STATISTICS for People Who (*Think They*) HATE STATISTICS

3RD EDITION

NEIL J. SALKIND

University of Kansas

SAGE Publications
Los Angeles • London • New Delhi • Singapore

2 Means to an End

Computing and Understanding Averages

Difficulty Scale ☺☺☺☺ (moderately easy)

What you'll learn about in this chapter

- Understanding measures of central tendency
- Computing the mean for a set of scores
- Computing the mode and the median for a set of scores
- Selecting a measure of central tendency

You've been very patient, and now it's finally time to get started working with some real, live data. That's exactly what you'll do in this chapter. Once data are collected, a usual first step is to organize the information using simple indexes to describe the data. The easiest way to do this is through computing an average, of which there are several different types.

An **average** is the one value that best represents an entire group of scores. It doesn't matter whether the group of scores is the number correct on a spelling test for 30 fifth graders or the batting percentage of each of the New York Yankees or the number of people who registered as Democrats or Republicans in the most recent election. In all of these examples, groups of data can be summarized using an average. Averages, also called **measures of central tendency**, come in three flavors: the mean, the median, and the mode. Each provides you with a different type of information about a distribution of scores and is simple to compute and interpret.

COMPUTING THE MEAN

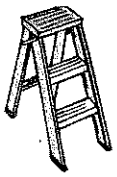
The **mean** is the most common type of average that is computed. It is simply the sum of all the values in a group, divided by the number of values in that group. So, if you had the spelling scores for 30 fifth graders, you would simply add up all the scores and get a total, and then divide by the number of students, which is 30.

The formula for computing the mean is shown in Formula 2.1.

$$\bar{X} = \frac{\Sigma X}{n} \tag{2.1}$$

where

- The letter X with a line above it (also sometimes called "X bar") is the mean value of the group of scores or the mean.
- The S, or the Greek letter sigma, is the summation sign, which tells you to add together whatever follows it.
- The X is each individual score in the group of scores.
- Finally, the n is the size of the sample from which you are computing the mean.



To compute the mean, follow these steps:

1. List the entire set of values in one or more columns. These are all the *X*s.
2. Compute the sum or total of all the values.
3. Divide the total or sum by the number of values.

For example, if you needed to compute the average number of shoppers at three different locations, you would compute a mean for that value.

Location	*Number of Annual Customers*
Lanham Park store	2,150
Williamsburg store	1,534
Downtown store	3,564

The mean or average number of shoppers in each store is 2,416. Formula 2.2 shows how it was computed using the formula you saw in Formula 2.1:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{2{,}150 + 1{,}534 + 3{,}564}{3} = \frac{7{,}248}{3} = 2{,}416 \qquad (2.2)$$

Or, if you needed to compute the average number of students in grades kindergarten through 6, you would follow the same procedure.

Grade	*Number of Students*
Kindergarten	18
1	21
2	24
3	23
4	22
5	24
6	25

The mean or average number of students in each class is 22.43. Formula 2.3 shows how it was computed using the formula you saw in Formula 2.1:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{18 + 21 + 24 + 23 + 22 + 24 + 25}{7} = 22.43 \qquad (2.3)$$

See, we told you it was easy. No big deal.

THINGS TO REMEMBER

The mean is sometimes represented by the letter M and is also called the typical, average, or most central score. If you are reading another statistics book or a research report, and you see something like $M = 45.87$, it probably means that the mean is equal to 45.87.

- In the formula, a small n represents the sample size for which the mean is being computed. A large N (like this) would represent the population size. In some books and in some journal articles, no distinction is made between the two.
- The sample mean is the measure of central tendency that most accurately reflects the population mean.
- The mean is like the fulcrum on a seesaw. It's the centermost point where all the values on one side of the mean are equal in weight to all the values on the other side of the mean.
- Finally, for better or worse, the mean is very sensitive to extreme scores. An extreme score can pull the mean in one or the other

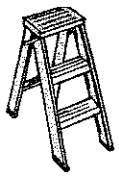
direction and make it less representative of the set of scores and less useful as a measure of central tendency. This, of course, all depends on the values for which the mean is being computed. More about this later.

The mean is also referred to as the **arithmetic mean**, and there are other types of means that you may read about, such as the harmonic mean. Those are used in special circumstances but need not concern you here. And if you want to be technical about it, the arithmetic mean (which is the one that we have discussed up to now) is also defined as the point at which the sum of the deviations is equal to zero (whew!). So, if you have scores like 3, 4, and 5 (where the mean is 4), the sum of the deviations about the mean (–1, 0, and +1) is 0.

Remember that the word *average* means only the one measure that best represents a set of scores, and that there are many different types of averages. Which type of average you use depends on the question that you are asking and the type of data that you are trying to summarize.

Computing a Weighted Mean

You've just seen an example of how to compute a simple mean. But there may be situations where you have the occurrence of more than one value and you want to compute a weighted mean. A weighted mean can be easily computed by multiplying the value by the frequency of its occurrence, adding the total of all the products and then dividing by the total number of occurrences.



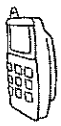
To compute a weighted mean, follow these steps:

1. List all the values in the sample for which the mean is being computed, such as those shown in the column labeled Value (the value of *X*) in the following table.
2. List the frequency with which each value occurs.
3. Multiply the value by the frequency, as shown in the third column.
4. Sum all the values in the Value × Frequency column.
5. Divide by the total frequency.

For example, here's a table that organizes the values and frequencies in a flying proficiency test for 100 airline pilots.

Value	*Frequency*	*Value × Frequency*
97	4	388
94	11	1,034
92	12	1,104
91	21	1,911
90	30	2,700
89	12	1,068
78	9	702
60 (don't fly with this guy)	1	60
Total	**100**	**8,967**

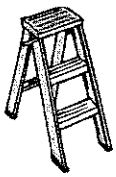
The weighted mean is 8,967/100, or 89.67. Computing the mean this way is much easier than entering 100 different scores into your calculator or computer program.



In basic statistics, an important distinction needs to be made between those values associated with samples (a part of a population) and those associated with populations. To do this, statisticians use the following conventions. For a sample statistic (such as the mean of a mean), Roman letters are used. For a population parameter (such as the mean of a population), Greek letters are used. So, the mean for the spelling score for a sample of 100 fifth graders is represented as $\overline{X}_s$, whereas the mean for the spelling score for the entire population of fifth graders is represented as μ_s, using the Greek letter mu, or μ.

COMPUTING THE MEDIAN

The median is also an average, but of a very different kind. The **median** is defined as the midpoint in a set of scores. It's the point at which one half, or 50%, of the scores fall above and one half, or 50%, fall below. It's got some special qualities that we will talk about later in this section, but for now, let's concentrate on how it is computed. There's no standard formula for computing the median.



To compute the median, follow these steps:

1. List the values in order, either from highest to lowest or lowest to highest.
2. Find the middle-most score. That's the median.

For example, here are the incomes from five different households:

$135,456
$25,500
$32,456
$54,365
$37,668

Here is the list ordered from highest to lowest:

$135,456
$54,365
$37,668
$32,456
$25,500

There are five values. The middle-most value is $37,668, and that's the median.

Now, what if the number of values is even? Let's add a value ($34,500) to the list so there are six income levels. Here they are.

$135,456
$54,365
$37,668
$34,500
$32,456
$25,500

When there is an even number of values, the median is simply the mean between the two middle values. In this case, the middle two cases are $34,500 and $37,668. The mean of those two values is $36,084. That's the median for that set of six values.

What if the two middle-most values are the same, such as in the following set of data?

$45,678
$25,567
$25,567
$13,234

Then the median is same as both of those middle-most values. In this case, it's $25,567.

If we had a series of values that was the number of days spent in rehabilitation for a sports-related injury for seven different patients, the numbers may look like this:

43
34
32
12
51
6
27

As we did before, we can order the values (51, 43, 34, 32, 27, 12, 6) and then select the middle value as the median, which in this case is 32. So, the median number of days spent in rehab is 32.

If you know about medians, you should know about **percentile points**. Percentile points are used to define the percentage of cases equal to and below a certain point in a distribution or set of scores. For example, if a score is "at the 75th percentile," it means that the score is at or above 75% of the other scores in the distribution. The median is also known as the 50th percentile, because it's the point below which 50% of the cases in the distribution fall. Other percentiles are useful as well, such as the 25th percentile, often called Q_1, and the 75th percentile, referred to as Q_3. So what's Q_2? The median, of course.

