

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

Table 2.24

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean \bar{x} is an example of a statistic which estimates the population mean μ .

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: $mean = \frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary + upper\ boundary}{2}$. We can now modify the mean definition to be

$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example 2.30

p105

"mean of a grouped frequency table"
Find "grouped mean"

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

(mdpt)

(Class)
Bin

f) frequency

we multiply midpt * f
& add up to get
the numerator of \bar{x}
(just like
last class p101)

Bin midpoints

$$(50+56.5)/2 = 53.25$$

$$(56.5+62.5)/2 = 59.5$$

$$(62.5+68.5)/2 = 65.5$$

$$+6$$
$$71.5$$

$$+6$$
$$77.5$$

$$+6$$
$$83.5$$

$$+6$$
$$89.5$$

$$+6$$
$$95.5$$

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Table 2.25

$$n = 19$$

$$\begin{aligned} 53.25 * 1 &= 53.25 \\ 59.5 * 0 &= 0 \\ 65.5 * 4 &= 262 \\ 71.5 * 4 &= 286 \\ 77.5 * 2 &= 155 \\ 83.5 * 3 &= 250.5 \\ 89.5 * 4 &= 358 \\ 95.5 * 1 &= 95.5 \\ \hline &+ 1460.25 \end{aligned}$$

$$\bar{x} = 1460.25 / 19 = 76.85$$

Solution 2.30

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Table 2.26

sanity-check:
this looks like a
reasonable \bar{x} given
the original data
table.

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$$

$$\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$$

Try It Σ

2.30 Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Table 2.27

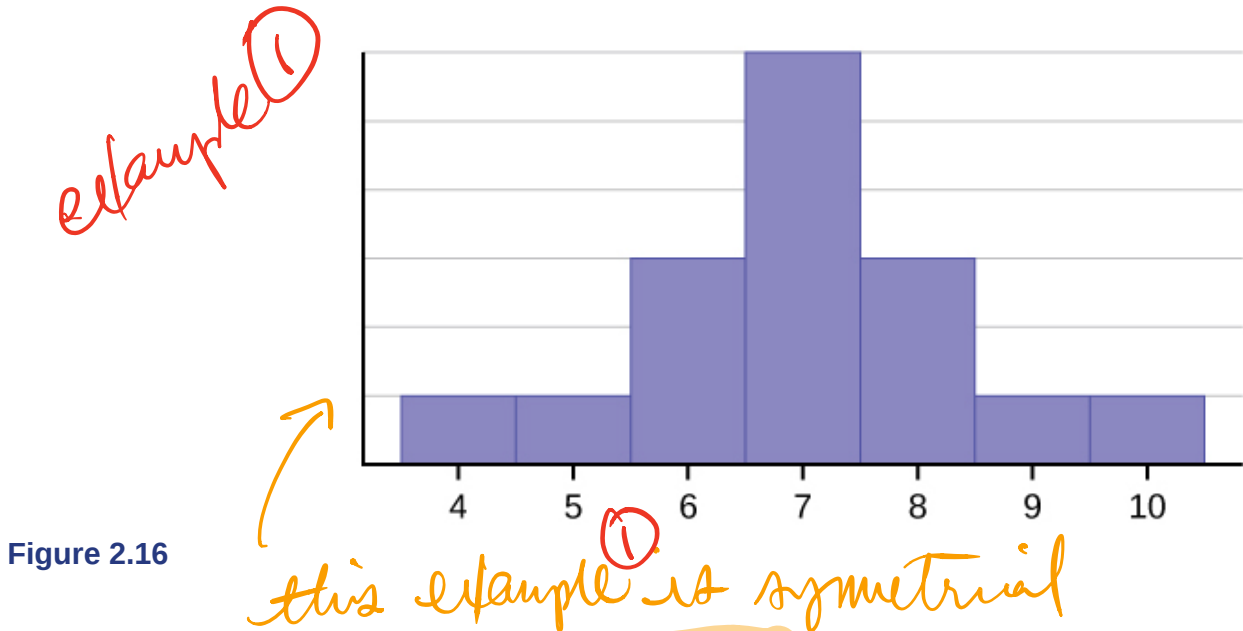
What is the best estimate for the mean number of hours spent playing video games?

p 106

2.6 | Skewness and the Mean, Median, and Mode

Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.



