3

The Shape of Distributions

- *Before We Begin*
- *The Basic Elements*
- *Beyond the Basics: Comparisons and Conclusions*
- *A Special Curve*
- *Chapter Summary*
- *Some Other Things You Should Know*
- *Key Terms*
- *Chapter Problems*

Up to this point, you've been looking at distributions presented as listings of scores or values. Now it's time to expand your horizons a bit. It's time to move beyond mere listings of scores or values and into the more visual world of graphs or curves.

As we take this next step, I'll ask you to do three things. First, I'll ask you to start thinking in a more abstract fashion. Sometimes I'll ask you to think about a concrete example that relates to a specific variable, but other times I'll ask you to think about a graph or curve in a very abstract sense. Second, I'll ask you to be very flexible in your thinking. I'll ask you to move from one

Up to this point, you've been looking at distributions presented as listings of scores or values. Now it's time to expand your horizons a bit. It's time to move beyond mere listings of scores or values and into the more visual world of graphs or curves.

As we take this next step, I'll ask you to do three things. First, I'll ask you to start thinking in a more abstract fashion. Sometimes I'll ask you to think about a concrete example that relates to a specific variable, but other times I'll ask you to think about a graph or curve in a very abstract sense. Second, I'll ask you to be very flexible in your thinking. I'll ask you to move from one type of graph to another, and sometimes I'll ask you to move back and forth between the two. Finally, I'll ask you to consider distributions with a larger number of cases than you've encountered so far. There's still no need to panic, though. Remember: The emphasis remains on the conceptual nature of the material.

Before We Begin

The last chapter allowed you to string together two very important concepts—namely, the standard deviation and the mean. Now it's time you expand your thinking by visualizing distributions and how they are influenced by the mean and standard deviation. For example, imagine two groups of test scores. Imagine that they have identical means but very different standard deviations. What about the reverse situation? What about a situation in which both classes have the same standard deviation, but they have radically different means?

Simple mental exercises along those lines can be very valuable, in your conceptual understanding of statistics. When you have reached the point where you can easily visualize different distributions (if only in a generalized form), I believe you've crossed an important milestone. I'm convinced that the ability to visualize distributions, particularly one distribution compared to another, is a talent that can be nurtured and developed. I'm also convinced that it's a significant asset when it comes to learning statistics. Therefore, try to visualize the various distributions that are discussed in this chapter. If that means that you can't read through the chapter in record time, so be it. Take your time. The goals are to learn the material *and* develop your visualization skills.

The Basic Elements

We'll start with an example that should be familiar to you by now—a situation in which some students have taken a test. This time, let's say that several thousand students took the test. Moreover, let's say you were given a chart or graph depicting the distribution of the test scores—something like Figure 3-1. A quick look at the chart tells you that it represents the distribution of scores by letter

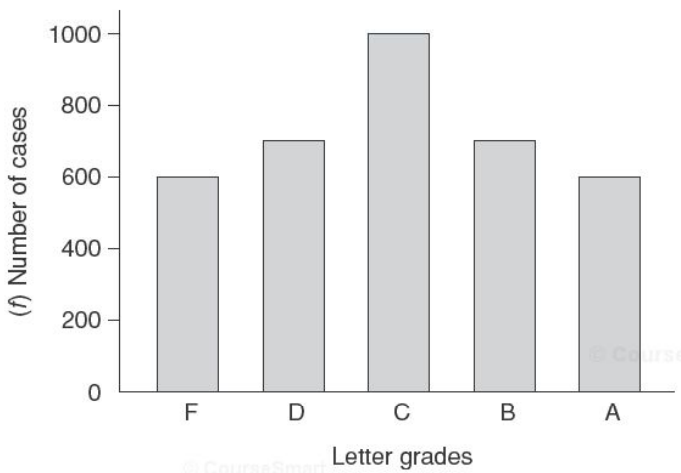


Figure 3-1 Distribution of Letter Grades

grade—the number of A's, B's, C's, and so forth. The illustration is probably very similar to many you've seen before. We refer to it as a *bar graph*.

A bar graph is particularly useful when the values or scores you want to represent fall into the category of nominal or ordinal data. Figure 3-1 is a perfect example. When the information about test scores is presented as letter grades (rather than actual test scores), you're dealing with ordinal level data; a letter grade of B is higher than a letter grade of C, but you don't really know how many points higher.

If, instead of letter grades, you had actual test scores expressed as numerical values, the measurement system would be far more refined, so to speak. That, in turn, would open the door to a more sophisticated method of illustrating the distribution of scores. Imagine for a moment that you had the actual scores for the same tests. Imagine that the measurement was very precise, with scores calculated to two decimal places (scores such as 73.28, 62.16, and 93.51). In this situation, the graph might look like the one shown in Figure 3-2.