Here comes the answer to the question you've probably had in the back of your mind since we started talking about the median. Why use the median instead of the mean? For one very good reason. The median is insensitive to extreme scores, whereas the mean is not.

When you have a set of scores in which one or more scores are extreme, the median better represents the centermost value of that set of scores than any other measure of central tendency. Yes, even better than the mean.

What do we mean by extreme? It's probably easiest to think of an extreme score as one that is very different from the group to which it belongs. For example, consider the list of five incomes that we worked with earlier (shown again here):

$135,456
$54,365
$37,668
$32,456
$25,500

The value $135,456 is more different from the other five than any other value in the set. We would consider that an extreme score.

The best way to illustrate how useful the median is as a measure of central tendency is to compute both the mean and the median for a set of data that contains one or more extreme scores and then compare them to see which one best represents the group. Here goes.

The mean of the set of five scores you see above is the sum of the set of five divided by five, which turns out to be $57,089. On the other hand, the median for this set of five scores is $37,668. Which is more representative of the group? The value $37,668, because it clearly lies more in the middle of the group, and we like to think about the average as being representative or assuming a central position. In fact, the mean value of $57,089 falls above the fourth highest value ($54,365) and is not very central or representative of the distribution.

It's for this reason that certain social and economic indicators (mostly involving income) are reported using a median as a measure of central tendency, such as "The median income of the average American family is . . . ," rather than using the mean to summarize the values. There are just too many extreme scores that would **skew**, or significantly distort, what is actually a central point in the set or distribution of scores.

You learned earlier that sometimes the mean is represented by the capital letter *M* instead of $\bar{X}$. Well, other symbols are used for the median as well. We like the letter *M*, but some people confuse it with the mean, so they use Med for median, or Mdn. Don't let that throw you—just remember what the median is and what it represents, and you'll have no trouble adapting to different symbols.

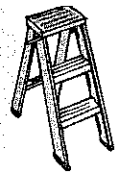
THINGS TO REMEMBER

Here are some interesting and important things to remember about the median.

- The mean is the middle point of a set of values, and the median is the middle point of a set of cases.
- Because the median cares about how many cases, and not the values of those cases, extreme scores (sometimes called **outliers**) don't count.

COMPUTING THE MODE

The third and last measure of central tendency that we'll cover, the mode, is the most general and least precise measure of central tendency, but it plays a very important part in understanding the characteristics of a special set of scores. The **mode** is the value that occurs most frequently. There is no formula for computing the mode.



To compute the mode, follow these steps:

1. List all the values in a distribution, but list each only once.
2. Tally the number of times that each value occurs.
3. The value that occurs most often is the mode.

For example, an examination of the political party affiliation of 300 people might result in the following distribution of scores.

Party Affiliation	*Number or Frequency*
Democrats	90
Republicans	70
Independents	140

The mode is the value that occurs most frequently, which in the above example is Independents. That's the mode for this distribution.

If we were looking at the modal response on a 100-item multiple-choice test, we might find that the A alternative was chosen more frequently than any other. The data might look like this.

Item Alternative Selected	*A*	*B*	*C*	*D*
Number of Times	57	20	12	11

On this 100-item multiple-choice test where each item has four choices (A, B, C, and D), A was the answer selected 57 times. It's the modal response.

Want to know what the easiest and most commonly made mistake is when computing the mode? It's selecting the number of times a category occurs, rather than the label of the category itself. Instead of the mode being Independents, it's easy for someone to conclude the mode is 140. Why? Because they are looking at the number of times the value occurred, and not the value that occurred most often! This is a simple mistake to make, so be on your toes when you are asked about these things.

Apple Pie à la Bimodal

If every value in a distribution contains the same number of occurrences, then there really isn't a mode. But if more than one value appears with equal frequency, the distribution is multimodal. The set of scores can be bimodal (with two modes), as the following set of data using hair color illustrates.

Hair Color	*Number or Frequency*
Red	7
Blond	12
Black	45
Brown	45

In the above example, the distribution is bimodal because the frequency of the values of black and brown hair occurs equally. You can even have a bimodal distribution when the modes are relatively close together, but not exactly the same, such as 45 people with black hair and 44 with brown hair. The question becomes, How much does one class of occurrences stand apart from another? Can you have a trimodal distribution? Sure—where three values have

the same frequency. It's unlikely, especially when you are dealing with a large set of **data points**, or observations, but certainly possible.

WHEN TO USE WHAT

OK, we've defined three different measures of central tendency and given you fairly clear examples of each. But the most important question remains unanswered. That is, "When do you use which measure?"

In general, which measure of central tendency you use depends on the type of data that you are describing. Unquestionably, a measure of central tendency for qualitative, categorical, or nominal data (such as racial group, eye color, income bracket, voting preference, and neighborhood location) can be described using only the mode.

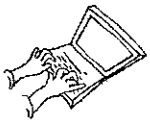
For example, you can't be looking at the most central measure that describes which political affiliation is most predominant in a group and use the mean—what in the world could you conclude, that everyone is half-Republican? Rather, that out of 300 people, almost half (140) are Independent seems to be the best way of describing the value of this variable. In general, the median and mean are best used with quantitative data, such as height, income level in dollars (not categories), age, test score, reaction, and number of hours completed for a degree.

It's also fair to say that the mean is a more precise measure than the median, and the median is a more precise measure than the mode. This means that all other things being equal, use the mean, and indeed, the mean is the most often used measure of central tendency. However, we do have occasions when the mean would not be appropriate as a measure of central tendency—for example, when we have categorical or nominal data, such as hair color. Then we use the mode. So, here is a set of three guidelines that may be of some help. And remember, there can always be exceptions.

1. Use the mode when the data are categorical in nature and values can fit into only one class, such as hair color, political affiliation, neighborhood location, and religion. When this is the case, these categories are called mutually exclusive.
2. Use the median when you have extreme scores and you don't want to distort the average (computed as the mean), such as when the variable of interest is income expressed in dollars.
3. Finally, use the mean when you have data that do not include extreme scores and are not categorical, such as the numerical score on a test or the number of seconds it takes to swim 50 yards.

USING THE COMPUTER AND COMPUTING DESCRIPTIVE STATISTICS

If you haven't already, now would be a good time to turn to Appendix A so you can become familiar with the basics of using SPSS. Then come back here.



Let's use SPSS to compute some descriptive statistics. The data set we are using is named Chapter 2 Data Set 1, which is a set of 20 scores on a test of prejudice. All of the data sets are available in Appendix C and from the Sage Internet site (www.sagepub.com/salkindstudy). There is one variable in this data set:

Variable	*Definition*
Prejudice	The value on a test of prejudice as measured on a scale from 1 to 100

Here are the steps to compute the measures of central tendency that we discussed in this chapter. Follow along and do it yourself. With this and all exercises, including data that you enter or download, we'll assume that the data set is already open in SPSS.

1. Click Analyze → Descriptive Statistics → Frequencies.
2. Double-click on the variable named Prejudice to move it to the Variable(s) box.
3. Click Statistics and you will see the Frequencies: Statistics dialog box shown in Figure 2.1.
4. Under Central Tendency, click the Mean, Median, and Mode boxes.
5. Click Continue.
6. Click OK.

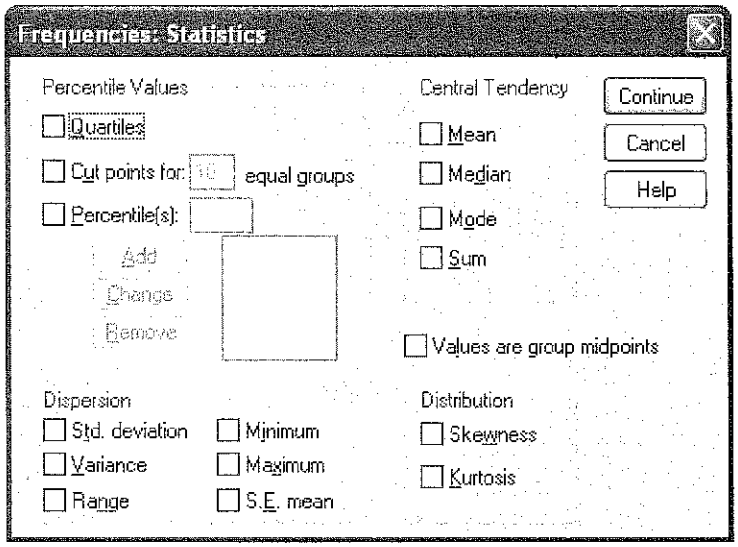


Figure 2.1 The Frequencies: Statistics Dialog Box From SPSS

The SPSS Output

Figure 2.2 shows you selected output from the SPSS procedure for the variable named Prejudice.