The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.

this example ② has left skew b/c long tail to left

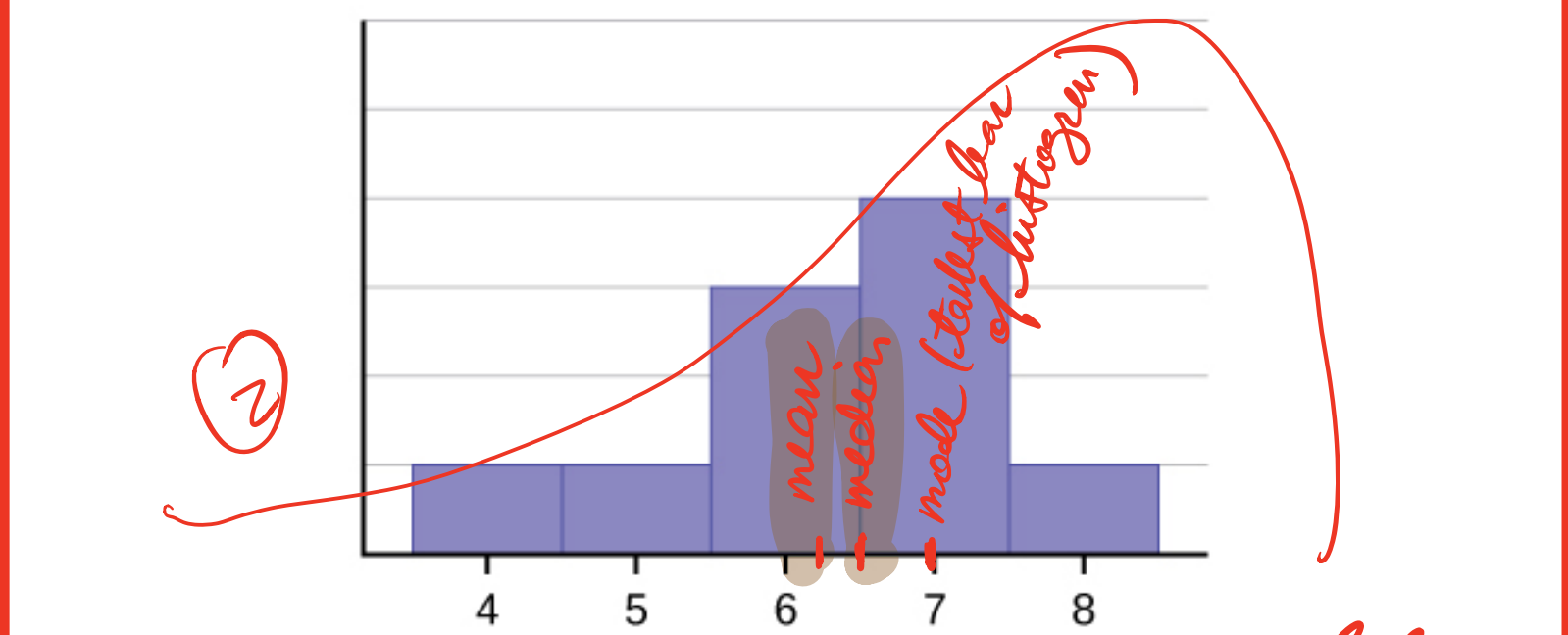


Figure 2.17 *Skew pulls median a little, mean a lot.*

The mean is 6.3, the median is 6.5, and the mode is seven. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 8; 9; 10, is also not symmetrical. It is skewed to the right.