Like the bar graph in Figure 3-1, the graph shown in Figure 3-2 is typical of what you might see in the way of data representation. Different values of the variable under consideration (in this case, test score) are shown along the baseline, and the frequency of occurrence is shown along the axis on the left side of the graph. The curve thus represents a **frequency distribution**—a table or graph that indicates how many times a value or score appears in a set of values or scores.

Instead of focusing on the specifics of the test scores presented in Figure 3-2, let's take a moment to reflect on curves or frequency distributions in general. Regardless of the specific information conveyed by the illustration, there are generally three important elements in a graph or plot of a frequency distribution.

First, there's the *X*-axis, or the baseline of the distribution. It reveals something about the range of values for the variable that you're considering. If you're looking at test scores, for example, the baseline or *X*-axis might show values ranging from 0 to 100. A frequency distribution of incomes might have a baseline with values ranging from, let's say, $15,000 to $84,000.

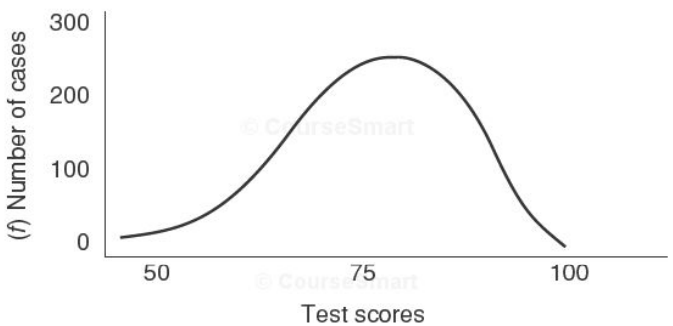


Figure 3-2 Distribution of Test Scores

Second, there's another axis—the *Y*-axis—usually running along the left side of the graph, with a symbol *f* to the side of it. The *f* stands for *frequency*—the number of times each value appears in the distribution, or the number of cases with a certain value (see Figure 3-3).

Now we add the third part of the graph—the curved line—as shown in Figure 3-4.

At this point, let me mention something that may strike you as obvious, but is worth mentioning nonetheless. It has to do with what is really represented by the space between the baseline and the curved line that forms the outline of the graph.

It's easy to look at a curve, such as the one shown in Figure 3-4, and forget that the area *under* the curve is actually filled with cases. Although the area under the curve may look empty, it is not. In fact, the area under the curve represents all the cases that were considered. Again, the area under the curve actually contains 100% of the cases (a point that will be important to consider later on). To understand this point, take a look at the graph shown in Figure 3-5, and think of each small dot as an individual case.

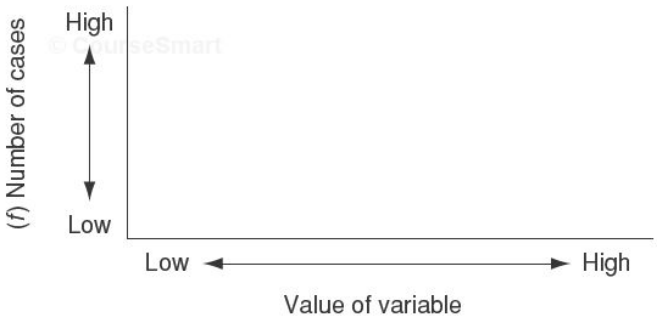


Figure 3-3 Components of a Frequency Distribution

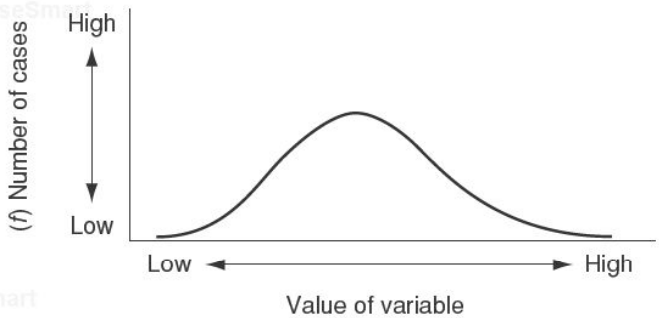


Figure 3-4 Components of a Frequency Distribution (Curve)

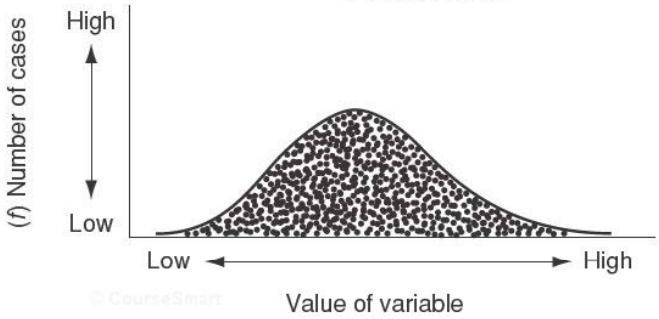


Figure 3-5 Cases/Observations Under a Curve

Remember: The area under the curve contains cases or observations! If necessary, take some time for a dark room moment at this point. Mentally visualize several different distributions. It doesn't make any difference what you think they represent. Just concentrate on the notion that cases or observations are under the curve—cases or observations stacked on top of one another (think of them as small dots, if need be, with all the dots stacked one upon the other).