In the Statistics part of the output, you can see how the mean, median, and mode are all computed along with the sample size and the fact that there were no missing data. SPSS does not use symbols such as X in its output. Also listed in the output are the frequency of each value and the percentage of times it occurs, all useful descriptive information.

It's a bit strange, but if you select Analyze → Descriptive Statistics → Descriptives in SPSS and then click Options, there's no option to select the median or the mode, which you might expect because they are basic descriptive statistics. The lesson here? Statistical analysis programs are usually quite different from one another, use different names for the same things, and make different assumptions about what's where. If you can't find what you want, it's probably there. Just keep hunting. Also, be sure to use the Help feature to help navigate through all this new information until you find what you need.

Statistics

Prejudice

N	Valid	20
	Missing	0
Mean		84.70
Median		87.00
Mode		87

Prejudice

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	55	1	5.0	5.0	5.0
	64	1	5.0	5.0	10.0
	67	1	5.0	5.0	15.0
	76	1	5.0	5.0	20.0
	77	1	5.0	5.0	25.0
	81	2	10.0	10.0	35.0
	82	1	5.0	5.0	40.0
	87	4	20.0	20.0	60.0
	89	1	5.0	5.0	65.0
	93	1	5.0	5.0	70.0
	94	2	10.0	10.0	80.0
	96	1	5.0	5.0	85.0
	99	3	15.0	15.0	100.0
	Total	20	100.0	100.0	

Figure 2.2 Descriptive Statistics From SPSS

Summary

No matter how fancy schmancy your statistical techniques are, you will still almost always start by simply describing what's there—hence the importance of understanding the simple notion of central tendency. From here, we go to another important descriptive construct: variability, or how different scores are from one another.

Time to Practice

1. By hand, compute the mean, median, and mode for the following set of 40 reading scores.

31	32	43	42
24	34	25	44
23	43	24	36
25	41	23	28
14	21	24	17

25	23	44	21
13	26	23	32
12	26	14	42
14	31	52	12
23	42	32	34

2. Compute the mean, median, and mode for the following three sets of scores saved as Chapter 2 Data Set 2. Do it by hand or use a computer program such as SPSS. Show your work, and if you use SPSS, print out a copy of the output.

Score 1	*Score 2*	*Score 3*
3	34	154
7	54	167
5	17	132

Score 1	*Score 2*	*Score 3*
4	26	145
5	34	154
6	25	145
7	14	113
8	24	156
6	25	154
5	23	123

3. Compute the means for the following set of scores saved as Chapter 2 Data Set 3 using SPSS. Print out a copy of the output.

Hospital Size (number of beds)	*Infection Rate (per 1,000 admissions)*
234	1.7
214	2.4
165	3.1
436	5.6
432	4.9
342	5.3
276	5.6
187	1.2
512	3.3
553	4.1

4. You are the manager of a fast food store. Part of your job is to report to the boss at the end of each day which special is selling best. Use your vast knowledge of descriptive statistics and write one paragraph to let the boss know what happened today. Here are the data. Don't use SPSS to compute important values; rather, do it by hand. Be sure to include a copy of your work.

Chapter 12 Data Set 1

Group	Language Score	Group	Language Score
1	87	2	81
1	86	2	82
1	76	2	78
1	56	2	85
1	78	2	91
1	98	3	89
1	77	3	91
1	66	3	96
1	75	3	87
1	67	3	89
2	87	3	90
2	85	3	89
2	99	3	96
2	85	3	96
2	79	3	93

Chapter 12 Data Set 2

Pratice	Time	Pratice	Time
1	58.7	2	54.6
1	55.3	2	51.5
1	61.8	2	54.7
1	49.5	2	61.4
1	64.5	2	56.9
1	61.0	3	68.0
1	65.7	3	65.9
1	51.4	3	54.7
1	53.6	3	53.6
1	59.0	3	58.7
2	64.4	3	58.7
2	55.8	3	65.7
2	58.7	3	66.5
2	54.7	3	56.7
2	52.7	3	55.4
2	67.8	3	51.5
2	61.6	3	54.8
2	58.7	3	57.2

Chapter 13 Data Set 1

Treatment	Gender	Loss	Treatment	Gender	Loss
1	1	76	2	1	88
1	1	78	2	1	76
1	1	76	2	1	76
1	1	76	2	1	76
1	1	76	2	1	56
1	1	74	2	1	76
1	1	74	2	1	76
1	1	76	2	1	98
1	1	76	2	1	88
1	1	55	2	1	78
1	2	65	2	2	65
1	2	90	2	2	67
1	2	65	2	2	67
1	2	90	2	2	87
1	2	65	2	2	78
1	2	90	2	2	56
1	2	90	2	2	54
1	2	79	2	2	56
1	2	70	2	2	54
1	2	90	2	2	56

Chapter 13 Data Set 2

Severity	Treatment	Pain Score	Severity	Treatment	Pain Score
1	Drug #1	6	2	Drug #2	7
1	Drug #1	6	2	Drug #2	5
1	Drug #1	7	2	Drug #2	4
1	Drug #1	7	2	Drug #2	3
1	Drug #1	7	2	Drug #2	4
1	Drug #1	6	2	Drug #2	5
1	Drug #1	5	2	Drug #2	4
1	Drug #1	6	2	Drug #2	4
1	Drug #1	7	2	Drug #2	3
1	Drug #1	8	2	Drug #2	3
1	Drug #1	7	2	Drug #2	4
1	Drug #1	6	2	Drug #2	5
1	Drug #1	5	2	Drug #2	6
1	Drug #1	6	2	Drug #2	7

(Continued)

(Continued)

1	Drug #1	7	2	Drug #2	7
1	Drug #1	8	2	Drug #2	6
1	Drug #1	9	2	Drug #2	5
1	Drug #1	8	2	Drug #2	4
1	Drug #1	7	2	Drug #2	4
1	Drug #1	7	2	Drug #2	5
2	Drug #1	7	1	Placebo	2
2	Drug #1	8	1	Placebo	1
2	Drug #1	8	1	Placebo	3
2	Drug #1	9	1	Placebo	4
2	Drug #1	8	1	Placebo	5
2	Drug #1	7	1	Placebo	4
2	Drug #1	6	1	Placebo	3
2	Drug #1	6	1	Placebo	3
2	Drug #1	6	1	Placebo	3
2	Drug #1	7	1	Placebo	4
2	Drug #1	7	1	Placebo	5
2	Drug #1	6	1	Placebo	3
2	Drug #1	7	1	Placebo	1
2	Drug #1	8	1	Placebo	2
2	Drug #1	8	1	Placebo	4
2	Drug #1	8	1	Placebo	3
2	Drug #1	9	1	Placebo	5
2	Drug #1	0	1	Placebo	4
2	Drug #1	9	1	Placebo	2
2	Drug #1	8	1	Placebo	3
1	Drug #2	6	2	Placebo	4
1	Drug #2	5	2	Placebo	5
1	Drug #2	4	2	Placebo	6
1	Drug #2	5	2	Placebo	5
1	Drug #2	4	2	Placebo	4
1	Drug #2	3	2	Placebo	4
1	Drug #2	3	2	Placebo	6
1	Drug #2	3	2	Placebo	5
1	Drug #2	4	2	Placebo	4
1	Drug #2	5	2	Placebo	2
1	Drug #2	5	2	Placebo	1
1	Drug #2	5	2	Placebo	3
1	Drug #2	6	2	Placebo	2
1	Drug #2	6	2	Placebo	2
1	Drug #2	7	2	Placebo	3
1	Drug #2	6	2	Placebo	4
1	Drug #2	5	2	Placebo	3
1	Drug #2	7	2	Placebo	2
1	Drug #2	6	2	Placebo	2
1	Drug #2	8	2	Placebo	1

Chapter 14 Data Set 1

Quality of Marriage	*Quality Parent-Child*	*Quality of Marriage*	*Quality Parent-Child*
1	58.7	2	54.6
1	55.3	2	51.5
1	61.8	2	54.7
1	49.5	2	61.4
1	64.5	2	56.9
1	61.0	3	68.0
1	65.7	3	65.9
1	51.4	3	54.7
1	53.6	3	53.6
1	59.0	3	58.7
2	64.4	3	58.7
2	55.8	3	65.7
2	58.7	3	66.5
2	54.7	3	56.7
2	52.7	3	55.4
2	67.8	3	51.5
2	61.6	3	54.8
2	58.7	3	57.2

Chapter 14 Data Set 2

Motivation	*GPA*	*Motivation*	*GPA*
1	3.4	6	2.6
6	3.4	7	2.5
2	2.5	7	2.8
7	3.1	2	1.8
5	2.8	9	3.7
4	2.6	8	3.1
3	2.1	8	2.5
1	1.6	7	2.4
8	3.1	6	2.1
6	2.6	9	4.0
5	3.2	7	3.9
6	3.1	8	3.1
5	3.2	7	3.3
5	2.7	8	3.0
6	2.8	9	2.0

smaller denominator lets us do so. Thus, instead of dividing by 10, we divide by 9. Or instead of dividing by 100, we divide by 99.

