3.4: Mean

The mean is often what people are referring to when they say “average,” though this is not the only average that can be used (as we have seen, the mode and median are also averages). The **mean** is found by adding all the scores in a dataset and dividing that total by the number of cases in the dataset. The mean is balanced such that the total of how far people were below the mean is equal to the total of how far people were above the mean. This is a very important feature of the mean.

Statistical Notation

Statistical notation refers to the various symbols which are used to represent concepts in formulas and when reporting results. The formulas for calculating the mean are shown below. One formula is for a sample and the other is for a population. The calculations are the same but we use different symbols to clarify whether our data were drawn from a sample (i.e. a subset, or part, of a population) or the population (i.e. all cases). When using a formula to find a sample mean, $\bar{x}$ is used. This symbol is often referred to as “X bar.” However, *M* can also be used as the symbol for the mean. When using a formula to find a population mean, μ is used; this is the lowercase Greek letter whose name is mu. The other difference between the formulas is that we use a lowercase *n* for a sample to show that it is not the whole group (i.e. it is smaller than could be possible if every case was included) and capital *N* for the population to show that it includes all cases. Though the symbols are different, the steps of the sample and population mean formulas are the same.

Formulas for the Mean

Sample Mean	Population Mean
$\bar{x} = \frac{\Sigma x}{n}$	$\mu = \frac{\Sigma x}{N}$

Calculating the Mean

The steps for calculating a mean are as follows:

1. Add up all of the raw scores to solve for Σx
2. Divide Σx by the sample size.

That is it! There are only two steps and they must be done in the order specified.

Try using the formula for the mean with the age dataset from Data Set 3.1. Follow the two steps:

1. **Add up all the raw scores**

$$\Sigma x = 14+16+18+19+19+20+23+25+27+28+29+29+29+32+33+33+34+36+39+42 \\ +46+47 = 636$$

$$\Sigma x = 638$$

2. **Divide Σx by the sample size**

$$\sum \boldsymbol{x} = 638 \text{ and } n = 22$$

$$\bar{x} = 638/22 = 29.00$$

M = 29.00

Let’s practice the formula again using Data Set 3.3. This time we will show all the data in the formula at once rather than broken out into steps:

Mean Calculations Using Data Set 3.3

$$\bar{x} = \frac{64+65+65+68+72+75+76+77+77+81}{10} = \frac{720}{10} = 72.00$$

When we calculated the means for Data Set 3.1 and Data Set 3.3, the results were whole numbers (though we should still report them to the hundredths place). However, this will not always be the case. Let’s demonstrate this by finding the mean for Data Set 3.4.

Mean Calculations Using Data Set 3.4

Data Set 3.4

Height in Inches
74
72
71
68
66
65
65
64
63
62
61
59

$$\bar{x} = \frac{59+61+62+63+64+65+65+66+68+71+72+74}{12} = \frac{790}{12} = 65.8333\overline{3}$$

Rounding and Symbol Use Guidelines

The mean for height in Data Set 3.4 has a repeating decimal (also known as a recurring decimal). Rational numbers with repeating decimals (like the mean for height) and irrational numbers continue infinitely. Generally, however, infinite specificity is beyond what is needed when reporting results. Without conventions, statisticians could choose to round and show this mean a variety of ways which would make comparisons of data from different

reports harder to compare. Some fields, therefore, have developed rules and guidelines for the place to which values should be rounded and shown. These guidelines help ensure consistency and comparability of results across studies. Psychology, for example, follows the guidelines set forth in the American Psychological Association (APA) style manual. APA guidelines state that most values found through a process or formula (which are often referred to as *obtained values*) should be rounded and reported to the hundredths place. An exception is made when data were measured using an interval scale, for which, rounding to the tens place is acceptable under some conditions (APA, 2022). This meets the two objectives of summarizing by providing some specificity (two decimal places) while gaining simplicity (by not having a number continue to trail infinitely).

Further, various symbols are used for the mean depending upon the situation. When reporting means in sentences or tables, APA guidelines state that the symbol *M* be used in place of $\bar{x}$ or μ. M is the uppercase version of the lowercase Greek letter μ and is also the first letter of the word “mean.” *M* is much easier to use when typing because it is on a Standard English keyboard and does not require inserting a special symbol into a document. For these reasons, it was a logical symbol to use for abbreviating the word “mean.”

