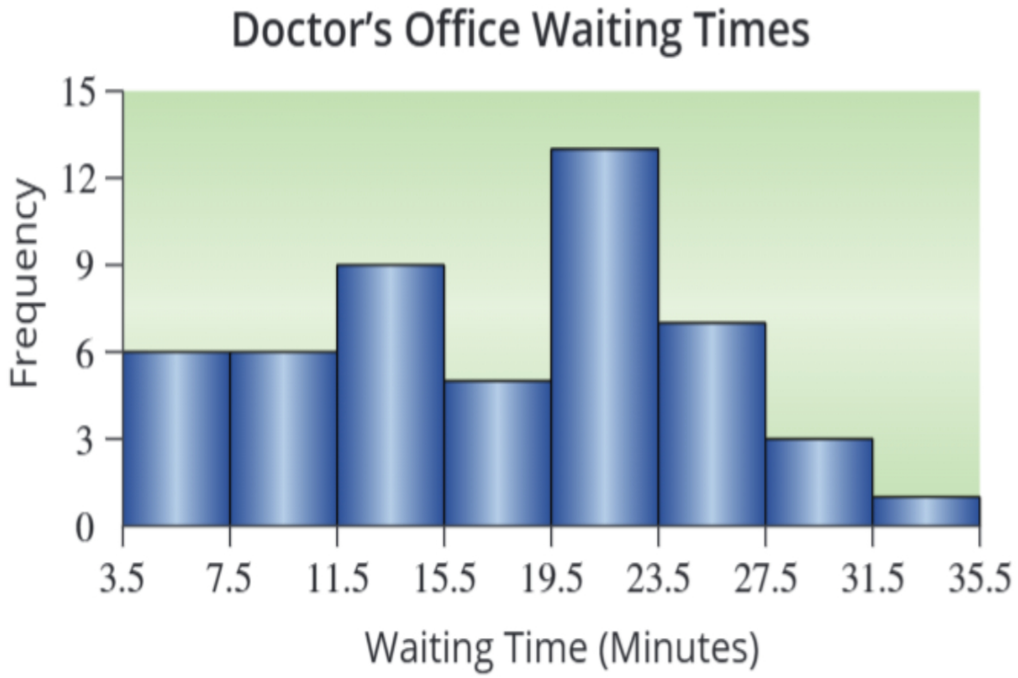
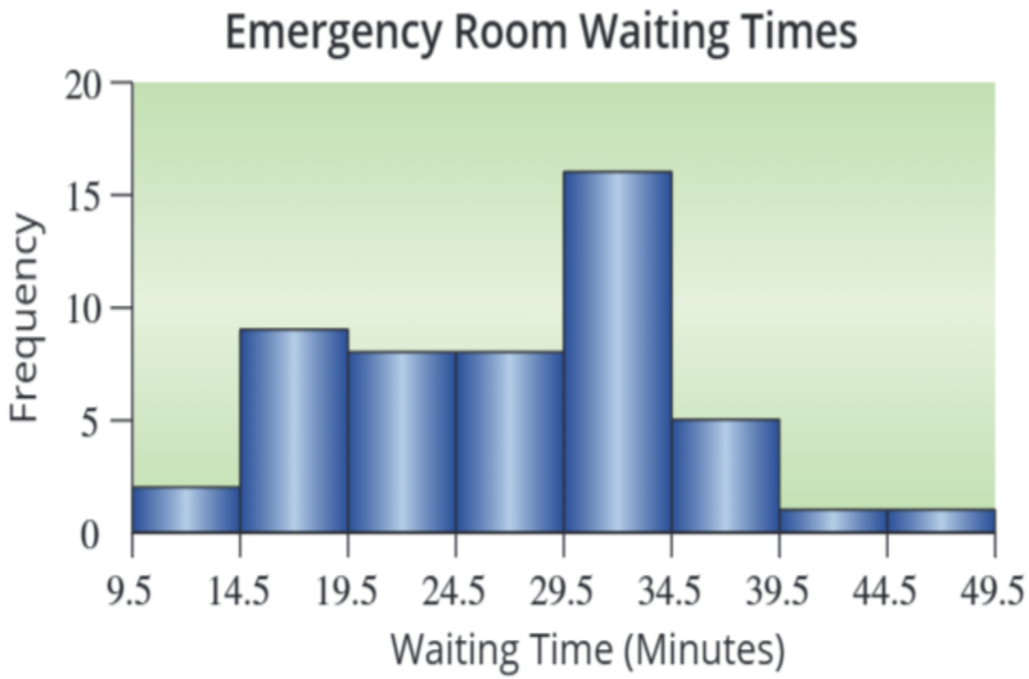
4 CHAPTER
Describing and Summarizing Data from One Variable