TECH TALK Biased estimates are appropriate if your intent is only to describe the characteristics of the sample. But if you intend to use the sample as an estimate of a population parameter, then the unbiased statistic is best to calculate.

Take a look in the following table and see what happens as the size of the sample gets larger (and moves closer to the population in size). The $n - 1$ adjustment has far less of an impact on the difference between the biased and the unbiased estimates of the standard deviation (the bold column in the table). All other things being equal, then, the larger the size of the sample, the less of a difference there is between the biased and the unbiased estimates of the standard deviation. Check out the following table, and you'll see what we mean.

Sample Size	Value of Numerator in Standard Deviation Formula	Biased Estimate of the Population Standard Deviation (dividing by n)	Unbiased Estimate of the Population Standard Deviation (dividing by n – 1)	Difference Between Biased and Unbiased Estimates
10	500	7.07	7.45	.38
100	500	2.24	2.25	.01
1,000	500	0.7071	0.7075	.0004

The moral of the story? When you compute the standard deviation for a sample, which is an estimate of the population, the closer to the size of the population the sample is, the more accurate the estimate will be.

What's the Big Deal?

The computation of the standard deviation is very straightforward. But what does it mean? As a measure of variability, all it tells us is how much each score in a set of scores, on the average, varies from the mean. But it has some very practical applications, as you

will find out in Chapter 4. Just to whet your appetite, consider this: The standard deviation can be used to help us compare scores from different distributions, *even when the means and standard deviations are different.* Amazing! This, as you will see, can be very cool.

THINGS TO REMEMBER

- The standard deviation is computed as the average distance from the mean. So, you will need to first compute the mean as a measure of central tendency. Don't fool around with the median or the mode in trying to compute the standard deviation.
- The larger the standard deviation, the more spread out the values are, and the more different they are from one another.
- Just like the mean, the standard deviation is sensitive to extreme scores. When you are computing the standard deviation of a sample and you have extreme scores, note that somewhere in your written report.
- If $s = 0$, there is absolutely no variability in the set of scores, and the scores are essentially identical in value. This will rarely happen.

COMPUTING THE VARIANCE

Here comes another measure of variability and a nice surprise. If you know the standard deviation of a set of scores and you can square a number, you can easily compute the variance of that same set of scores. This third measure of variability, the **variance**, is simply the standard deviation squared.

In other words, it's the same formula you saw earlier but without the square root bracket, like the one shown in Formula 3.3:

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} \tag{3.3}$$

If you take the standard deviation and never complete the last step (taking the square root), you have the variance. In other words, $s^2 = s \times s$, or the variance equals the standard deviation times itself

(or squared). In our earlier example, where the standard deviation was equal to 1.76, the variance is equal to 1.76^2 or 3.11. As another example, let's say that the standard deviation of a set of 150 scores is 2.34. Then, the variance would be 2.34^2 or 5.48.

You are not likely to see the variance mentioned by itself in a journal article or see it used as a descriptive statistic. This is because the variance is a difficult number to interpret and apply to a set of data. After all, it is based on squared deviation scores.

But the variance is important because it is used both as a concept and as a practical measure of variability in many statistical formulas and techniques. You will learn about these later in *Statistics for People Who (Think They) Hate Statistics.*

The Standard Deviation Versus the Variance

How are standard deviation and the variance the same, and how are they different?

Well, they are both measures of variability, dispersion, or spread. The formulas used to compute them are very similar. You see them all over the place in the "Results" sections of journals.

They are also quite different.

First, and most important, the standard deviation (because we take the square root of the average summed squared deviation) is stated in the original units from which it was derived. The variance is stated in units that are squared (the square root is never taken).

What does this mean? Let's say that we need to know the variability of a group of production workers assembling circuit boards. Let's say that they average 8.6 boards per hour, and the standard deviation is 1.59. The value 1.59 means that the difference in the average number of boards assembled per hour is about 1.59 circuit boards from the mean.

Let's look at an interpretation of the variance, which is 1.59^2, or 2.53. This would be interpreted as meaning that the average difference between the workers is about 2.53 circuit boards *squared* from the mean. Which of these two makes more sense?

USING THE COMPUTER TO COMPUTE MEASURES OF VARIABILITY

Let's use SPSS to compute some measures of variability. We are using the file named Chapter 3 Data Set 1.

There is one variable in this data set:

Variable	*Definition*
ReactionTime	Reaction time on a tapping task

Here are the steps to compute the measures of variability that we discussed in this chapter.

1. Open the file named Chapter 3 Data Set 1.
2. Click Analyze → Descriptive Statistics → Frequencies.
3. Double-click on the ReactionTime variable to move it to the Variable(s) box.
4. Click Statistics, and you will see the Frequencies: Statistics dialog box. Use this dialog box to select the variables and procedures you want to perform.
5. Under Dispersion, click Std. Deviation.
6. Under Dispersion, click Variance.
7. Under Dispersion, click Range.
8. Click Continue.
9. Click OK.

The SPSS Output

Figure 3.1 shows selected output from the SPSS procedure for ReactionTime. There are 30 valid cases with no missing cases, and the standard deviation is .70255. The variance equals .494 (or s^2), and the range is 2.60.

Statistics

Reaction Time

N	Valid	30
	Missing	0
Std. Deviation		.70255
Variance		.494
Range		2.60

Figure 3.1 Output for the Variable ReactionTime

Let's try another one, titled Chapter 3 Data Set 2. There are two variables in this data set:

Variable	*Definition*
MathScore	Score on a mathematics test
ReadingScore	Score on a reading test

Follow the same set of instructions as given previously, only in Step 3, you select both variables. The SPSS output is shown in Figure 3.2, where you can see selected output from the SPSS procedure for these two variables. There are 30 valid cases with no missing cases, and the standard deviation for math scores is 12.36 with a variance of 152.7 and a range of 43. For reading scores, the standard deviation is 18.700, the variance is a whopping 349.689 (that's pretty big), and the range is 76 (which is large as well, reflecting the similarly large variance).

Statistics

		Math_Score	Reading_Score
N	Valid	30	30
	Missing	0	0
Std. Deviation		12.357	18.700
Variance		152.700	349.689
Range		43	76

Figure 3.2 Output for the Variables MathScore and ReadingScore

Summary

Measures of variability help us even more fully understand what a distribution of data points looks like. Along with a measure of central tendency, we can use these values to distinguish distributions from one another and effectively describe what a collection of test scores, heights, or measures of personality looks like. Now that we can think and talk about distributions, let's explore ways we can look at them.

Time to Practice

1. Why is the range the most convenient measure of dispersion, yet the most imprecise measure of variability? When would you use the range?

2. Compute the exclusive and inclusive ranges for the following items.

High Score	*Low Score*	*Inclusive Range*	*Exclusive Range*
7	6		
89	45		
34	17		
15	2		
1	1		

3. Why would you expect more variability on a measure of personality in college freshman graders than you would on a measure of height?
4. Why does the standard deviation get smaller as the individuals in a group score more similarly on a test?
5. For the following set of scores, compute the range, the unbiased and the biased standard deviations, and the variance. Do the exercise by hand.

   31, 42, 35, 55, 54, 34, 25, 44, 35

   Why is the unbiased estimate greater than the biased estimate?
6. Use SPSS to compute all the descriptive statistics for the following set of three test scores over the course of a semester. Which test had the highest average score? Which test had the lowest amount of variability?