If the APA guidelines are followed, the obtained value for Data Set 3.4 could be written as such:

$$M = \mathbf{65.83}$$

This guideline for rounding only applies to the final result, not the raw data or steps of the formula. Keep in mind that when we round numbers we are losing some specificity and can be introducing a little bit of error in the process. Thus, the fewer decimal places you show when rounding, the greater the potential rounding error can be. The sooner this error is introduced in the computations, the greater its potential to impact the final result. Therefore, though final results are often rounded to the hundredths place, if a value is being used as a step in another calculation or formula it is best to use the exact, unrounded value or to keep steps to at least four decimal places if rounding.

Comparing the Mean to the Mode and Median

The mean, median, and mode are three options for finding and reporting central tendencies. Though all three can be used to summarize data for quantitative variables, only one is typically used at a time. There are two main reasons for this. First, reporting all three goes against one of the goals of summarizing which is to simplify. Second, the mean, median, and mode are often expected to yield similar results, for reasons we will review in detail in Chapter 5. Therefore, reporting all three measures of central tendency will yield redundant summaries in some cases. Third, when the mean, median, and mode do differ, it can be due to a relevant feature of the variable which is better captured or dealt with using one of the three measures of central tendency. Thus, it is

necessary to understand the differences which underlie the three measures to know which one is the best fit for each data set or variable.

Balancing Deviations

The mean is the point at which deviations are balanced. **Deviation** refers to the difference between the mean and any of the raw scores**.** The formula for the mean produces a number that is above some raw scores and below others. This means that some raw scores will have a positive deviation because they are greater in value than the mean while others will have a negative deviation because they are lesser in value than the mean. The deviation balances positive and negative deviations such that if all the deviations are added the sum will be 0. This is what is meant when we say the mean balances deviations.

Look back at Data Set 3.1 keeping in mind that the mean was 29.00. Notice that several of the raw scores for age such as 16 and 20 are below the mean, and others such as 32 and 44 are above the mean. If you find the deviations for all of the raw scores below the mean and then add them up, you will get the same absolute value as when you find all the deviations for the raw scores above the mean and add them up. Therefore, the mean is the point (or value) that balances between the total deviations below it and the total deviations above it.

Let's look at the balance point property of the mean using our data for age (Data Set 3.1). We can demonstrate the way the mean functions as a balance point by subtracting the mean from each raw score to find the deviations (see Table 1). The first column of Table 1 lists all of the raw scores. The second column lists the mean for the dataset. The third column lists the deviation for each raw score. When you sum the deviations, the result is 0. This is because the deviations below the mean balance with the deviations above the mean and will always sum to 0. This is what is being stated when we refer to the mean as a balance point; the mean is a number that causes the sum of the deviations to balance out to 0.

Formula for Deviation

The deviation for an individual score is calculate as:

$\text{dev} = x - \bar{x}$

x refers to an individual score and $\bar{x}$ refers to the mean.

The total deviation for a dataset is calculated as:

$\text{Sum of dev } = \sum(x - \bar{x})$

$\sum$ indicates that all deviations should be added.

Table 1 Deviations for Data Set 3.1

Raw Score	Mean	Deviation*
47	29	18
46	29	17
42	29	13
39	29	10
36	29	7
34	29	5
33	29	4
33	29	4
32	29	3
29	29	0
29	29	0
29	29	0

Raw Score	Mean	Deviation*
28	29	-1
27	29	-2
25	29	-4
23	29	-6
20	29	-9
19	29	-10
19	29	-10
18	29	-11
16	29	-13
14	29	-15
		Sum of Deviations = 0

*Note

Raw scores are shown to the whole number to reflect how they appeared in the data set and the steps are also shown to the whole number because all decimal places were zeros.

The mean and median are both balance points but focus on balancing two different things. The mean balances deviations, whereas the median balances the number of cases. Take a look at the data. We see that 12 scores were above the mean and 10 scores were below the mean. Thus, unlike the median, the mean doesn't always balance the number of cases that are above or below the mean; instead, the mean refers to how far those scores are from the mean. In contrast there is a balance of cases from the median, with 11 cases above the median and 11 cases below the median.