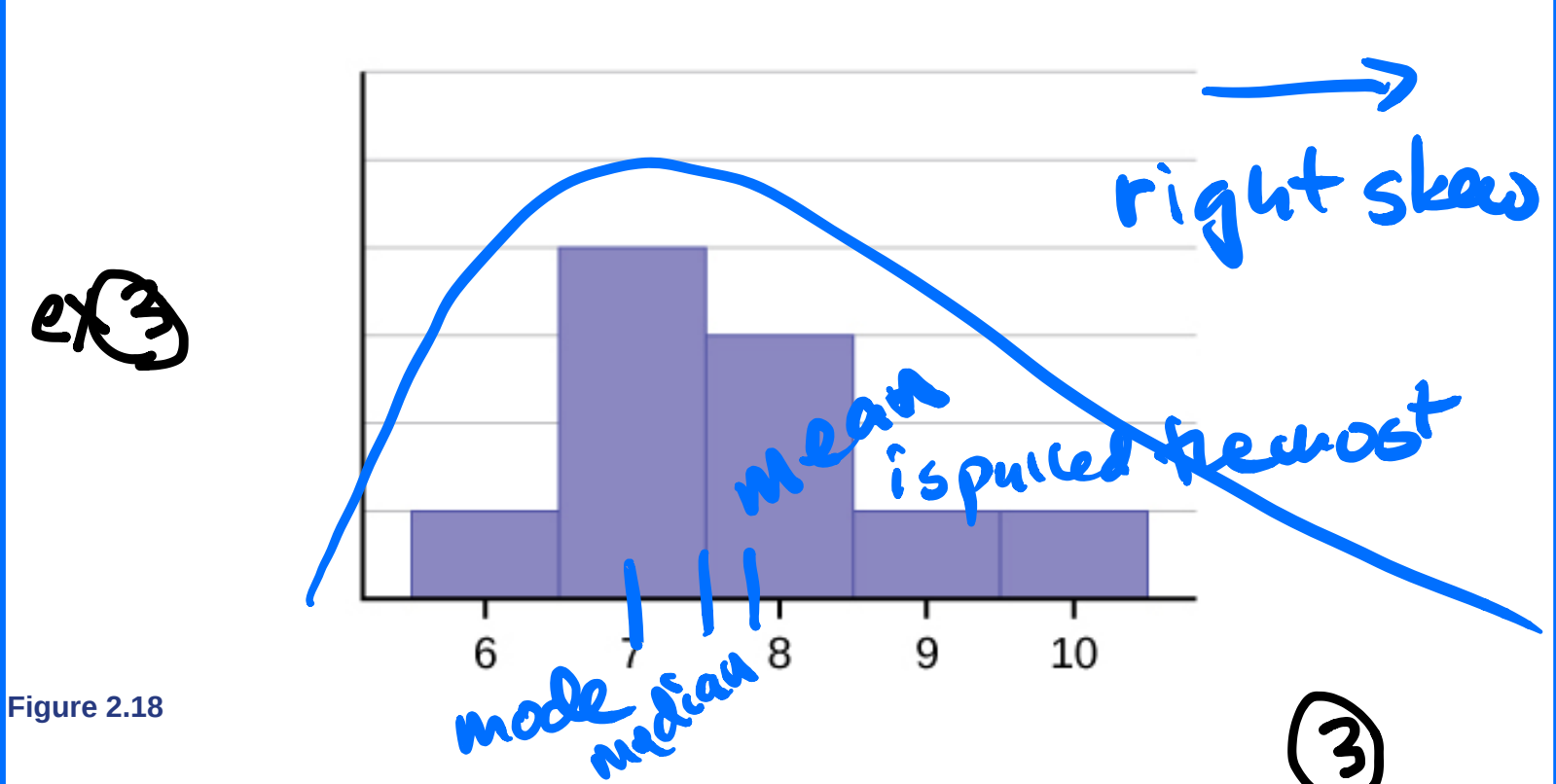


Figure 2.18

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

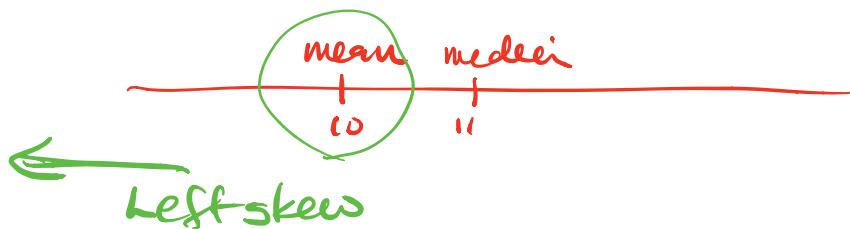
Example 2.31

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Exercise. Our data has mean $\bar{x} = 10$
and median = 11.

Is the data most likely
left or right skewed?

Left skewed because skew
pulls mean more than ^{it pulls} median



2.7 | Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

measures how spread out our data is

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. the average wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes; at supermarket B the standard deviation for the wait time is four minutes.

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket A. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket A, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because $5 + (1)(2) = 7$.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because $5 + (-2)(2) = 1$.

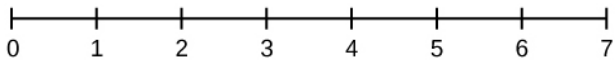


Figure 2.24

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer

Ways of measuring spread of data:

- standard deviation (we cover below)
- range

Example of range

Data: { 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, }

$$\begin{aligned}\text{range} &= \text{largest} - \text{smallest} \\ &= 10 - 6 \\ &= \boxed{4}\end{aligned}$$


Range only relies on 2 numbers
(min datapoint & max).

So they invented a measure
of spread using all the data,

a measure called

standard deviation (s)

NOTE

 In practice, **USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION**. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 2.32

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s .

