Chapter 4: Numerical Methods for Describing Data

Describing Quantitative Data with Numbers

Chapter 4.2 & 4.3 Numerical Methods for Describing 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**

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:

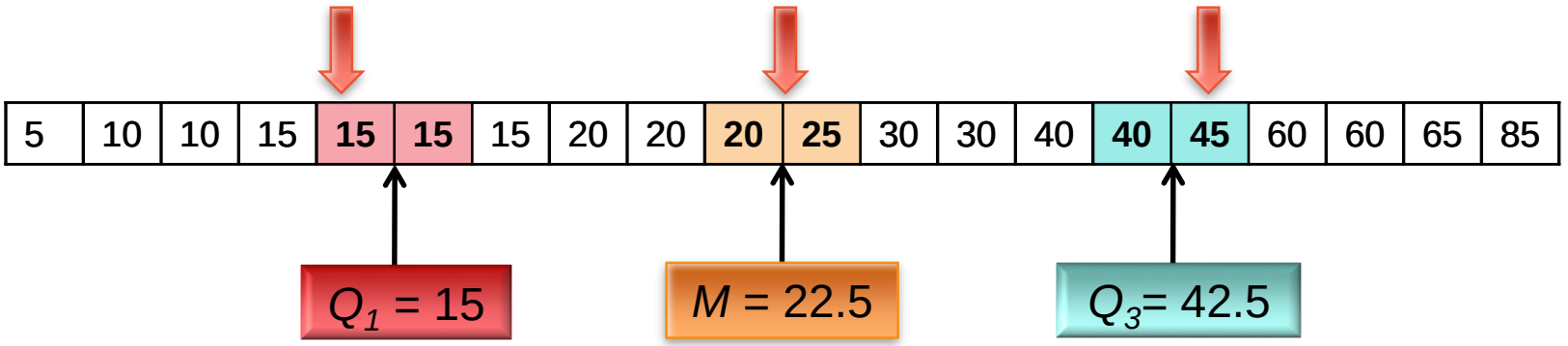
$$IQR = Q_3 - Q_1$$

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



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

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

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 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.

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

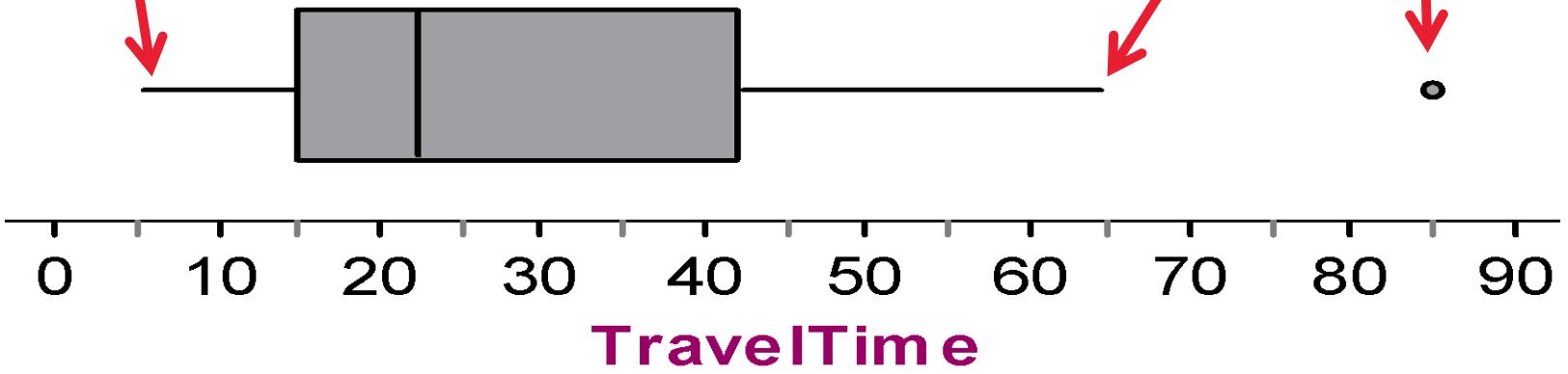
Min=5

$Q_1 = 15$

$M = 22.5$

$Q_3 = 42.5$

Max=85
Recall, this is an outlier by the 1.5 x IQR rule



Boxplots (Box-and-Whisker Plots) review

- The five-number summary divides the distribution in to numerical values that leads to a **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.