LEARNING CHECK

Question: Although it appears to be empty, what is represented by the area under a curve?

Answer: The area under the curve represents cases or observations.

Beyond the Basics: Comparisons and Conclusions

Let's now turn our attention to a question that involves some material from the previous chapter—namely, the mean and the standard deviation. Instead of thinking about the distribution of a specific variable, let's consider two distributions—Distribution A and Distribution B—in an abstract sense. These two distributions have the same mean score (50), but beyond that, they are very different. In Distribution A, the scores are widely dispersed, ranging from 10 to 90. In Distribution B, the scores are tightly clustered about the mean, ranging from 30 to 70. These two distributions are represented by the two curves shown in Figure 3-6.

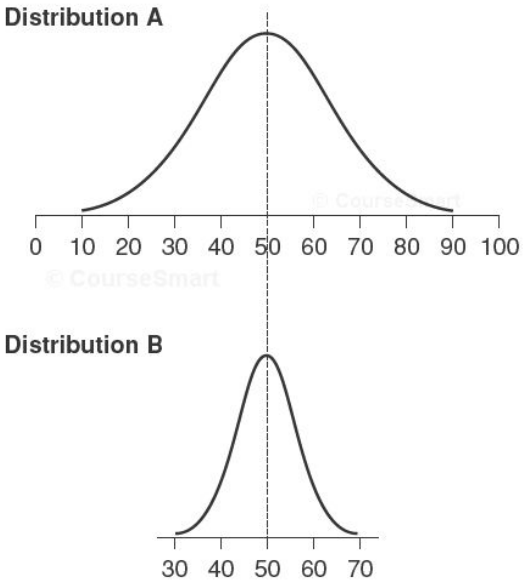


Figure 3-6 Comparison of Two Distributions With Same Mean but Different Standard Deviations

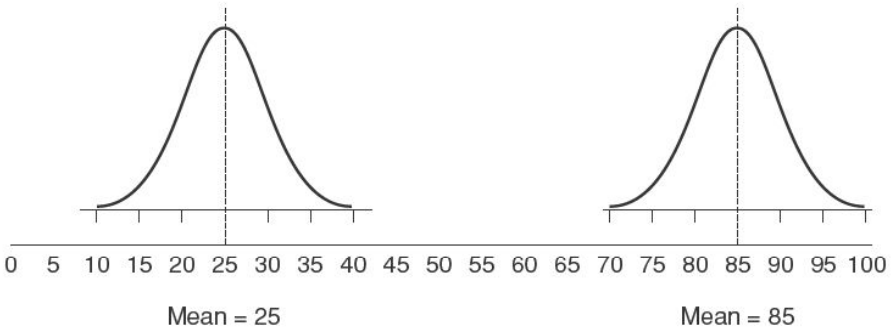


Figure 3-7 Comparison of Two Distributions With Same Standard Deviations but Different Means

Simple visual inspection of the curves should tell you that the standard deviation of Distribution B is smaller than the standard deviation of Distribution A. The edges of the curve in Distribution B don't extend out as far as they do in Distribution A.

Now consider the examples shown in Figure 3-7. Here, the two curves represent two distributions with the same standard deviation but very different mean values.

Assuming you're getting the hang of visualizing curves in your mind, let's now consider the matter of extreme scores in a distribution. Imagine for a