Test 1	*Test 2*	*Test 3*
50	50	49
48	49	47
51	51	51
46	46	55
49	48	55
48	53	45
49	49	47
49	52	45
50	48	46
50	55	53

7. For the following set of scores, compute by hand the unbiased estimates of the standard deviation and variance.

   4, 5, 6, 2, 5, 7, 5, 6, 8, 5

8. The variance for a set of scores is 25. What is the standard deviation and what is the range?

9. This practice problem uses the data contained in the file named Chapter 3 Data Set 3. There are two variables in this data set.

Variable	*Definition*
Height	height in inches
Weight	weight in pounds

Using SPSS, compute all of the measures of variability you can for height and weight.

10. How can you tell whether SPSS produces a biased or an unbiased estimate of the standard deviation?.

4 A Picture Really Is Worth a Thousand Words

Difficulty Scale ☺☺☺☺ (pretty easy, but not a cinch)

What you'll learn about in this chapter

- Why a picture is really worth a thousand words
- How to create a histogram and polygon
- Different types of charts and their uses
- Using Excel and SPSS to create charts

WHY ILLUSTRATE DATA?

In the previous two chapters, you learned about two important types of descriptive statistics—measures of central tendency and measures of variability. Both of these provide you with the one best score for describing a group of data (central tendency) and a measure of how diverse, or different, scores are from one another (variability).

What we did not do, and what we will do here, is examine how differences in these two measures result in different-looking distributions. Numbers alone (such as $\bar{X} = 10$ and $s = 3$) may be important, but a visual representation is a much more effective way of examining the characteristics of a distribution as well as the characteristics of any set of data.

So, in this chapter, we'll learn how to visually represent a distribution of scores as well as how to use different types of graphs to represent different types of data.

TEN WAYS TO A GREAT FIGURE (EAT LESS AND EXERCISE MORE?)

Whether you create illustrations by hand or use a computer program, the principles of decent design still apply. Here are 10 to copy and put above your desk.

1. **Minimize chart or graph junk.** "Chart junk" (a close cousin to "word junk") is where you use every function, every graph, and every feature a computer program has to make your charts busy, full, and uninformative. Less is definitely more.

2. **Plan out your chart before you start creating the final copy.** Use graph paper even if you will be using a computer program to generate the graph.

3. **Say what you mean and mean what you say—no more and no less.** There's nothing worse than a cluttered (with too much text and fancy features) graph to confuse the reader.

4. **Label everything so nothing is left to the misunderstanding of the audience.**

5. **A graph should communicate only one idea.**

6. **Keep things balanced.** When you construct a graph, center titles and axis labels.

7. **Maintain the scale in a graph.** The scale refers to the relationship between the horizontal and vertical axes. This ratio should be about three to four, so a graph that is three inches wide will be about four inches tall.

8. **Simple is best and less is more.** Keep the chart simple, but not simplistic. Convey the one idea as straightforwardly as possible, with distracting information saved for the accompanying text. Remember, a chart or graph should be able to stand alone, and the reader should be able to understand the message.

9. **Limit the number of words you use.** Too many words, or words that are too large, can detract from the visual message your chart should convey.

10. **A chart alone should convey what you want to say.** If it doesn't, go back to your plan and try it again.

FIRST THINGS FIRST: CREATING A FREQUENCY DISTRIBUTION

The most basic way to illustrate data is through the creation of a frequency distribution. A **frequency distribution** is a method of tallying and representing, how often certain scores occur. In the creation of a frequency distribution, scores are usually grouped into class intervals, or ranges of numbers.

Here are 50 scores on a test of reading comprehension and what the frequency distribution for these scores looks like.

Here are the raw data on which it is based:

47	10	31	25	20
2	11	31	25	21
44	14	15	26	21
41	14	16	26	21
7	30	17	27	24
6	30	16	29	24
35	32	15	29	23
38	33	19	28	20
35	34	18	29	21
36	32	16	27	20

And here's the frequency distribution:

Class Interval	*Frequency*
45–49	1
40–44	2
35–39	4
30–34	8
25–29	10
20–24	10
15–19	8
10–14	4
5–9	2
0–4	1

The Classiest of Intervals

As you can see from the above table, a **class interval** is a range of numbers, and the first step in the creation of a frequency distribution is to define how large each interval will be. As you can see in the frequency distribution that we created, each interval spans five possible scores such as 5–9 (which includes scores 5, 6, 7, 8, and 9) and

40–44 (which includes scores 40, 41, 42, 43, and 44). How did we decide to have an interval that includes only five scores? Why not five intervals, each consisting of 10 scores? Or two intervals, each consisting of 25 scores?

Here are some general rules to follow in the creation of a class interval, regardless of the size of values in the data set you are dealing with.

1. Select a class interval that has a range of 2, 5, 10, 15, or 20 data points. In our example, we chose 5.
2. Select a class interval so that 10 to 20 such intervals cover the entire range of data. A convenient way to do this is to compute the range, then divide by a number that represents the number of intervals you want to use (between 10 and 20). In our example, there are 50 scores and we wanted 10 intervals: 50/10 = 5, which is the size of each class interval. If you had a set of scores ranging from 100 to 400, you can start with the following estimate and work from there: 300/20 = 15, so 15 would be the class interval.
3. Begin listing the class interval with a multiple of that interval. In our frequency distribution shown earlier, the class interval is 5 and we started with the lowest class interval of 0.
4. Finally, the largest interval goes at the top of the frequency distribution.

Once class intervals are created, it's time to complete the frequency part of the frequency distribution. That's simply counting the number of times a score occurs in the raw data and entering that number in each of the class intervals represented by the count.

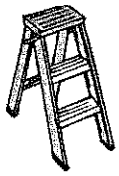
In the frequency distribution that we created on page 50, the number of scores that occur between 30 and 34 and are in the 30–34 class interval is 8. So, an 8 goes in the column marked Frequency. There's your frequency distribution.

THE PLOT THICKENS: CREATING A HISTOGRAM

Now that we've got a tally of how many scores fall in what class intervals, we'll go to the next step and create what is called a **histogram**, a visual representation of the frequency distribution where the frequencies are represented by bars.

Depending on the book you read and the software you use, visual representations of data are called graphs (such as in SPSS) or charts (such as in the Microsoft spreadsheet Excel). It really makes no difference. All you need to know is that a graph or a chart is the visual representation of data.

To create a histogram, do the following:



1. Using a piece of graph paper, place values at equal distances along the *x*-axis, as shown in Figure 4.1. Now, identify the midpoint of the class intervals, which is the middle point in the class interval. It's pretty easy to just eyeball, but you can also just add the top and bottom values of the class interval and divide by 2. For example, the midpoint of the class interval 0–4 is the average of 0 and 4, or 4/2 = 2.

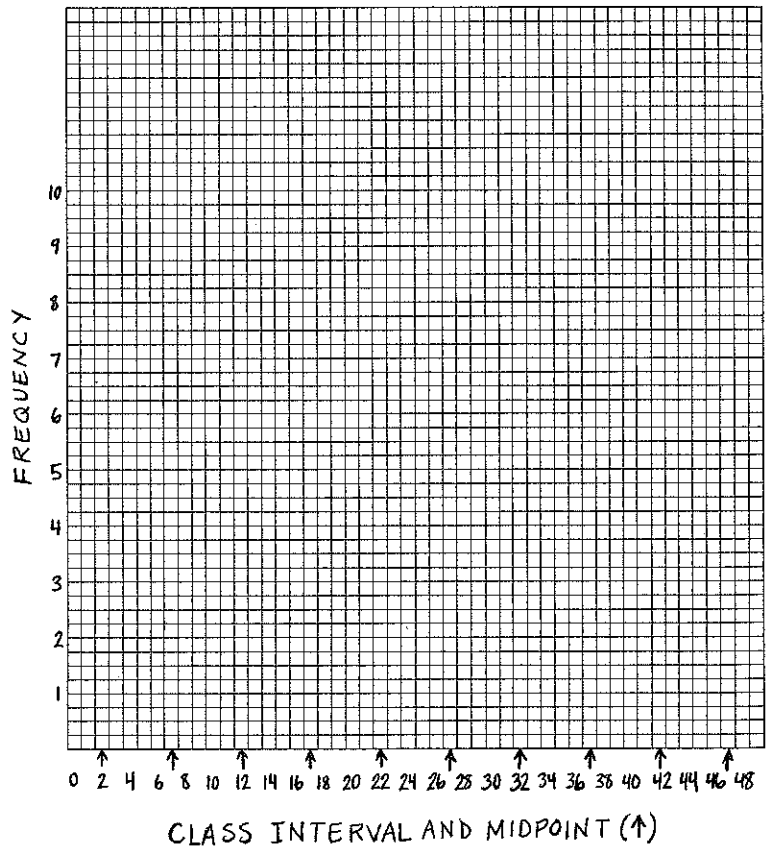


Figure 4.1 Class Intervals Along the *x*-Axis

2. Draw a bar or column around each **midpoint** that represents the entire class interval to the height representing the frequency of that class interval. For example, in Figure 4.2, you can see our first entry where the class interval of 0–4 is represented by the frequency of 1 (representing the one time a value between 0 and 4 occurs). Continue drawing bars or columns until each of the frequencies for each of the class intervals is represented. Here's a nice hand-drawn (really!) histogram for the frequency distribution of the 50 scores that we have been working with so far.