Sept 18th/19th AGENDA

- Warm-Up: Review of Box plots
- Deviation Activity
- Continue Chapter 4 (Finish material thru Section 4.3)
- *FRAPPY Time (YEA! 2006 Ques.#5)*

Sept 18th /19th Warm UP

Practicing Box and Whisker Plots with the 5 Number summary

1. Find the five number summary for each set of data:

Data Set A: 4, 5, 7, 9, 11

Data Set B: 4, 5, 7, 8, 9, 11

2. Find the IQR for each set

3. Make a box plot for each data set.

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***

Warm UP

Practicing Box and Whisker Plots with the 5 Number summary

1. Five number summary:

Data Set A: 4, 5, 7, 9, 11

Min	Q_1	Med	Q_3	Max
4	**4.5**	**7**	**10**	**11**

Data Set B: 4, 5, 7, 8, 9, 11

Min	Q_1	Med	Q_3	Max
4	**5**	**7.5**	**9**	**11**

IQR

$(Q_3 - Q_1)$

IQR = **5.5**

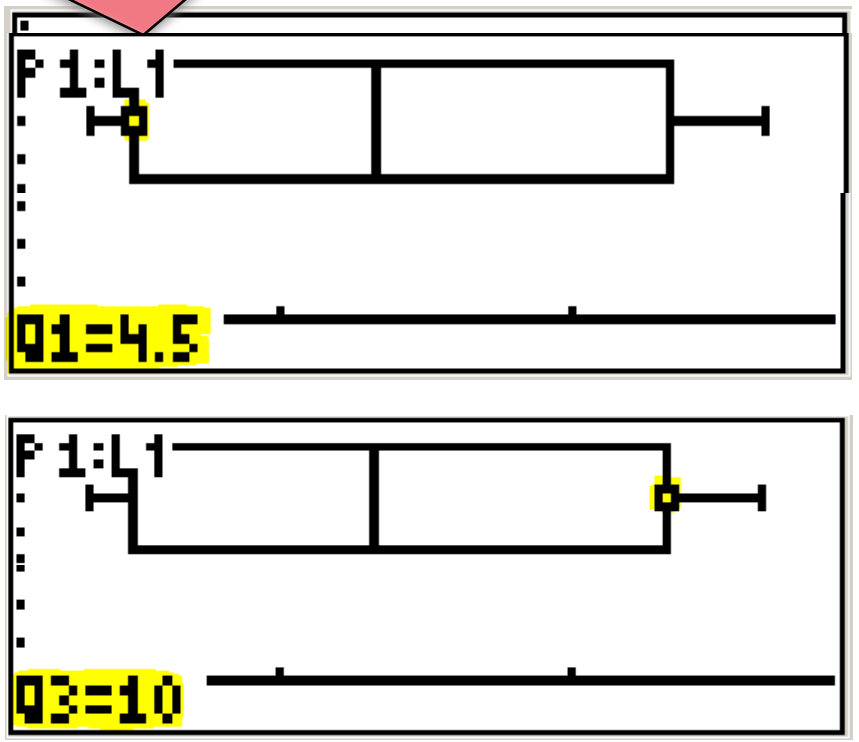
IQR = **4**

Comparing Box Plots & Five Number summary

Sample Data

L1	L2
4	4
5	5
7	7
9	8
11	9
------	11

Sample Box and Whisker Plots



Deviation Activity (already completed?)

- Answer the first 5 questions. Be prepared to share/discuss with your peers.
- Work on the remainder of problems if you can figure them out on your own.

Deviation Activity

- Day Shift Mean: **8.1** Median: **8.1**
- Night Shift Mean: **8.0** Median: **8.0**
- 1) Which shift average was closer?