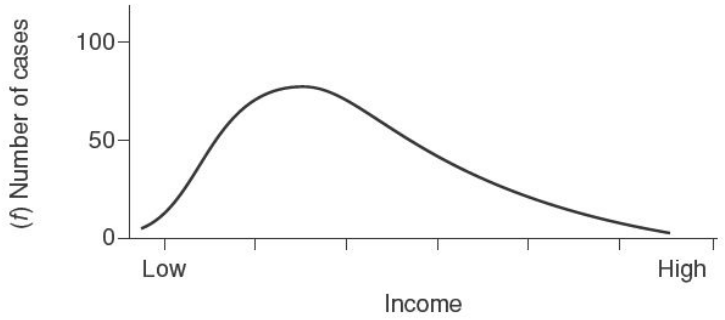


Figure 3-8 Positive Skew on Income

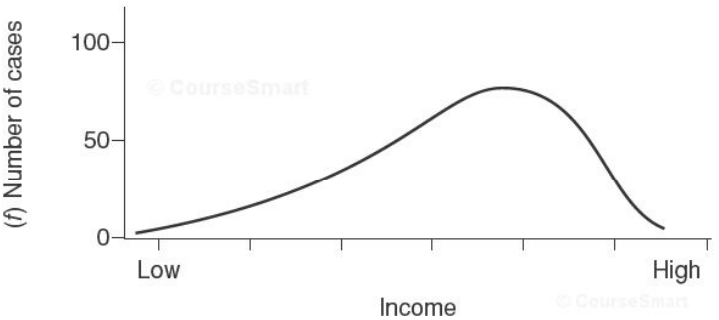


Figure 3-9 Negative Skew on Income

moment a distribution of income data—individual income information collected from a large number of people. Assume that most of the people have incomes that are close to the center of the distribution, but a few people have extremely high incomes. As you develop a mental picture, you should begin to visualize something that looks like the curve shown in Figure 3-8.

If, on the other hand, the extreme incomes were low incomes, the curve might look something like the one shown in Figure 3-9.

In statistics, we have a term for distributions like these. We refer to them as **skewed distributions.** When a distribution is skewed, it departs from *symmetry* in the sense that most of the cases are concentrated at one end of the distribution. We'll eventually have a closer look at this matter of skewness, but first let's consider some curves that lack those extremes. In other words, let's start by considering **symmetrical distributions.**

To understand the idea of symmetry (or a symmetrical distribution), imagine a situation in which you had height measurements from a large number of people. Height is a variable generally assumed to be distributed in a symmetrical fashion. Accordingly, the measurements would probably reflect roughly equal proportions of short and tall people in the sample. There might be just

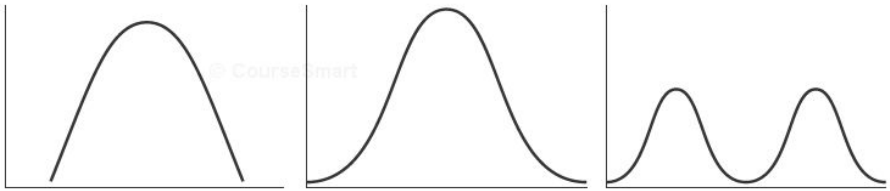


Figure 3-10 Symmetrical Curves/Distributions

a few very tall people, but, by the same token, there would be just a few very short people in the sample.

When a distribution is truly symmetrical, a line can be placed through the center of the distribution and the two halves will be mirror images. Fifty percent of the cases will be found on each side of the center line, and the shapes of the two sides of the distribution will be identical. When you think about it for just a moment, you'll realize that an infinite number of symmetrical shapes are possible. Figure 3-10 presents just a few for you to consider.

LEARNING CHECK

Question: What is a symmetrical distribution?
Answer: A symmetrical distribution is one in which the two halves of the distribution are mirror images of each other.

Question: What is a skewed distribution?
Answer: A skewed distribution is a distribution that departs from symmetry in the sense that most of the cases are concentrated at one end of the distribution.

As I mentioned before, a curve that departs from symmetry (one that is not symmetrical) is referred to as a skewed distribution. Think back to some of the examples involving data on income. In a distribution with some extremely high incomes (relative to the other incomes in the distribution), the distribution was skewed to the right. In the case of the distribution with some extremely low incomes (relative to the other incomes in the distribution), the distribution was skewed to the left.

When a distribution is skewed to the right, we say it has a **positive skew.** When a distribution is skewed to the left, we say it has a **negative skew.** To understand why we use the terms *positive* and *negative,* just think of it this way: If you have an imaginary line of numbers with a 0 in the middle, all values to the right of 0 are positive values, and all values to the left of 0 are negative. The terms relate to the elongated portion of the curve, which statisticians refer to as the **tail of the distribution** (see Figure 3-11). If the tail of the distribution