   Notice how each class interval is represented by a range of scores along the *x*-axis.

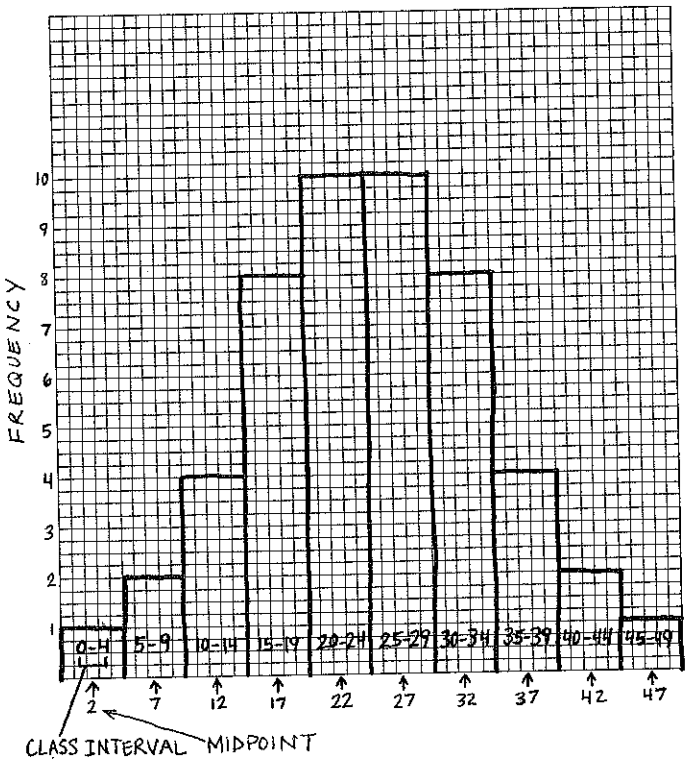


Figure 4.2 A Hand-Drawn Histogram

The Tally-Ho Method

You can see by the simple frequency distribution that you saw at the beginning of the chapter that you already know more about the distribution of scores than just a simple listing of them. You have a good idea of what values occur with what frequency. But another visual representation (besides a histogram) can be done by using tallies for each of the occurrences, as shown in Figure 4.3.

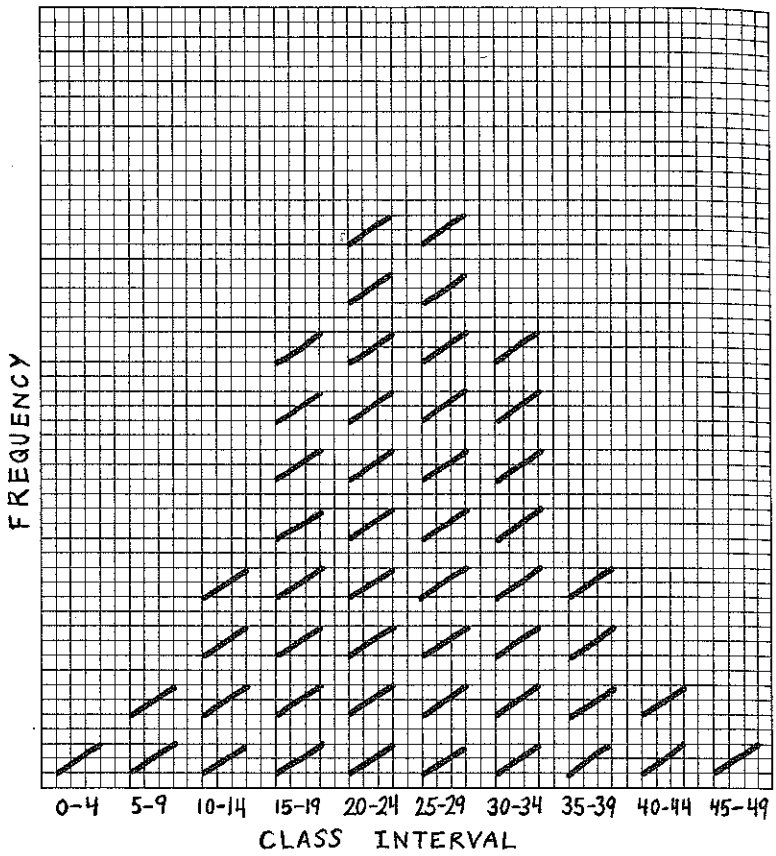


Figure 4.3 Tallying Scores

We used tallies that correspond with the frequency of scores that occur within a certain class. This gives you an even better visual representation of how often certain scores occur relative to other scores.

THE NEXT STEP: A FREQUENCY POLYGON

Creating a histogram or a tally of scores wasn't so difficult, and the next step (and the next way of illustrating data) is even easier. We're going to use the same data—and, in fact, the histogram that you just saw created—to create a frequency polygon. A **frequency polygon** is a continuous line that represents the frequencies of scores within a class interval, as shown in Figure 4.4.

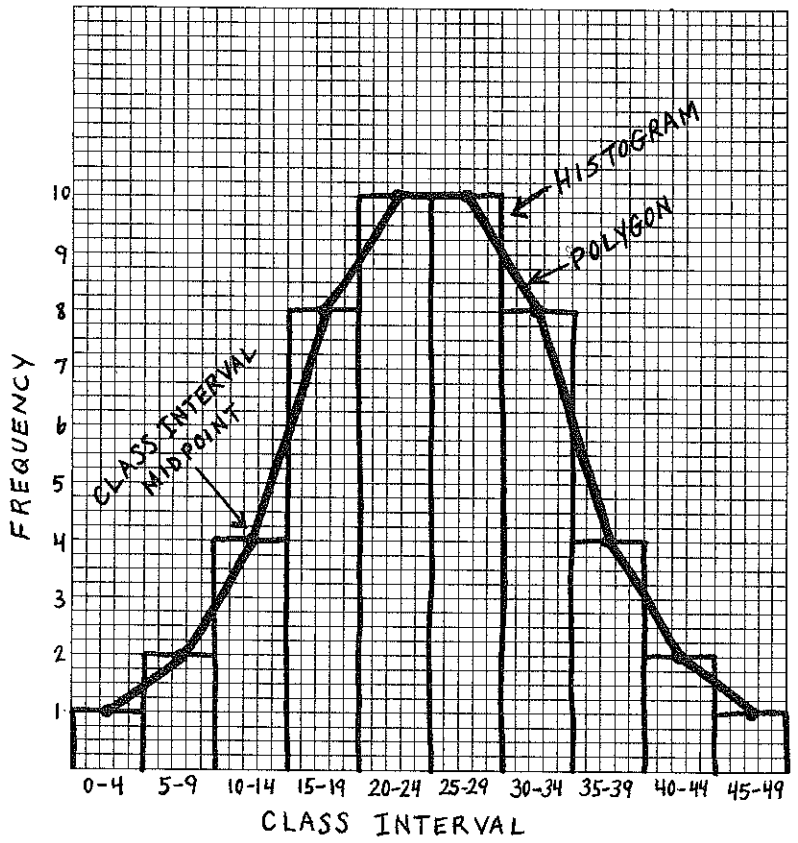
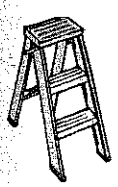


Figure 4.4 A Hand-Drawn Frequency Polygon



How did we draw this? Here's how.

1. Place a midpoint at the top of each bar or column in a histogram (see Figure 4.2).
2. Connect the lines and you've got it—a frequency polygon!

