**

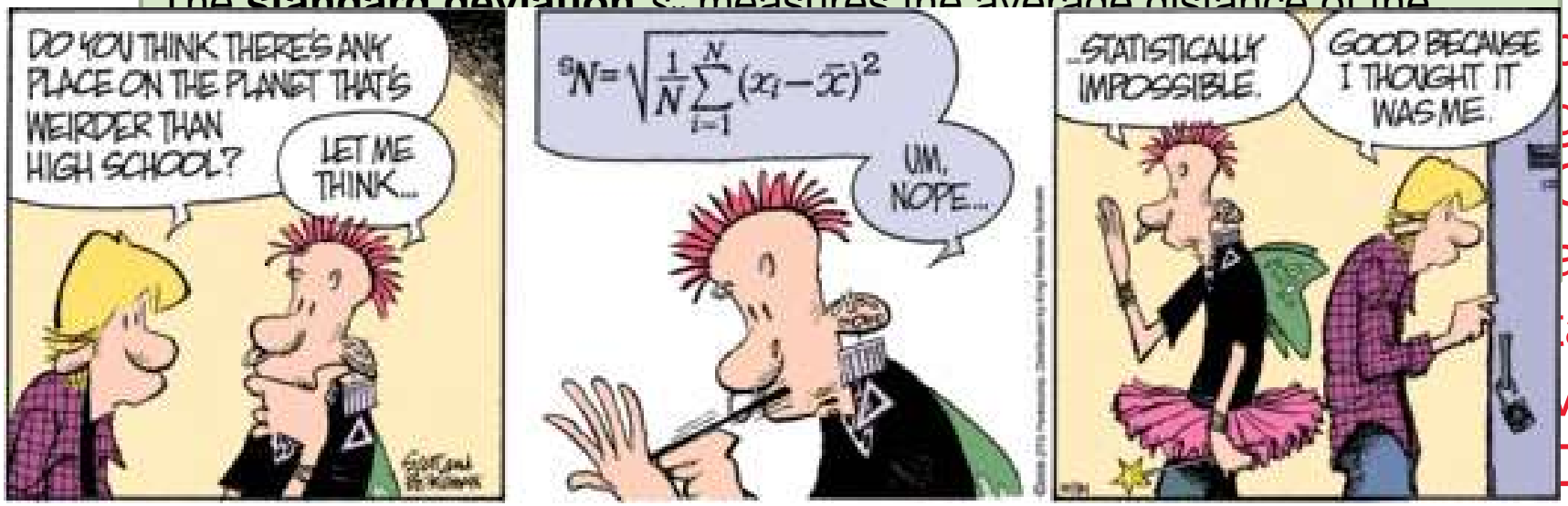
"average" squared deviation = 52/(9-1) = 6.5 This is the **variance.**

Standard deviation = square root of variance = $\sqrt{6.5} = 2.55$

ing
The
rd
on

Definition:

The **standard deviation** s_x measures the average distance of the



$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

- **Choosing Measures of Center and Spread**
  - We now have a choice between two descriptions for center and spread
    - Mean and Standard Deviation
    - Median and Interquartile Range

Choosing Measures of Center and Spread

•The median and *IQR* are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.

•Use mean and standard deviation only for reasonably symmetric distributions that don't have outliers.

•**NOTE: Numerical summaries do not fully describe the shape of a distribution.** ***ALWAYS PLOT YOUR DATA!***

Section 1.3
Describing Quantitative Data with Numbers

Summary

In this section, we learned that…

- ✓ A numerical summary of a distribution should report at least its **center** and **spread**.
- ✓ The **mean** and **median** describe the center of a distribution in different ways. The mean is the average and the median is the midpoint of the values.
- ✓ When you use the median to indicate the center of a distribution, describe its spread using the **quartiles**.
- ✓ The **interquartile range (*IQR*)** is the range of the middle 50% of the observations: $IQR = Q_3 - Q_1$.

Section 1.3
Describing Quantitative Data with Numbers

Summary

In this section, we learned that…

- ✓ An extreme observation is an **outlier** if it is smaller than Q_1–(1.5x*IQR*) or larger than Q_3+(1.5x*IQR*) .
- ✓ The **five-number summary** (*min,* Q_1*, M,* Q_3*, max*) provides a quick overall description of distribution and can be pictured using a **boxplot.**
- ✓ The **variance** and its square root, the **standard deviation** are common measures of spread about the mean as center.
- ✓ The mean and standard deviation are good descriptions for symmetric distributions without outliers. The median and *IQR* are a better description for skewed distributions.

Looking Ahead…

In the next Chapter…

We'll learn how to model distributions of data…

- **Describing Location in a Distribution**
- **Normal Distributions**