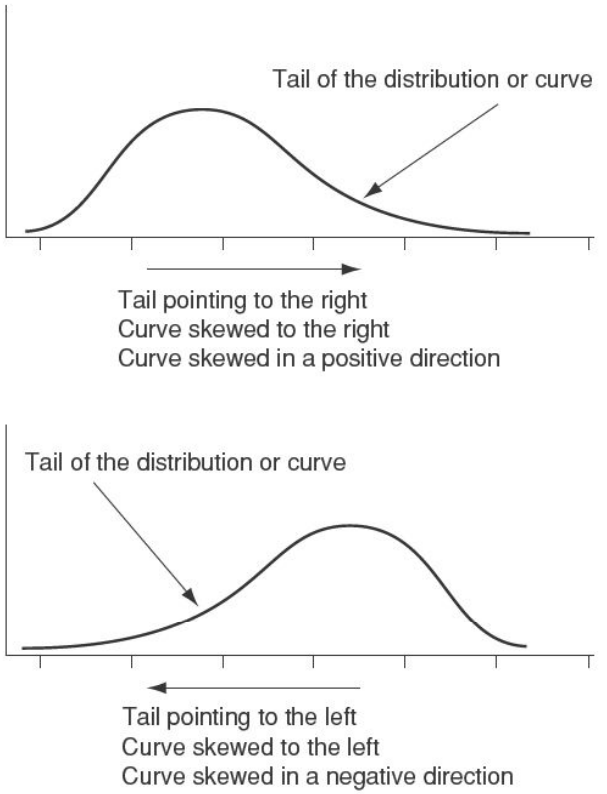


Figure 3-11 Skewed Curves/Distributions

extends toward the right, we say that the curve has a positive skew or is skewed to the right. Conversely, a curve with a tail that extends to the left is said to be skewed to the left or negatively skewed.

I've asked you to move from skewed to symmetrical distributions and back to skewed distributions—all in an effort to get you familiar with the basic difference, and primarily so you'll develop an appreciation for symmetrical distributions. Now I'm going to ask you to make the leap again—back to symmetrical distributions—but this time, we will consider a very special case.

A Special Curve

If you paid close attention when you looked at some of the illustrations of symmetrical curves, you probably noticed that symmetrical curves can take on

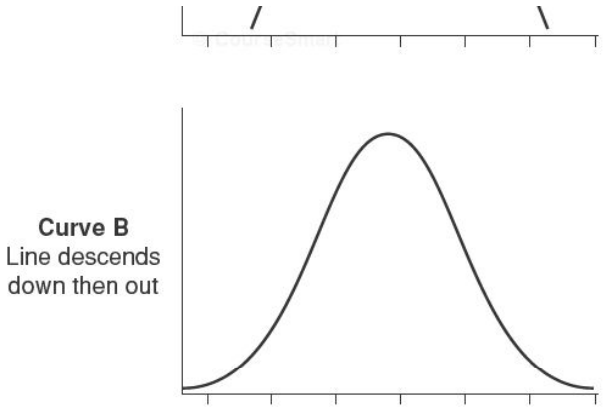


Figure 3-12 Comparison of Unimodal Symmetrical Curves

Consider, for example, Curve A in Figure 3-12. Focus on the highest point (the midpoint) of the distribution, and take note of how the line descends on either side of the midpoint. In a sense, the curved line descends out and down toward the baseline. Now focus on Curve B. Like Curve A, Curve B is a unimodal symmetrical curve (it has only one mode), but the manner in which the curved line descends toward the baseline is very different. Starting at the high point of the curve, the path of the curved line descends and then begins to turn outward. The curved line doesn't just drop to the baseline. Instead, it shows a pattern of gradual descent that moves in an outward direction.

Obviously, any number of curves could show this general pattern of descending down and then out. To statisticians, though, there's a particular type of curve that's of special interest. They refer to this very special sort of symmetrical curve as a **normal curve.**

A normal curve is symmetrical, and it descends down and then out. Moreover, the mean, median, and mode all coincide on a normal curve. But the special characteristics of a normal curve go beyond that. Indeed, a normal curve is one that conforms to a precise mathematical function. When a curve is, in fact, a normal curve, the mean and the standard deviation define the total shape of the curve. The curve may be relatively flat; it may be sharply peaked; or it may have a more moderate shape. The point is that the shape is predictable

because a normal curve is defined by a precise mathematical function. Once again, the mean and the standard deviation will define the exact shape of a normal curve.

LEARNING CHECK

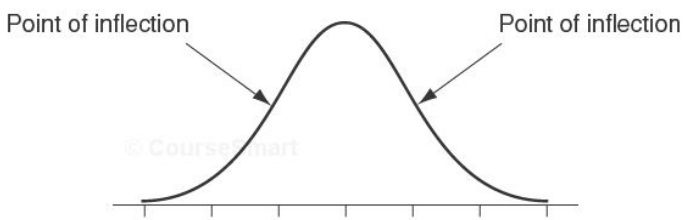
Question: What type of symmetrical curve is of particular interest to statisticians?

Answer: A normal curve.

Take a close look at Figure 3-13. Starting at the top of the curve (which happens to be where the mean, median, and mode coincide), you can trace the line of the curve on one side. The line descends downward at a fairly steady rate, but the line eventually reaches a point at which it begins to turn in a more outward direction. From that point—known as the **point of inflection**—the rate of descent of the curve toward the baseline is more gradual. To appreciate this element, take the time to trace the curved line in Figure 3-13, either visually or with your index finger or a pencil. Concentrate on the point at which the curve begins to change directions—the point of inflection.

In normal curves with a small standard deviation, the curve will be fairly peaked in shape, and the degree of initial downward descent of the curve will be very noticeable. In normal curves with a larger standard deviation, the curve will be flatter, and the degree of initial downward descent of the curve will be less pronounced. Either way, the entire shape of the distribution is defined by the mean and standard deviation. Figure 3-14 shows some examples of normal curves.