Attributes for Summarizing Data

To adequately summarize a set of data, a data analyst might ask the following questions.

- Location: Where is the center of the data?
- Dispersion: Is the data widely scattered or tightly grouped around the central point?
- Shape: Is the data spread symmetrically about the central value? Is the data unbalanced (e.g., are the values much larger than the mean, but not much smaller)?
- Does the data tend to cluster in several groups?

PROPERTIES

- Poverty rates in 2016 ranged from a low of 7.3% in New Hampshire to a high of 20.8% in Mississippi.[1]
- In 2016, Maryland ($78,945) and Alaska ($76,440) had median household incomes that were among the highest; Mississippi had the lowest ($41,754).[2]
- In 2016, 22.9% of 18 to 34-year-olds living in households, lived in their parents' home.[3] (Compare this to 14.7% in 1975.)
- The median price of an existing home in the U.S., as of August 2017, was $253,500.[4]
- The average hourly manufacturing earnings in the U.S. in August 2017 was $20.90. (Compare this to the average of $14.84 in 2002.)[5]

Numerical Descriptive Statistics

Numerical descriptive statistics are numerical summaries of quantitative data.

DEFINITION

Inferential Statistics

Inferential statistics is concerned with making conclusions or inferences about population parameters using sample statistics.

DEFINITION

Parameter vs. Statistic

A **parameter** is a numerical measure that describes a characteristic of a *population*.

A **statistic** is a numerical measure that describes a characteristic of a *sample*.

DEFINITION

Arithmetic Mean

Suppose there are n observations in a data set, consisting of the observations $x_1, x_2, \ldots, x_n$; then the **arithmetic mean** is defined to be

$$\frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right).$$

DEFINITION

Calculate the sample mean of the following sample data values: 4, 10, 7, 15.

$x_1 = 4,\ x_2 = 10,\ x_3 = 7,\ x_4 = 15,$ and $n = 4.$

$\overline{X} \rightarrow$ mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4 + 10 + 7 + 15}{4} = \frac{36}{4} = 9.$$

$\sum \rightarrow$ sumation

Weighted Mean

The weighted mean of a data set with values $x_1, x_2, x_3, \ldots, x_n$ is given by

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum(w_ix_i)}{\sum w_i}$$

where w_i is the weight of observation x_i.

FORMULA

Meghan's Grades		
Course	**Grade**	**Credit Hours**
Psychology 101	B 3	3
Probability and Statistics	A 4	4
Anatomy I	C 2	5
English 101	A 4	3

→ 9
→ 16
→ 10
→ 12

15

47

~~$\frac{47}{4} = 11.75$~~

$\frac{47}{15} = 3.13$

Trimmed Mean

The **trimmed mean** is a modification of the arithmetic mean which ignores an equal percentage of the highest and lowest data values in calculating the mean.

DEFINITION

Consider the same data set, except the last data value is replaced with an outlier.

~~16~~ 18 20 21 23 23 24 32 36 ~~490~~

mean = 70.3

Find the 10% trimmed mean.

24.625

Median

The **median** of a set of observations is the measure of center that is the middle value of the data when it is arranged in ascending order. The same number of data values lie on either side of the median.

DEFINITION

To determine the median of a set of data, we use the following steps.

Finding the Median of a Data Set

1. Arrange the data in ascending order.
2. Determine the number of values in the data.
3. Find the data value in the middle of the data set.
4. If the number of data values is odd, then the median is the data value that is exactly in the middle of the data set.
5. If the number of data values is even, then the median is the mean of the two middle observations in the data set.

PROCEDURE

Consider the following ten test scores from a student taking a high school calculus class.