Note that the median point represents the point between the bottom two scores of 29 causing one of the 29s to represent a value below 29 and three to represent values above 29. This is because the real score limits (or boundaries of what rounds to 29) are from 28.50 to 29.50 and, thus, the 29s are considered to represent various places in this range. Thus, sometimes a different version of the median known as the interpolated median (also known as the precise median) are used to better represent this. However, the standard median is used more often in the behavioral sciences and, thus, is the focus of this chapter. Those interested in the way a median serves as a balance of cases when the same value appears on each side of the middle value (like we see in Data Set 3.1) are encouraged to review the interpolated median online or, if you have a professor like Dr. Peter who loves explaining how numbers are conceptualized when using medians, ask your professor during their office hours. If they have read and recommended this book to you, they should see this question coming.

The Impact of Outliers and Sample Size on the Mean. Because the mean balances deviations, it is sensitive to lack of symmetry in data or something known as an outlier. Outliers are rare (meaning infrequently occurring) scores that are more extreme than the other scores in a dataset. When an outlier is present, the mean has to shift towards the outlier to balance the deviations. This shift is called *skew*. Let's compare two data sets to see an example of how an outlier shifts data. Data Set 3.5 includes data for the variable Household Income from a small sample of 10 cases. The incomes range from $47,000 to $112,000. Though the range of incomes is fairly

large, none of the incomes reported are extremely different than the next closest score or the data set overall. The mean of income for Data Set 3.5 is $80,000. The scores vary with some incomes being close to the mean and others farther from the mean.

Data Set 3.5 Household Income in Dollars (n = 10)

Raw Score	Mean	Deviation
112,000	80,000	32,000
103,000	80,000	23,000

91,000	80,000	11,000
87,000	80,000	7,000
80,000	80,000	0
80,000	80,000	0
75,000	80,000	-5,000
66,000	80,000	-14,000
59,000	80,000	-21,000
47,000	80,000	-33,000
		Sum of Deviations = 0

Now let's look at Data Set 3.6. Data Set 3.6 includes all 10 scores for the variable Household Income as Data Set 3.5 but with the addition of an 11th case with an income of $1,950,000. The score of $1,950,000 is an outlier. Let's take a look at what happens to the statistics for the data when this outlier is present. The incomes in Data Set 3.6 range from $31,000 to $1,950,000. The mean of income for Data Set 3.6 is $250,000, which is much higher than the mean was when the outlier was not in the data (as shown in Data Set 3.5). This is because the mean had to move closer to the outlier to allow the deviations to balance out to zero. The deviations for all raw scores in data set 3.6 are also very large. This is because the outlier has dramatically impacted the mean.

Data Set 3.6 Household Income in Dollars (n = 10)

Raw Score	Mean	Deviation
1,950,000	250,000	1,700,000
112,000	250,000	-138,000
103,000	250,000	-147,000
91,000	250,000	-159,000
87,000	250,000	-163,000
80,000	250,000	-170,000
80,000	250,000	-170,000
75,000	250,000	-175,000
66,000	250,000	-184,000
59,000	250,000	-191,000
47,000	250,000	-203,000
		Sum of Deviations = 0

The name for this kind of impact is skew. **Skew** refers to the asymmetry in quantitative data which can have a greater impact on the mean than the median or the mode. The mean can be greatly impacted by an outlier because outliers pull the mean toward them in order to balance deviations. The measures of central tendency for data sets 3.5 and 3.6 are summarized in Table 2. Compare the mean for the sample without an outlier (Data Set 3.5) which was $80,000 to the mean for the sample with an outlier (Data Set 3.6) which was $250,000. The outlier is a very high score and when it was added to the data set, it caused the mean to increase. This is known as positive skew. **Positive skew** refers to an increase in the mean due to asymmetry in data caused by a high score or high scores. The mean is pulled up toward the high outlier. Consistent with this, a low outlier will cause negative skew. **Negative skew** refers to a decrease in the mean due to asymmetry in data caused by a low score or low scores.

The median, by comparison, is less likely to be dramatically impacted by outliers and asymmetry because it balances cases rather than deviations. This allows the median to experience more stability when an outlier is added to a data set. To illustrate this, let's

compare the medians for the data when with (Data Set 3.6) and without the outlier present (Data Set 3.5). The medians are the same for Data Set 3.5 and Data Set 3.6 (see Table 2). Thus, we can see that, unlike the mean, the median is not very sensitive to the impact of an outlier.

The presence of skew can be identified by comparing means to medians. When the mean and median are equal, it indicates that the mean has not been skewed. When the mean is higher than the median, it indicates that there is positive skew. This is because the mean gets pulled in the direction of an outlier more so than the median is pulled. For this same reason, when a mean is lower than a median, it indicates that there is negative skew in the data set.

