

# 2 | DESCRIPTIVE STATISTICS



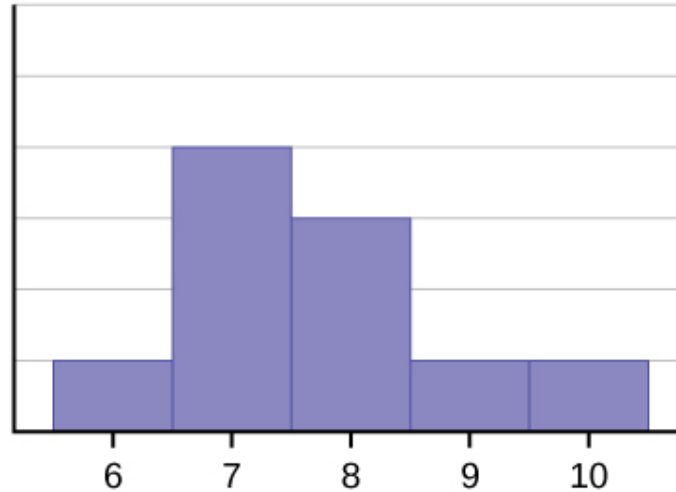
**Figure 2.1** When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

## Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "**Descriptive Statistics.**" You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.



**Figure 2.13**

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.

As with the mean, median and mode, and as we will see shortly, the variance, there are mathematical formulas that give us precise measures of these characteristics of the distribution of the data. Again looking at the formula for skewness we see that this is a relationship between the mean of the data and the individual observations cubed.

$$a_3 = \frac{\sum (x_i - \bar{x})^3}{ns^3}$$

where  $s$  is the sample standard deviation of the data,  $X_i$ , and  $\bar{x}$  is the arithmetic mean and  $n$  is the sample size.

Formally the arithmetic mean is known as the first moment of the distribution. The second moment we will see is the variance, and skewness is the third moment. The variance measures the squared differences of the data from the mean and skewness measures the cubed differences of the data from the mean. While a variance can never be a negative number, the measure of skewness can and this is how we determine if the data are skewed right or left. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean and standard deviation are *dimensional* quantities (this is why we will take the square root of the variance) that is, have the same units as the measured quantities  $X_i$ , the skewness is conventionally defined in such a way as to make it *nondimensional*. It is a pure number that characterizes only the shape of the distribution. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive  $X$  and a negative value signifies a distribution whose tail extends out towards more negative  $X$ . A zero measure of skewness will indicate a symmetrical distribution.

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

## 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.

## 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.

### Calculating the Standard Deviation

If  $x$  is a number, then the difference " $x$  minus the mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is  $x - \mu$ . For sample data, in symbols a deviation is  $x - \bar{x}$ .

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter  $s$  represents the sample standard deviation and the Greek letter  $\sigma$  (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then  $s$  should be a good estimate of  $\sigma$ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the  $x - \bar{x}$  values for a sample, or the  $x - \mu$  values for a population). The symbol  $\sigma^2$  represents the population variance; the population standard deviation  $\sigma$  is the square root of the population variance. The symbol  $s^2$  represents the sample variance; the sample standard deviation  $s$  is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations. Formally, the variance is the second moment of the distribution or the first moment around the mean. Remember that the mean is the first moment of the distribution.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by  $N$ , the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by  $n - 1$ , one less than the number of items in the sample.

### Formulas for the Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \text{ or } s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \text{ or } s = \sqrt{\frac{\left(\sum_{i=1}^n x_i^2\right) - n\bar{x}^2}{n - 1}}$$

- For the sample standard deviation, the denominator is  $n - 1$ , that is the sample size minus 1.

### Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \text{ or } \sigma = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2}$$

- For the population standard deviation, the denominator is  $N$ , the number of items in the population.

In these formulas,  $f$  represents the frequency with which a value appears. For example, if a value appears once,  $f$  is one. If a value appears three times in the data set or population,  $f$  is three. Two important observations concerning the variance and standard deviation: the deviations are measured from the mean and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data, what is the "normal" or most usual value of the observation. Later we will be trying to measure the "unusualness" of an observation or a sample mean and thus we need a measure from the mean. The second observation is that the deviations are squared. This does two things, first it makes the deviations all positive and second it changes the units of measurement from that

of the mean and the original observations. If the data are weights then the mean is measured in pounds, but the variance is measured in pounds-squared. One reason to use the standard deviation is to return to the original units of measurement by taking the square root of the variance. Further, when the deviations are squared it explodes their value. For example, a deviation of 10 from the mean when squared is 100, but a deviation of 100 from the mean is 10,000. What this does is place great weight on outliers when calculating the variance.

### Types of Variability in Samples

When trying to study a population, a sample is often used, either for convenience or because it is not possible to access the entire population. Variability is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability
- Induced variability
- Sample variability

Here are some examples to describe each type of variability.