Note that in Figure 4.4, the histogram on which the frequency polygon is based is drawn using vertical and horizontal lines, and the polygon is drawn using curved lines. That's because, although we want you to see what a frequency polygon is based on, you usually don't see the underlying histogram.

Why use a frequency polygon rather than a histogram to represent data? It's more a matter of preference than anything else. A frequency

polygon appears more dynamic than a histogram (a line that represents change in frequency always looks neat), but you are basically conveying the same information.

Cumulating Frequencies

Once you have created a frequency distribution and have visually represented those data using a histogram or a frequency polygon, another option is to create a visual representation of the cumulative frequency of occurrences by class intervals. This is called a **cumulative frequency distribution.**

A cumulative frequency distribution is based on the same data as a frequency distribution, but with an added column (Cumulative Frequency), as shown below.

Class Interval	*Frequency*	*Cumulative Frequency*
45–49	1	50
40–44	2	49
35–39	4	47
30–34	8	43
25–29	10	35
20–24	10	25
15–19	8	15
10–14	4	7
5–9	2	3
0–4	1	1

The cumulative frequency distribution begins by the creation of a new column labeled Cumulative Frequency. Then, we add the frequency in a class interval to all the frequencies below it. For example, for the class interval of 0–4, there is 1 occurrence and none below it, so the cumulative frequency is 1. For the class interval of 5–9, there are 2 occurrences in that class interval and one below it for a total of 3 (2 + 1) occurrences in that class interval or below it. The last class interval (45–49) contains 1 occurrence and there is a total of 50 occurrences at or below that class interval.

Once we create the cumulative frequency distribution, then the data can be plotted just as they were for a histogram or a frequency polygon. Only this time, we'll skip right ahead and plot the midpoint of each class interval as a function of the cumulative frequency of that class interval. You can see the cumulative frequency distribution in Figure 4.5 based on the 50 scores from the beginning of this chapter.

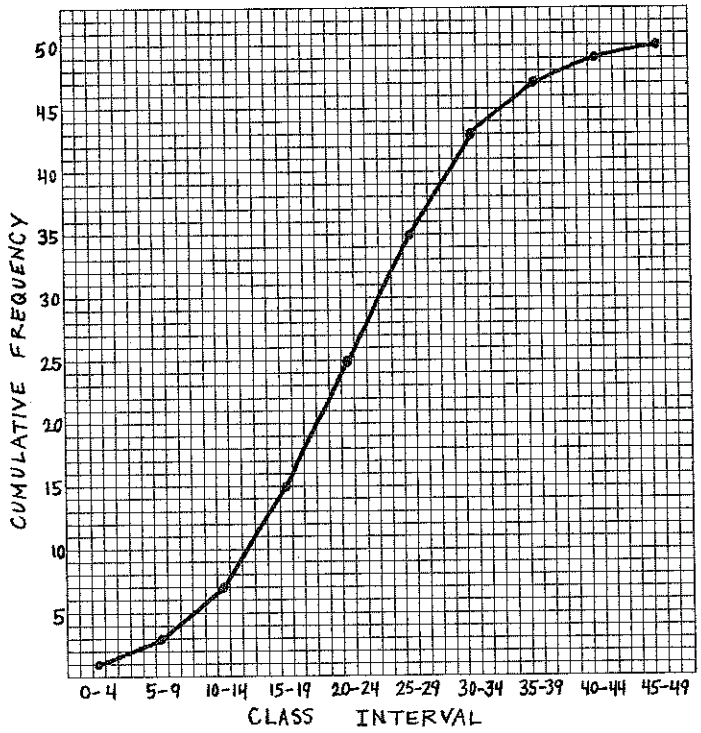


Figure 4.5 A Hand-Drawn Cumulative Frequency Distribution

TECH TALK

Another name for a cumulative frequency polygon is an **ogive.** And, if the distribution of the data is normal or bell shaped (see Chapter 8 for more on this), then the ogive represents what is popularly known as a bell curve or a normal distribution. SPSS creates a really nice ogive—it's called a P-P plot (for probability plot) and is really easy to create. See Appendix A for an introduction to creating graphs using SPSS, and also see the material toward the end of this chapter.

FAT AND SKINNY FREQUENCY DISTRIBUTIONS

You could certainly surmise by now that distributions can be very different from one another in a variety of ways. In fact, there are four different ways: average value, variability, skewness, and kurtosis. Those last two are new terms, and we'll define them as we show you what they look like. Let's define each of the four characteristics and then illustrate them.

Average Value

We're back once again to measures of central tendency. You can see in Figure 4.6 how three different distributions can differ in their average value. Notice that the average for Distribution C is more than the average for Distribution B, which, in turn, is more than the average for Distribution A.

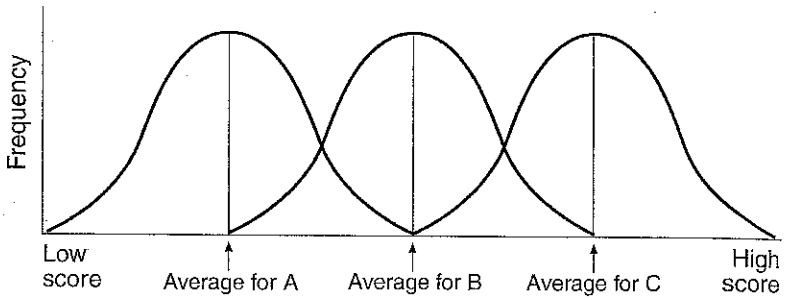


Figure 4.6 How Distributions Can Differ in Their Average Score

Variability

In Figure 4.7, you can see three distributions that all have the same average value, but differ in variability. The variability in Distribution A is less than that in Distribution B, and, in turn, less than that found in C. Another way to say this is that Distribution C has the largest amount of variability of the three distributions, and A has the least.

Skewness

Skewness is a measure of the lack of symmetry, or the lopsidedness, of a distribution. In other words, one "tail" of the distribution is longer than another. For example, in Figure 4.8, Distribution A's right tail is longer than its left tail, corresponding to a smaller number of occurrences at the high end of the distribution. This is a positively skewed distribution. This might be the case when you have a test that is very difficult, and few people get scores that are relatively high and many more get scores that are relatively low. Distribution C's right tail is shorter than its left tail, corresponding to a larger number of occurrences at the high end of the distribution. This is a negatively skewed

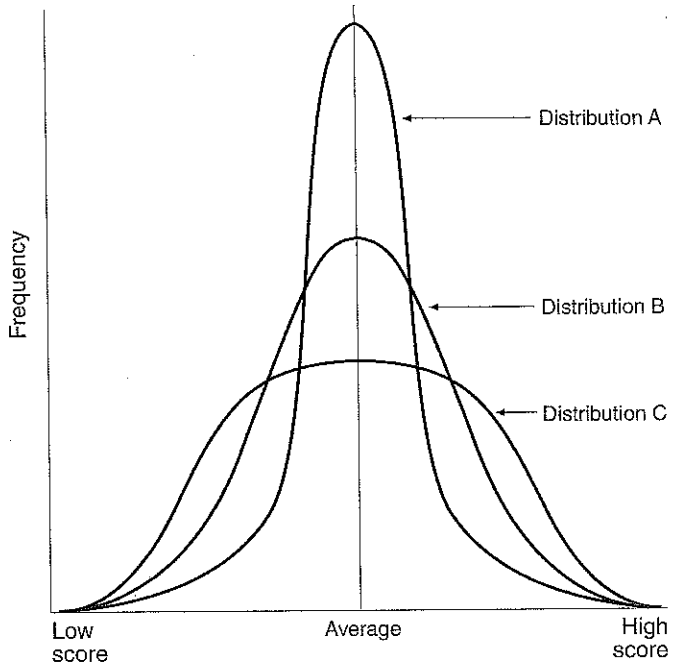


Figure 4.7 How Distributions Can Differ in Variability

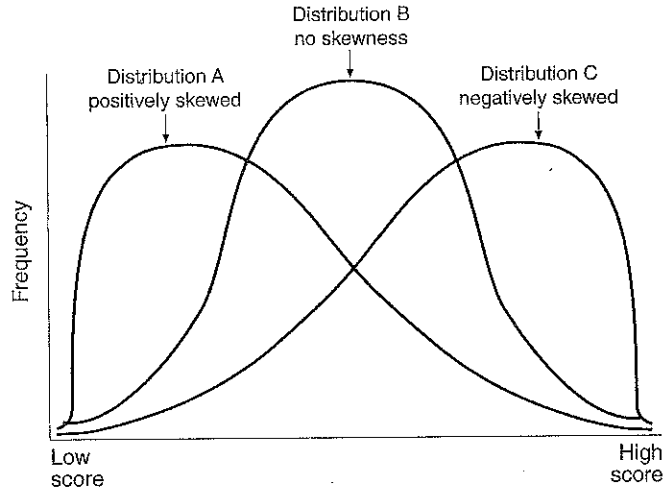


Figure 4.8 Degree of Skewness in Different Distributions

distribution and would be the case for an easy test (lots of high scores and relatively few low scores). And Distribution B—well, it's just right, equal lengths of tails and no skewness. If the mean is greater than the median, the distribution is positively skewed. If the median is greater than the mean, the distribution is negatively skewed.