65, 98, 76, 83, 94, 79, 88, 72, 90, 85

Find the median.

65, 72, 76, 79, 83, 85, 88, 90, 94, 98

median = 84

Consider the following goal tallies from eleven games played by the Charleston Battery soccer team.

2, 3, 5, 4, 1, 7, 3, 3, 1, 2, 6

Find the median.

1, 1, 2, 2, 3, 3, 3, 4, 5, 6, 7

Mode

The **mode** of a data set is the most frequently occurring value.

DEFINITION

Find the mode of the following data regarding the number of power outages reported over a period of eleven days.

0, 1, 4, 3, 9, 8, 10, 0, 1, 3

0, 1, 3

Outliers and Resistant Measures

An **outlier** is a data value that is extremely different from other measurements in the data set. Statistical measures which are not affected by outliers are said to be **resistant**.

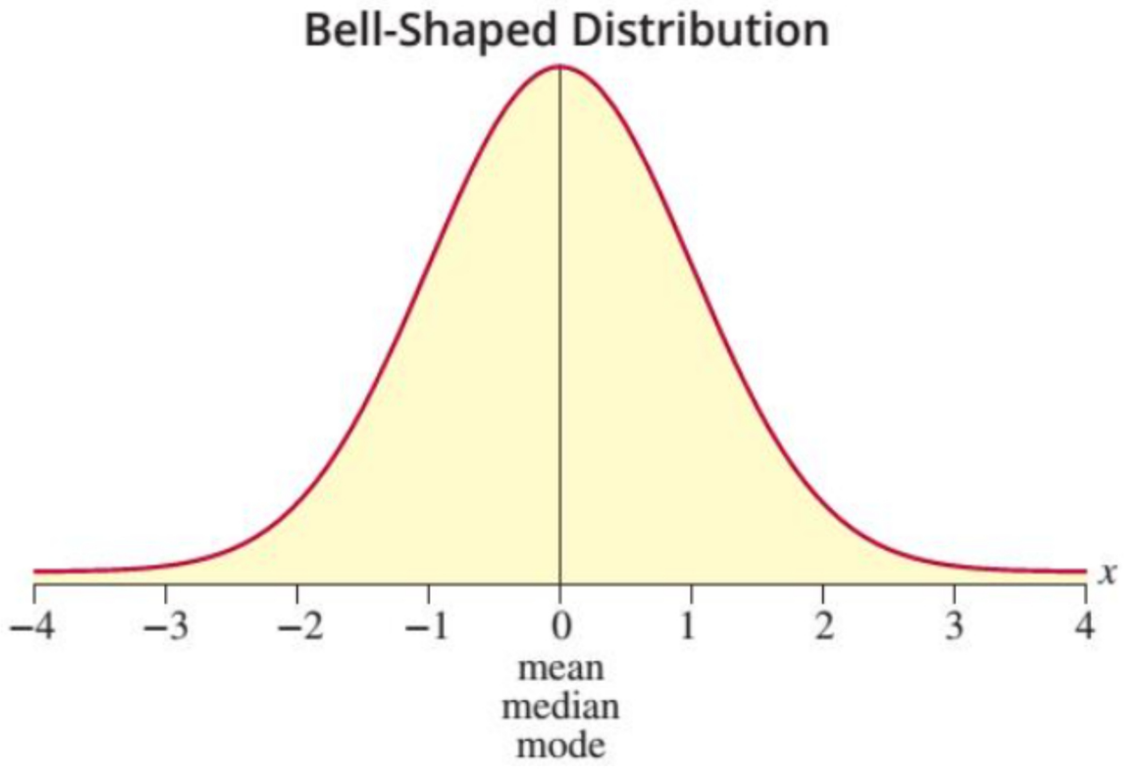
DEFINITION

Table 4.1.3 – Applicable Level of Measurement

	Qualitative		Quantitative	
	Nominal	Ordinal	Interval	Ratio
Mean			✔	✔
Median		✔	✔	✔
Mode	✔	✔	✔	✔
Trimmed Mean			✔	✔