#### Example 1: Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers: For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

#### Example 2: Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

#### Example 3: Induced variability

Induced variability is the counterpart to natural variability; this occurs because we have artificially induced an element of variation (that, by definition, was not present naturally): For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

#### Example 4: Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

### Example 2.29

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> <sup>2</sup>	(Freq.)( <i>Deviations</i> <sup>2</sup> )
$x$	$f$	$(x - \bar{x})$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
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 2.28

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:

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

### Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is  $-1.525$  for the data value nine. **If you add the deviations, the sum is always zero.** (For **Example 2.29**, there are  $n = 20$  deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation. By squaring the deviations we are placing an extreme penalty on observations that are far from the mean; these observations get greater weight in the calculations of the variance. We will see later on that the variance (standard deviation) plays the critical role in determining our conclusions in inferential statistics. We can begin now by using the standard deviation as a measure of "unusualness." "How did you do on the test?" "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by  $n = 20$ , the calculation divided by  $n - 1 = 20 - 1 = 19$  because the data is a sample. For the **sample** variance, we divide by the sample size minus one ( $n - 1$ ). Why not divide by  $n$ ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** This estimate requires us to use an estimate of the population mean rather than the actual population mean. Based on the theoretical mathematics that lies behind these calculations, dividing by  $(n - 1)$  gives a better estimate of the population variance.

The standard deviation,  $s$  or  $\sigma$ , is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make  $s$  or  $\sigma$  very large.

## Example 2.30

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- b. Calculate the following to one decimal place:
  - i. The sample mean
  - ii. The sample standard deviation
  - iii. The median
  - iv. The first quartile
  - v. The third quartile
  - vi. *IQR*

### Solution 2.30

- a. See **Table 2.29**
- b.
  - i. The sample mean = 73.5
  - ii. The sample standard deviation = 17.9
  - iii. The median = 73
  - iv. The first quartile = 61
  - v. The third quartile = 90
  - vi.  $IQR = 90 - 61 = 29$

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

**Table 2.29**

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1? ANSWER: Rounding)

Table 2.29

## Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of

the measures of center by finding the mean of the grouped data with the formula:  $\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f}$

where  $f$  = interval frequencies and  $m$  = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

### Example 2.31

Find the standard deviation for the data in **Table 2.30**.

Class	Frequency, $f$	Midpoint, $m$	$f * m$	$f(m - \bar{x})^2$
0-2	1	1	$1 * 1 = 1$	$1(1 - 7.58)^2 = 43.26$
3-5	6	4	$6 * 4 = 24$	$6(4 - 7.58)^2 = 76.77$
6-8	10	7	$10 * 7 = 70$	$10(7 - 7.58)^2 = 3.33$
9-11	7	10	$7 * 10 = 70$	$7(10 - 7.58)^2 = 41.10$
12-14	0	13	$0 * 13 = 0$	$0(13 - 7.58)^2 = 0$
	26=n		$\bar{x} = \frac{197}{26} = 7.58$	$s^2 = \frac{306.35}{26 - 1} = 12.25$

Table 2.30

For this data set, we have the mean,  $\bar{x} = 7.58$  and the standard deviation,  $s_x = 3.5$ . This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since  $7.58 - 3.5 =$

0.58. While the formula for calculating the standard deviation is not complicated,  $s_x = \sqrt{\frac{\sum(m - \bar{x})^2 f}{n - 1}}$  where

$s_x$  = sample standard deviation,  $\bar{x}$  = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

## Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value  $x$ , calculate how many standard deviations away from its mean the value is.
- Use the formula:  $x = \text{mean} + (\text{\#ofSTDEVs})(\text{standard deviation})$ ; solve for #ofSTDEVs.
- $\text{\# of STDEVs} = \frac{x - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol  $z$ . In symbols, the formulas become:

Sample	$x = \bar{x} + zS$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

**Table 2.31**



### Example 2.32

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Table 2.32

#### Solution 2.32

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$z = \# \text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

$$\text{For John, } z = \# \text{ of STDEVs} = \frac{2.85 - 3.0}{0.7} = -0.21$$

$$\text{For Ali, } z = \# \text{ of STDEVs} = \frac{77 - 80}{10} = -0.3$$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of  $-0.21$  is higher than Ali's z-score of  $-0.3$ . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

### Try It $\Sigma$

**2.32** Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Table 2.33

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a Normal Distribution, which we will examine in great detail later:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

## Coefficient of Variation

Another useful way to compare distributions besides simple comparisons of means or standard deviations is to adjust for differences in the scale of the data being measured. Quite simply, a large variation in data with a large mean is different than the same variation in data with a small mean. To adjust for the scale of the underlying data the Coefficient of Variation (CV) has been developed. Mathematically:

$$CV = \frac{s}{\bar{x}} * 100 \text{ conditioned upon } \bar{x} \neq 0, \text{ where } s \text{ is the standard deviation of the data and } \bar{x} \text{ is the mean.}$$

We can see that this measures the variability of the underlying data as a percentage of the mean value; the center weight of the data set. This measure is useful in comparing risk where an adjustment is warranted because of differences in scale of two data sets. In effect, the scale is changed to common scale, percentage differences, and allows direct comparison of the two or more magnitudes of variation of different data sets.