Kurtosis

Even though this sounds like a medical condition, it's the last of the four ways that we can classify how distributions differ from one another. **Kurtosis** has to do with how flat or peaked a distribution appears, and the terms used to describe this characteristic are relative ones. For example, the term **platykurtic** refers to a distribution that is relatively flat compared to a normal, or bell-shaped, distribution. The term **leptokurtic** refers to a distribution that is relatively peaked compared to a normal, or bell-shaped, distribution. In Figure 4.9, Distribution A is platykurtic compared to Distribution B. Distribution C is leptokurtic compared to Distribution B. Figure 4.9 looks similar to Figure 4.7 for a good reason—distributions that are platykurtic, for example, are relatively more disperse than those that are not. Similarly, a distribution that is leptokurtic is less variable or dispersed relative to others.

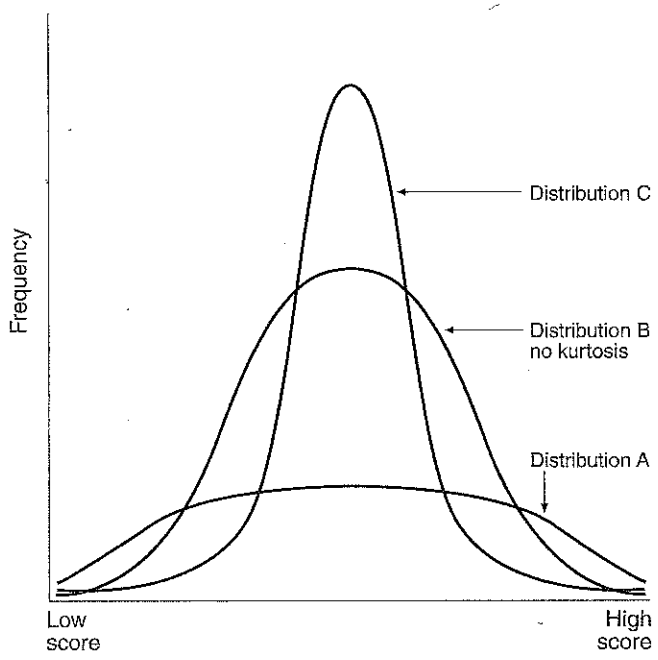


Figure 4.9 Degrees of Kurtosis in Different Distributions

While skewness and kurtosis are used mostly as descriptive terms (such as "That distribution is negatively skewed"), there are mathematical indicators of how skewed or kurtotic a distribution is.

For example, skewness is computed by subtracting the value of the median from the mean. If the mean of a distribution is 100 and the median is 95, the skewness value is 100 – 95 = 5, and the distribution is positively skewed. If the mean of a distribution is 85 and the median is 90, the skewness value is 85 – 90 = –5, and the distribution is negatively skewed. There's an even more sophisticated formula, which is not relative, but takes the standard deviation of the distribution into account so that skewness indicators can be compared to one another (see Formula 4.1).

$$Sk = \frac{3(\bar{X} - M)}{s} \tag{4.1}$$

where

- Sk is Pearson's (he's the correlation guy you'll learn about in Chapter 5) measure of skewness
- $\bar{X}$ is the mean
- M is the median

Here's an example: The mean of Distribution A is 100, the median is 105, and the standard deviation is 10. For Distribution B, the mean is 120, the median is 116, and the standard deviation is 10. Using Pearson's formula, the skewness of Distribution A is –1.5, and the skewness of Distribution B is 1.2. Distribution A is negatively skewed, and Distribution B is positively skewed. However, Distribution A is more skewed than Distribution B, regardless of the direction.

Let's not leave kurtosis out of this discussion. It too can be computed using a fancy formula as follows . . .

$$K = \frac{\sum\left(\frac{X - \bar{X}}{s}\right)^4}{n} - 3$$

where

- Σ = sum
- X = the individual score
- $\bar{X}$ = the mean of the sample
- s = the standard deviation
- n = the sample size

This is a pretty complicated formula that basically looks at how flat or peaked a set of scores is. You can see that if each score is the same, then the numerator is zero and $K = 0$. K equals zero when the distribution is normal or *mesokurtic* (now there's a word to throw around). If the individual scores (the Xs in the formula) differ greatly from the mean (and there is lots of variability), then the curve will probably be quite peaked.

OTHER COOL WAYS TO CHART DATA

What we did so far in this chapter is take some data and show how charts such as histograms and polygons can be used to communicate visually. But there are several other types of charts that are used in the behavioral and social sciences, and although it's not necessary for you to know exactly how to create them (manually), you should at least be familiar with their names and what they do. So here are some popular charts, what they do, and how they do it.

There are several very good personal computer applications for creating charts, among them the spreadsheet Excel (a Microsoft product) and, of course, SPSS. For your information, the charts that you see in Figures 4.10, 4.11, and 4.12 were created using Excel. The charts in the "Using the Computer to Illustrate Data" section were created using SPSS.

Column Charts

A column chart should be used when you want to compare the frequencies of different categories with one another. Categories are organized horizontally on the x-axis, and values are shown vertically on the y-axis. Here are some examples of when you might want to use a column chart:

- Number of voters by political affiliation
- The sales of three different types of products
- Number of children in each of six different grades

Figure 4.10 shows a graph of number of voters by political affiliation.

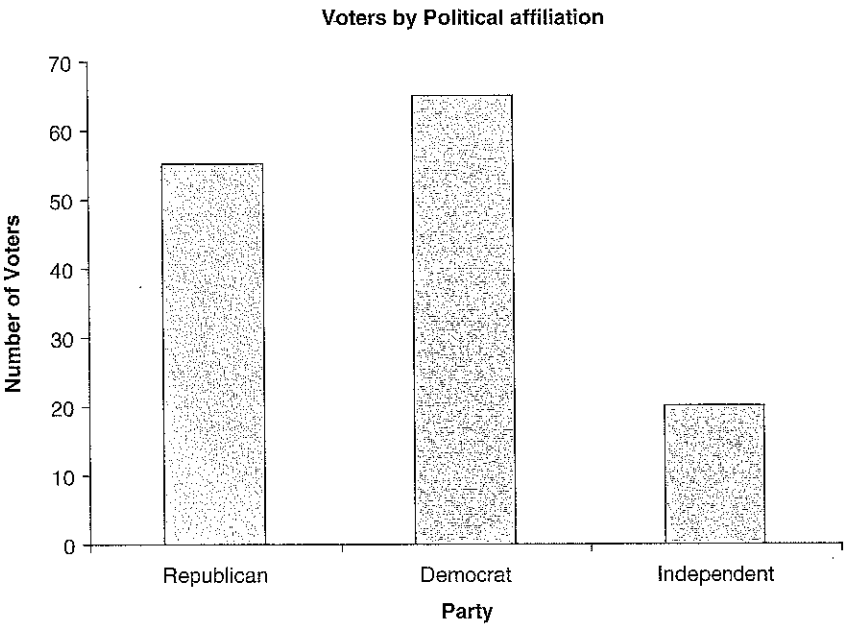


Figure 4.10 A Column Chart That Compares Different Categories With One Another

Bar Charts

A bar chart is identical to a column chart, but in this chart, categories are organized vertically on the *y*-axis and values are shown horizontally on the *x*-axis.

Line Charts

A line chart should be used when you want to show a trend in the data at equal intervals. Here are some examples of when you might want to use a line chart:

- Number of cases of mononucleosis (mono) per season among college students at three state universities
- Change in student enrollment over the school year
- Number of travelers on two different airlines for each quarter