Table 2 The Impact of an Outlier on Measures of Central Tendency

	Without Outlier	With Outlier
Mode	80,000	80,000
Median	80,000	82,000
Mean	80,000	250,000

The Role of Sample Size

Sample size can mitigate or aggravate the impact of an outlier. The impact of an outlier is based on two things:

1. How extreme the outlier is relative to the rest of the scores and
2. Sample size.

The more extreme the outlier is, the greater its impact can be. However, the impact of an outlier is proportional to how much of the sample is comprised of the outlier. When the sample size is small, there are fewer values to balance against an outlier which allows the outlier to have a greater impact. You can see this demonstrated in Data Set 3.6 which includes an extreme outlier in a small data set. The only score above the mean is the outlier and all other scores had to balance it out by being below the mean. When the sample size is large, there are more values to balance against the outlier which mitigates its impact. Thus, the outlier in Data Set 3.6 would have more ability to skew the mean in a sample of 11 cases where it represents one eleventh of the observations than in a sample of 100 cases where it represents only one-hundredth of the observations. You can summarize it this way: When the sample size is small, the outlier represents a larger proportion of the scores and, thus, has more power to move the mean; however, when the sample size is large, the outlier represents a smaller proportion of he scores and, thus, has less power to move the mean.

Reading Review 3.2

1. What is the sample size for the test score data?
2. What is the median for the test score data?
3. What is the mean for the test score data?
4. Does there appear to be any positive or negative skew in the test score data?

Test Scores	
96	82
94	80
91	79
90	73
89	71
88	65
85	52
84	47

Choosing between the Mode, Median, and Mean. There are several things a statistician should consider when deciding which measure of central tendency to use to summarize a variable. Here, we will review a few of the main considerations.

A statistician must ensure that the measure of central tendency is appropriate for the way a variable was measured. If data are qualitative, the mode is the only option. The mode is also often considered appropriate for ordinal data because of quantitative limitations with this scale of measurement. The mode is particularly useful when a score or qualitative response repeats several times in a data set. If there is no mode, it is generally best to choose the median or the mean for interval- or ratio-level data unless the goal of the summary is to indicate that there were no repeating scores or qualitative responses.

If the variable is measured on the interval or ratio scale, the mode, median, or mean could all be appropriate so additional factors should be considered. The statistician should consider which measure of central tendency will provide the most appropriate information for their current goals. Thus, the mean is most commonly chosen for quantitative data because it can be used to serve many different goals. This is because the mean is a foundational descriptive statistic on which more advanced, inferential statistical procedures such as t-tests (which are covered in Chapters 7 through 9) and ANOVAs (which are covered in Chapters 10 and 11) rely. However, when there is problematic skew, the median is generally recommended and used instead of the mean.

Measures of Central Tendency Compared to Real Scores. The mode has one benefit that the median and mean do not. When there is a mode, it will always be the same as at least two observed scores or qualitative responses. The mode for a sample cannot be something that was not observed in the sample data. Keep in mind that measures of central tendency are meant to summarize the data. Another way of thinking about a summary is that it is meant to describe what was generally true. However, only the mode is guaranteed to provide a summary that was true for at least two cases.

In contrast, the median and the mean for sample data can be values that were not true of any observed cases. Take the median for Data Set 3.3 which was 73.50. We can summarize those data by saying the median speed was 73.50. However, there isn't a single case where the speed was exactly 73.50 miles per hour. Take a look at

the mean for Data Set 3.5. Therefore, it is important to keep in mind what each summary is and is not able to tell us. The measures of central tendency are not designed to tell what is always true. They can't because they are built for summarizing things that vary not things which are constant. The median and mean aren't even designed to tell us what was at least sometimes true (though they do sometimes yield values that occurred in the data set). Instead, they are each designed to focus on a different way of stating what was approximately (or tended to be) true. This can lead to some confusion when we read statements such as "The average household has 2.50 children" because no household can actually have 2.50 children as you cannot have a fraction of a person. What is meant when an "average" is stated in this way is a theoretical and approximate representation of what tended to be true. Note also that the word "average" in the sentence is vague as the mode, median, and mean are all averages. We know that the average being referred to is not the mode because the mode can only be a value or qualitative response which can and did occur. That means that the average being used to summarize number of children is either the median or the mean, but we cannot be sure which. This is why it is always important to specify when reporting results.