Because a normal curve is one that conforms to a precise mathematical function, it's possible to know a great deal of information about the data



The point of inflection is the point at which the curved line begins to change direction.

Figure 3-13 Locating the Point of Inflection

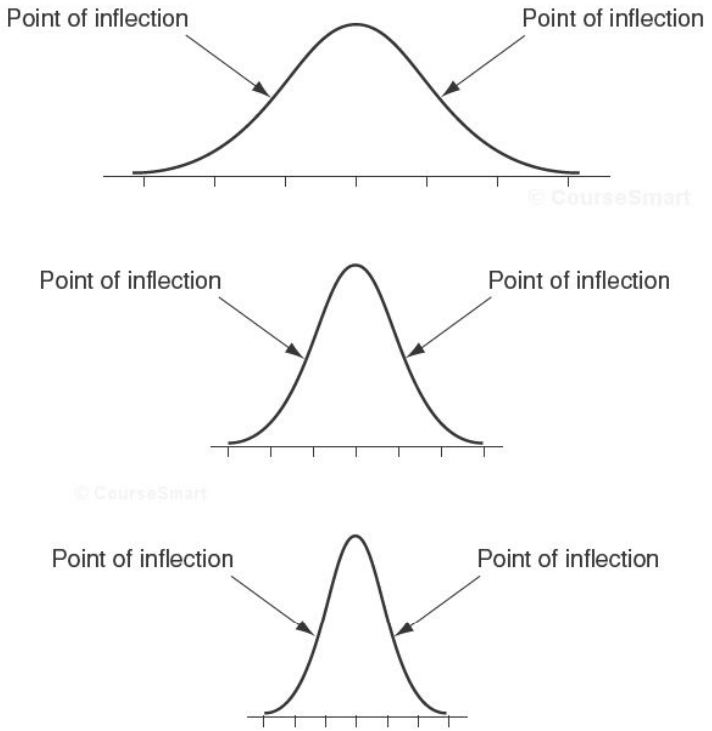


Figure 3-14 More Examples for Locating the Point of Inflection

distribution that underlies any normal curve. As a matter of fact, the point at which the curve begins to turn outward—the point of inflection—will be one standard deviation away from the mean.

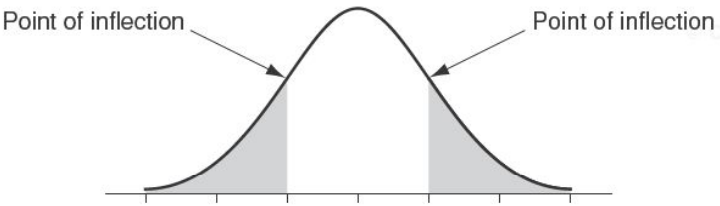
For example, let's say we have a distribution of test scores, and the test scores are normally distributed. What this means is that the distribution of scores, if plotted in a graph, will form a normal curve. Now let's say that same distribution of scores has a mean of 60 and a standard deviation of 3. Since we know that the points of inflection will always be one standard deviation above and below the mean on a normal curve, we know that the points of inflection will correspond to scores of 63 and 57, respectively.

LEARNING CHECK

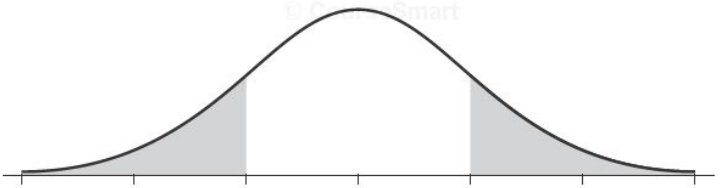
Question: What is the point of inflection of a normal curve?
Answer: It is the point at which the curve begins to change direction. This point is also one standard deviation away from the mean.

As it turns out, we're in a position to know a lot more than that. For example, imagine that you're looking at a normal curve, and you mark the inflection points on both sides of the mean—the points on either side of the mean where the curve begins to change direction (even if ever so slightly). You now know that you have marked off the points that correspond to one standard deviation above and below the mean. In addition, however, if you draw lines down from the inflection points to the baseline, you will be marking off a portion of the normal curve that contains slightly more than 68% of the total cases, or 68% of the area under the curve. Why? Because that's the way a normal curve is mathematically defined.

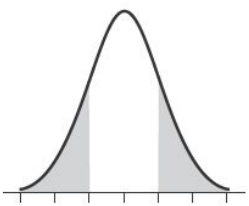
This point is central to everything that follows, so take a close look at Figure 3-15.



Approximately 68% of cases in a normal distribution are between one standard deviation above and below the mean.