Table 4.1.4 Sensitivity to Outliers

	Not Sensitive	Very Sensitive
Mean		✔
Median	✔	
Mode	✔	
Trimmed Mean	✔	



right

Positively Skewed Curve

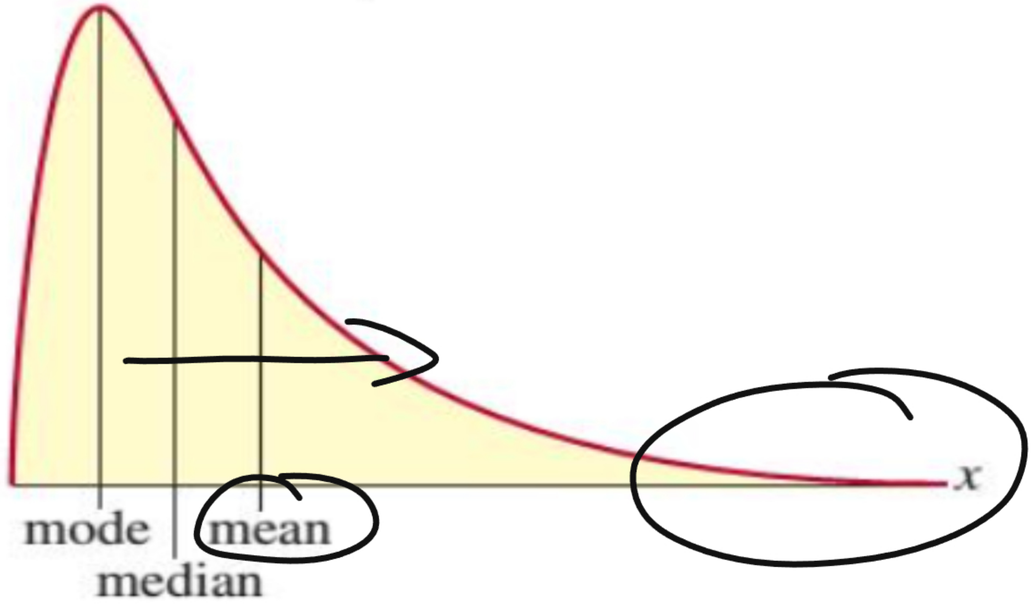


Figure 4.1.4

Left

Negatively Skewed Curve

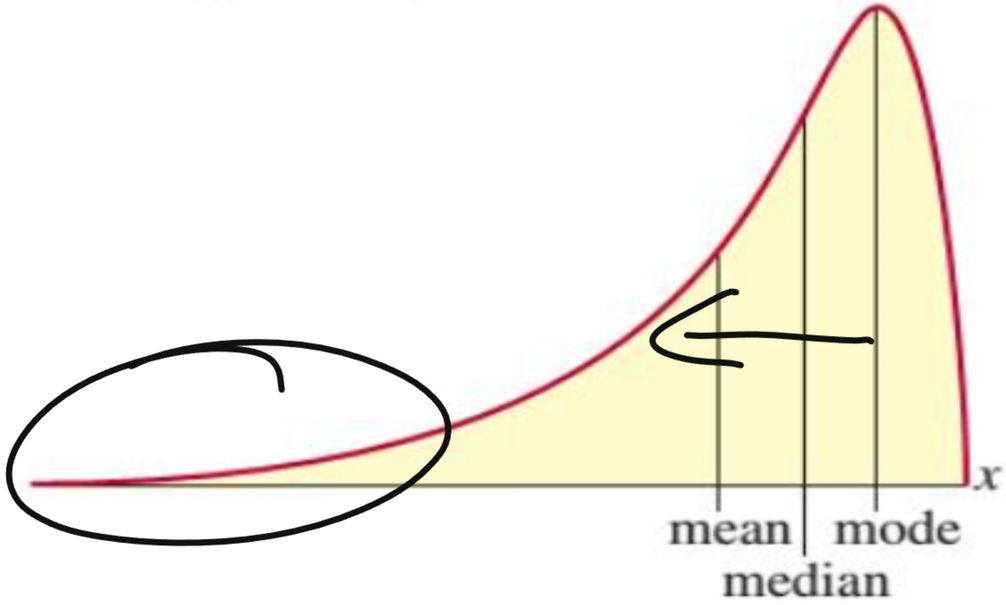


Figure 4.1.5

Many of the good measures of dispersion use the concept of deviation from the mean. If the mean is a focal point or base, use it as a common basis from which to measure variation. The distance that a point is from its mean is called a **deviation from the mean**. A data set and its deviations from the mean are calculated in Table 4.2.1.