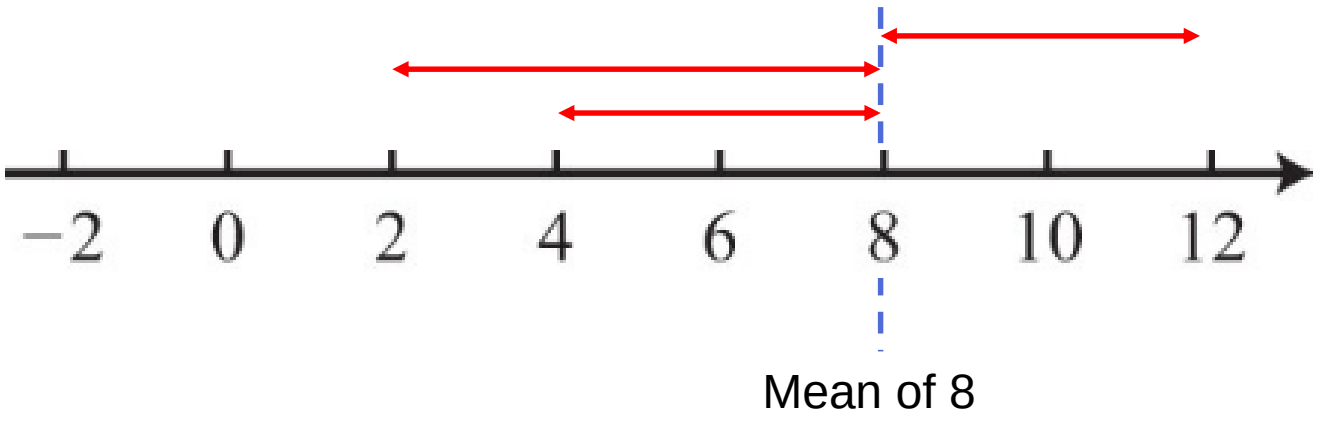
Night shift

- 2) What shift do you want to manufacture your part?

Day shift? why?

Deviation Activity

- Line plot of the data, showing the deviations from the mean.



Deviation from the mean

- What do you notice about the individual deviations $(x_i - \bar{x})$?
- What properties will be true for all data sets?
- What is the meaning of "***deviation from the mean***"?
- Can we create a "**standard deviation**" for all of the data within any given sample?

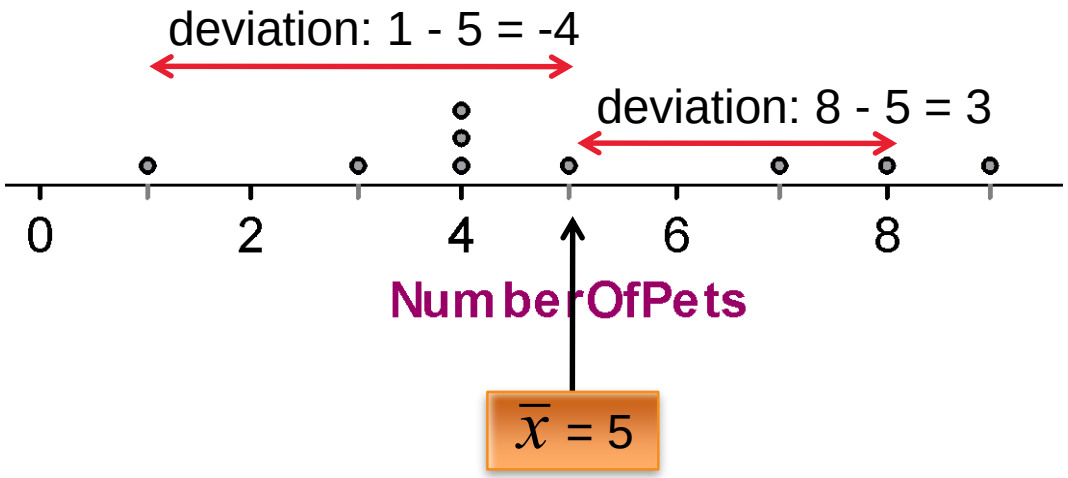
Deviations from the mean

- What kind of average is the MAD (mean absolute deviation)?
- Why does taking absolute value improve our outcome?
- What kind of average is the S.D. (standard deviation)?

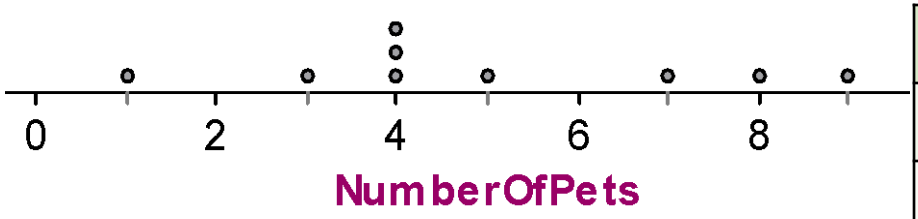
Measuring Spread: The Standard Deviation

- 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



Measuring Spread: The Standard Deviation



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 = $\sqrt{6.5} = 2.55$

Standard deviation – class activity:
Practice using your calculators

- Use the data set given from Mr. L.
- Find the sample mean ($\bar{x}$)
- Find the deviations from the mean ($x_i - \bar{x}$)
- Square the deviations and find the sum of these values
- Find the "average" of this sum (divide by *n* or $n - 1$)