Data	Freq.	Deviations	<i>Deviations</i> ²	(Freq.)(<i>Deviations</i> ²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times 0.275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times 0.000625 = 0.0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times 0.225625 = 1.35375$
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times 0.950625 = 2.851875$
				The total is 9.7375

Table 2.29

The sample variance, s^2 , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one ($20 - 1$):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891, \text{ which is rounded to two decimal places, } s = 0.72.$$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation on a calculator or computer.
- For a sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$

p112 Standard deviation is "average square deviation"

$x * f$	Data x	Freq. f	Deviations $(x - \bar{x})$	Deviations ² $(x - \bar{x})^2$	(Freq.)(Deviations ²) $(f)(x - \bar{x})^2$
$9 \times 1 = 9$	9	1	$9 - 10.5 = -1.5$	$(-1.5)^2 = 2.25$	$1 * 2.25 = 2.25$
$9.5 \times 2 = 19$	9.5	2	$9.5 - 10.5 = -1$	$(-1)^2 = 1$	$2 * 1 = 2$
$10 \times 4 = 40$	10	4	$10 - 10.5 = -0.5$	$(-0.5)^2 = 0.25$	$4 * 0.25 = 1$
42	10.5	4	$10.5 - 10.5 = 0$	$(0)^2 = 0$	$4 * 0 = 0$
66	11	6	$11 - 10.5 = 0.5$	$(0.5)^2 = 0.25$	$6 * 0.25 = 1.5$
34.5	11.5	3	$11.5 - 10.5 = 1$	$(1)^2 = 1$	$3 * 1 = 3$
$\frac{210.5}{20}$		20			

Table 2.29

sum of square deviations = 9.75

$$\bar{x} = 210.5 / 20$$

$$= 10.525$$

$$\approx \boxed{10.5}$$

$$s^2 = \text{sample variance} = \frac{9.75}{n-1} = \frac{9.75}{19} = 0.51$$

$$s = \text{sample standard deviation} = \sqrt{s^2} = \sqrt{0.51} = \boxed{0.72}$$

The units of this final answer s will be the same as whatever the units of the given original data was