CHAPTER 1
Exploring Data

1.3
Describing Quantitative Data with Numbers



The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore

Describing Quantitative Data with Numbers

Learning Objectives

After this section, you should be able to:

- ✓ CALCULATE measures of center (mean, median).
- ✓ CALCULATE and INTERPRET measures of spread (range, *IQR*, standard deviation).
- ✓ CHOOSE the most appropriate measure of center and spread in a given setting.
- ✓ IDENTIFY outliers using the 1.5 × *IQR* rule.
- ✓ MAKE and INTERPRET boxplots of quantitative data.
- ✓ USE appropriate graphs and numerical summaries to compare distributions of quantitative variables.

Measuring Center: The Mean

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

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 Σ is 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**. The median describes the midpoint of a distribution.

The **median** 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 is the center observation in the ordered list.
3. If the number of observations *n* is even, the median 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.

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.

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

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 *IQR:*

To calculate the **quartiles**:

1. Arrange the observations in increasing order and locate the median.

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*

Travel times for 20 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

$Q_1 = 15$ | *Median* = 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.

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.

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.

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 ***Median*** 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:

- A central box is drawn from the first quartile (Q_1) to the third quartile (Q_3).
- A line in the box marks the median.
- Lines (called whiskers) extend from the box out to the smallest and largest observations that are not outliers.
- Outliers are marked with a special symbol such as an asterisk (*).

Construct a Boxplot

Consider our New York travel time data:

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

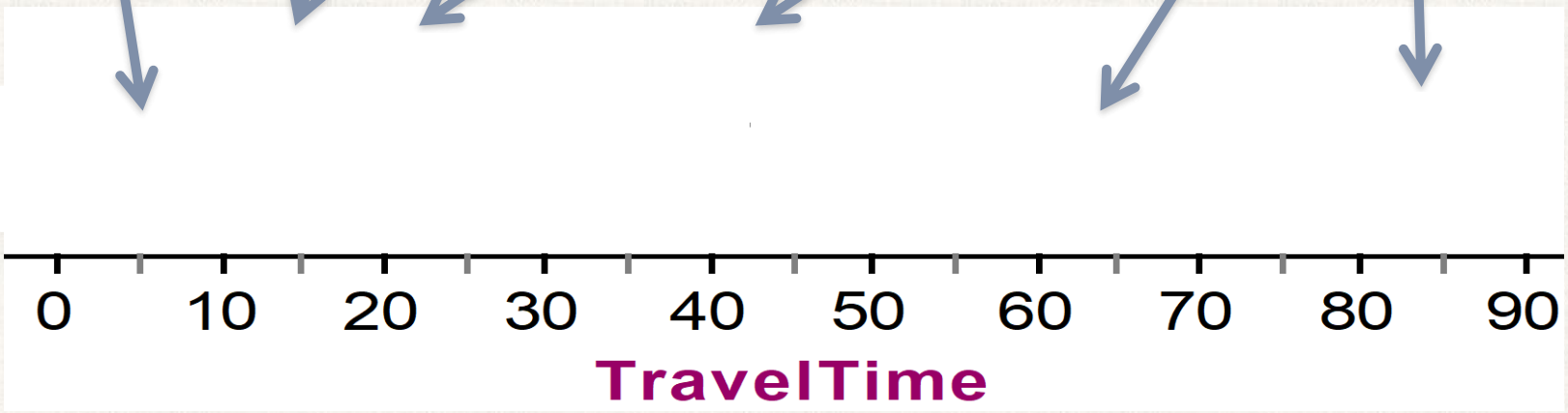
Q_1 = 15

Median = 22.5

Q_3= 42.5

Max=85

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

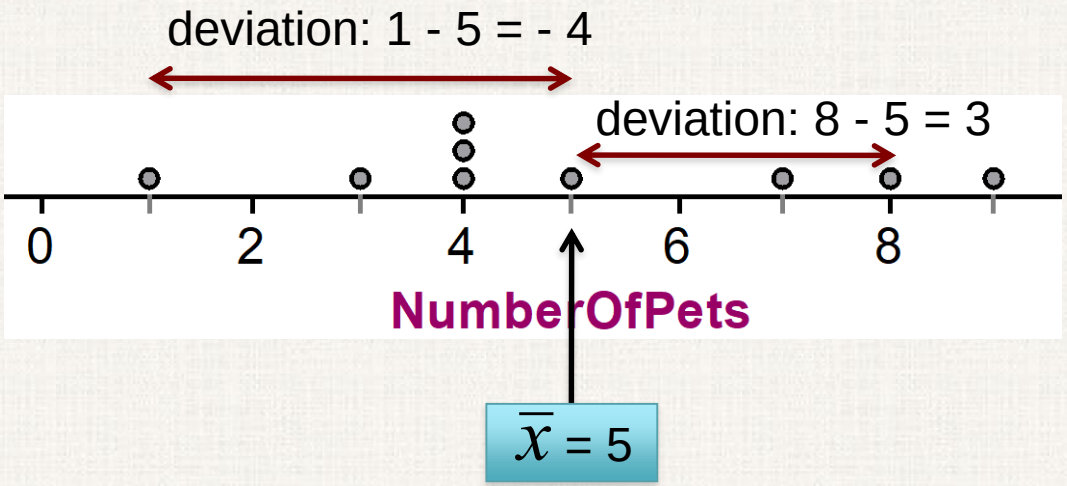


Measuring Spread: The Standard Deviation

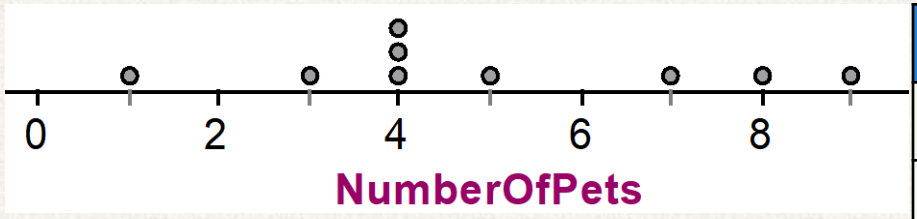
The most common measure of spread looks at how far each observation is from the mean. This measure is called the standard deviation.

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



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.**

x_i	(x_i-mean)
1	1 - 5 = -4
3	3 - 5 = -2
4	4 - 5 = -1
4	4 - 5 = -1
4	4 - 5 = -1
5	5 - 5 = 0
7	7 - 5 = 2
8	8 - 5 = 3
9	9 - 5 = 4
	Sum=?

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

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

Measuring Spread: The Standard Deviation

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.

The average squared distance is called the **variance**.

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

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

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!**

Organizing a Statistical Problem

As you learn more about statistics, you will be asked to solve more complex problems. Here is a four-step process you can follow.

How to Organize a Statistical Problem: A Four-Step Process

- **State**: What's the question that you're trying to answer?
- **Plan**: How will you go about answering the question? What statistical techniques does this problem call for?
- **Do**: Make graphs and carry out needed calculations.
- **Conclude**: Give your conclusion in the setting of the real-world problem.

Data Analysis: Making Sense of Data

Section Summary

In this section, we learned how to…

- ✓ CALCULATE measures of center (mean, median).
- ✓ CALCULATE and INTERPRET measures of spread (range, *IQR*, standard deviation).
- ✓ CHOOSE the most appropriate measure of center and spread in a given setting.
- ✓ IDENTIFY outliers using the 1.5 × *IQR* rule.
- ✓ MAKE and INTERPRET boxplots of quantitative data.
- ✓ USE appropriate graphs and numerical summaries to compare distributions of quantitative variables.