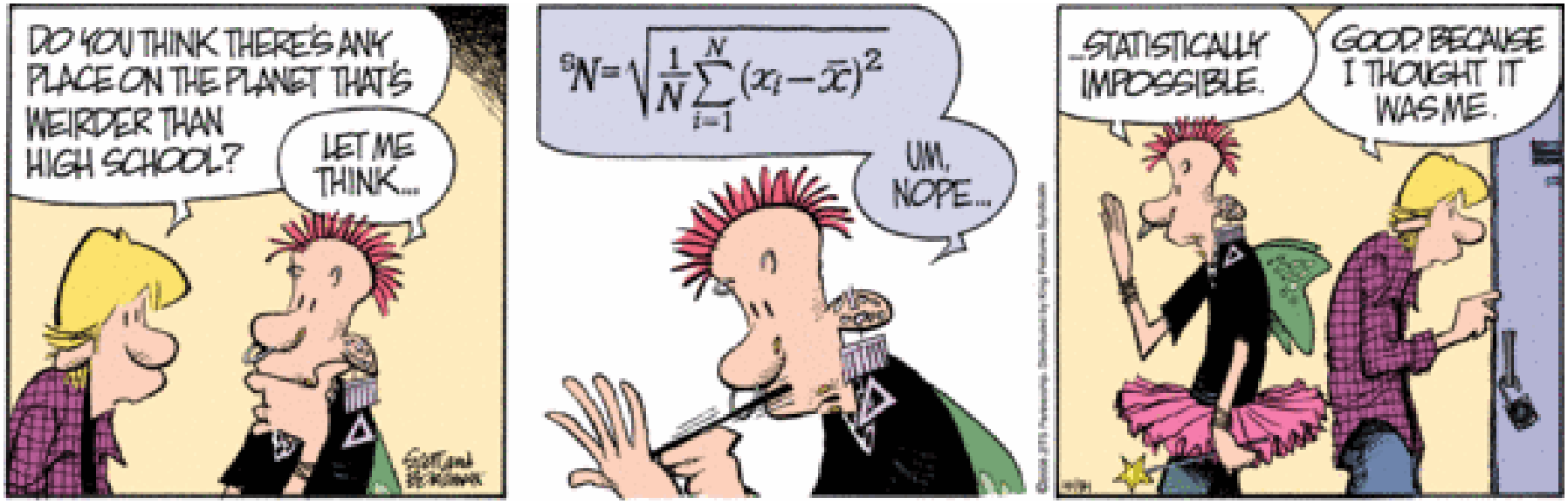
Measuring Spread: The Standard Deviation

Definition:

The **standard deviation** s_x measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. This average squared distance is called the **variance**.

$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

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


DO YOU THINK THERE'S ANY PLACE ON THE PLANET THAT'S WEIRDER THAN HIGH SCHOOL?
LET ME THINK...
$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
UM, NOPE...
...STATISTICALLY IMPOSSIBLE.
GOOD. BECAUSE I THOUGHT IT WAS ME.

Notation for Variance and Standard Deviation: Sample vs. Population

SAMPLE Notation

- Sample mean: $\bar{x}$
- Sample Variance:

$$s_x^2 \text{ or } s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Sample S.D.

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

Population Notation

- Population Mean: μ
- Population Variance:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

- Population S.D.:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

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 4.1 & 4.2
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 4.2 & 4.3
Describing Quantitative Data with Numbers

Summary

In this section, we learned that…

- ✓ An extreme observation is an **outlier** if it is smaller than $Q_1 - (1.5 \times IQR)$ or larger than $Q_3 + (1.5 \times IQR)$.
- ✓ 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.

History of Variance (FYI)

- The term *variance* was first introduced by Ronald Fisher in his 1918 paper *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*:[15]

- **Summarizing spread of distributions: Khan Academy videos** - https://www.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/v/range-variance-and-standard-deviation-as-measures-of-dispersion

- **History, and students' understanding of variance in statistics by** Michael Kourkoulos & Tzanakis

- In this article we examine students' difficulties in understanding variance and standard deviation in introductory statistics, taking into consideration relevant elements from the historical evolution of these concepts.

Looking Ahead…

In the next part of Chapter 4…

We'll learn how to model distributions of data…

- **Describing Location in a Distribution**
- **Normal Distributions**
- **Calculating Z scores**