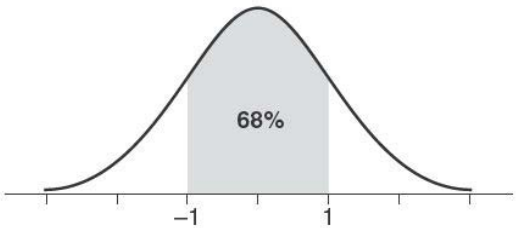
Even if the curve is relatively flat, approximately 68% of the cases will be found ±1 standard deviation from the mean.



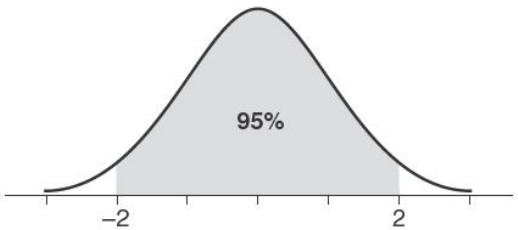
Even if the curve is relatively peaked, approximately 68% of the cases will be found ±1 standard deviation from the mean.

Figure 3-15 General Shape of a Normal Curve

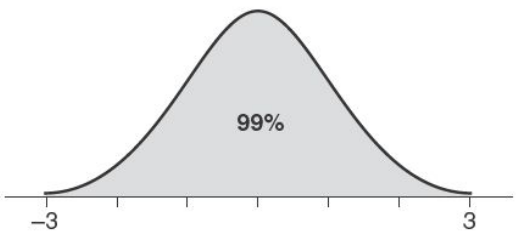
If you marked off two standard deviations from the mean, you would have marked off a portion of a normal curve that contains slightly more than 95% of the total cases. And lines drawn at three standard deviations above and below a normal curve will enclose an area that contains more than 99% of the cases (see Figure 3-16). If you're still wondering why, the answer remains the same: That's how a normal curve is mathematically defined.



Approximately 68% of cases in a normal distribution are between one standard deviation above and below the mean.



Approximately 95% of cases in a normal distribution are between two standard deviations above and below the mean.



Approximately 99% of cases in a normal distribution are between three standard deviations above and below the mean.

Figure 3-16 Distribution of Cases or Area Under a Normal Curve

This information—relating standard deviations to the area under the normal curve—is so fundamental to statistical inference that statisticians often think of it as the **1-2-3 Rule.** Here it is again, just for good measure:

> One standard deviation on either side of the mean of a normal curve will encompass approximately 68% of the area under the curve.
>
> Two standard deviations on either side of the mean of a normal curve will encompass approximately 95% of the area under the curve.
>
> Three standard deviations above and below the mean will encompass slightly more than 99% of the area under the curve.

At this point, let me suggest that you take a moment or two to digest this material. Start with an understanding that a normal curve is one that follows a precise mathematical function. Then concentrate on the notion that for any truly normal curve, there is a known area under the curve between standard deviations (for example, 68% of the area under the curve is between one standard deviation above and below the mean). Fix the critical values in your mind: ±1 standard deviation encloses approximately 68% of cases; ±2 standard deviations encompasses approximately 95%; and ±3 standard deviations encompasses slightly more than 99%.

LEARNING CHECK

Question: What does the 1-2-3 Rule tell us?

Answer: It tells us the amount of area under the normal curve that is located between certain points (expressed in standard deviation units). Approximately 68% of the area is found between one standard deviation above and below the mean. Approximately 95% of the area is found between two standard deviations above and below the mean. Slightly more than 99% of the area is found between three standard deviations above and below the mean.

Whether you realize it or not, you're actually accumulating quite a bit of knowledge about normal curves. Indeed, if you throw in the fact that 50% of the area, or cases, under the curve are going to be found on either side of the mean, you're actually in a position to begin answering a few questions.

For example, let's say you know that some test results are normally distributed. In other words, a plot of the scores reveals a distribution that

conforms to a normal curve. Let's also say that you know you scored one standard deviation above the mean. Now here's a reasonable question, given what you already know: Approximately what percentage of the test scores would be below yours? Approximately what percentage of the scores would be above yours? There's no need to hit the panic button. Just think it through.

Start with what you know about the percentage of cases (or scores) that fall between one standard deviation above and one standard deviation below the mean. You know (from what you read earlier) that approximately 68% are found between these two points. Since a normal curve is symmetrical, this means that approximately 34% of the scores will be found between the mean and one standard deviation *above* the mean. In other words, the 68% (approximately) will be equally divided between the two halves of the curve. Therefore, you will find approximately 34% of the cases (or area) between the mean and one standard deviation (either one standard deviation above or one standard deviation below). You also know (because of symmetry) that the lower half of the curve will include 50% of the cases. So, all that remains to answer the question is some simple addition:

50%	the lower half of the curve
+ 34%	the percentage between the mean and one standard deviation above the mean
84%	the percentage of cases below one standard deviation above the mean (therefore, aproximately 84% of the cases would be below your score, and approximately 16% would be above your score)

Figure 3-17 shows the same solution in graphic form.