Table 4.2.1 Calculating Deviations from the Mean

Mean = 10	
Data Values	**Deviations from the Mean (Data − Mean = Deviation)**
3	$3 - 10 = -7$
12	$12 - 10 = 2$
20	$20 - 10 = 10$
15	$15 - 10 = 5$
0	$0 - 10 = -10$

Range

The **range** is the difference between the largest and smallest data values.

DEFINITION

Calculate the range of the following data set consisting of the number of times the television channel was changed in a 1-hour time span.

4, 6, 16, 9, 24, 8, 12, 1

Variance

The **variance** of a data set containing the complete set of *population* data is given by

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

and is called the **population variance**.

σ^2 → variance
σ → standard Deviation
μ → mean

The **variance** of a data set containing *sample* data is given by

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

and is called the **sample variance**.

s^2 → variance
s → Standard Deviation
$\bar{x}$ → mean

FORMULA

Given the following times in minutes of 6 persons running a 1000-meter course, compute the sample variance.

5, 10, 9, 11, 9, 7

Calculating the Sample Variance

Data	Deviation $x_i - \bar{x}$	Squared Deviation $(x_i - \bar{x})^2$
5	$5 - 8.5 = -3.5$	12.25
10	$10 - 8.5 = 1.5$	2.25
9	$9 - 8.5 = 0.5$	0.25
11	$11 - 8.5 = 2.5$	6.25
9	$9 - 8.5 = 0.5$	0.25
7	$7 - 8.5 = -1.5$	2.25
	TOTAL	**23.5**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{23.5}{5} = 4.7 \text{ squared mi}$$

Varriance
4.7
Standard deviation
$\sqrt{4.7} = 2.17$

Given the following times in minutes of 6 persons running a 1000-meter course, compute the sample variance.

7, 8, 8, 9, 9, 10

7 - 8.5	-1.5	2.25
8 - 8.5	-.5	.25
8 - 8.5	-.5	.25
9 - 8.5	.5	.25
9 - 8.5	.5	.25
10 - 8.5	1.5	2.25

5.5 / 5

1.1 Variance

1.05 Standard deviation

Standard Deviation

The **standard deviation** is also a measure of how much the data varies around the mean. It is found by taking the square root of the variance.

DEFINITION

Properties of the Standard Deviation

- The standard deviation is always nonnegative. It is zero only if all the data values are exactly the same.
- The standard deviation can increase dramatically if there are one or more outliers in the data.
- The standard deviation is expressed in the same units as the original data values.

PROPERTIES

745	789	712	764	736
758	722	773	751	741

Standard Deviation, s: **23.173020711355**

Count, N: 10
Sum, Σx: 7491
Mean, $\bar{x}$: 749.1
Variance, s^2: 536.98888888889

Coefficient of Variation

For population data, the measure is defined as $CV = \left(\frac{\sigma}{\mu} \cdot 100 \right) \%$,

and for sample data, $CV = \left(\frac{s}{\bar{x}} \cdot 100 \right) \%$.

FORMULA

A consumer interest group is interested in comparing two brands of vitamin C. One brand of vitamin C advertises that its tablets contain 500 mg of vitamin C. The other brand advertises that its tablets contain 250 mg of vitamin C. Tablets for each brand are randomly selected and the milligrams of vitamin C for each tablet are measured with the following results.

Vitamin C Content (mg)		
	Brand A (500 mg)	**Brand B (250 mg)**
$\bar{x}$	500	250
s	10	7

a. Calculate the coefficient of variation for Brand A.

b. Calculate the coefficient of variation for Brand B.

c. Which brand more consistently produces tablets as advertised? Explain.

$\frac{s}{\bar{x}} = \frac{10}{500}$

$A = 0.02 \rightarrow 2\%$

$\frac{s}{\bar{x}} = \frac{7}{250}$

$B = 0.028 \rightarrow 2.8\%$