By now you're probably getting the idea that normal curves and distributions have an important place in the world of statistical analysis. Indeed, the idea of a normal curve is central to many statistical procedures. As a matter of fact, the notion of a normal curve or distribution is so fundamental to statistical inference that statisticians long ago developed a special case normal curve as a point of reference. We refer to it as the *standardized normal curve,* and that's our topic in Chapter 4.

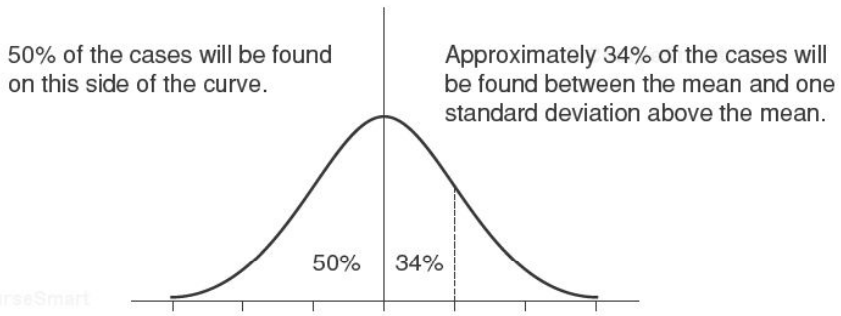


Figure 3-17 The Logic Behind the Problem Solution

Chapter Summary

In your exploration of data distributions, curves, and such, you've taken a very important step toward statistical reasoning. You took your first step in that regard as soon as you began to visualize a curve. Your ability to visualize data distributions in the form of curves is something that will come into play throughout your statistical education, so there's no such thing as too much practice at the outset.

Ideally, you've learned more than what a data distribution might look like if it were plotted or graphed. For example, you've learned about symmetrical curves, and you've learned about skewed curves. You've been introduced to the notion of a normal distribution and what normal distributions look like when they are graphed.

You've learned, for example, that a normal distribution can take on any number of different shapes (from very flat to very peaked), but the exact shape is always determined by two values—the mean and standard deviation of the underlying distribution. You've also learned that this mathematical definition of a normal curve's shape (based on the mean and standard deviation) makes the shape of a normal curve predictable.

You've learned that the points of inflection on a normal curve are the points that correspond to one standard deviation above and below the mean. And you've come to understand that, given a normal curve, a predictable amount of area (or cases) under the curve corresponds to specific points along the baseline (the 1-2-3 Rule).

As we move to the next chapter, the material that you've learned about normal curves in general will come into play in a major way. As you're about to discover, the notion of a normal distribution or normal curve is central to statistical analysis, so much so that it becomes the basis for a good amount of statistical inference.

Some Other Things You Should Know

The curves and distributions presented in this chapter were, in many instances, somewhat abstract. Sometimes actual values or scores were represented, but other times they were not. At this point, let me call your attention to a distinction that is often made in the world of numbers—namely, the distinction between discrete and continuous distributions. The difference is perhaps best illustrated by way of examples.

Consider a variable such as the number of children in a family. Respondents to a survey might answer that they had 0, 1, 2, 3, or some other number of children. The scale of measurement is clearly interval/ratio, but the only possible responses are integer values, or whole numbers. Those are considered discrete values, and a distribution based on those values is a *discrete distribution.*

Now consider a variable such as weight. Assuming that you had a very sophisticated scale, you could conceivably obtain very refined measurements—maybe so refined that ounces could be expressed to one or more decimal places. Such a system of measurement would result in what's known as a *continuous distribution*—a distribution based on such refined measurement that one value could, in effect, blend into the next.

You should take note that curves are often stylized presentations of data. A smooth curve may not be an accurate reflection of an underlying distribution based on discrete values. An accurate representation of a discrete distribution would actually be a little jagged or bumpy, because only integer values are possible, and there is no way for one integer value to blend into the next. That said, you should also know that this is really a minor point, and it doesn't reduce the overall utility of statistical analysis.

On the technology side of the ledger, you should know that a wide variety of statistical analysis software is available, all of which can reduce the task of statistical analysis to mere button pushing if you're not careful. There's no doubt that the availability of statistical software has simplified certain aspects of statistical analysis, but an overreliance on such software can work against you in the long run. There's still no substitute for fundamental brainpower when it comes to a thorough look at your data in the form of distributions and graphs before you really get started. That's why the process of visualization remains so important.

Key Terms

frequency distribution
negative skew
normal curve
1-2-3 rule
point of inflection
positive skew
skewed distribution
symmetrical distribution
tail of the distribution

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. When a line can be drawn through the middle of a curve and both sides of the curve are mirror images of each other, the curve is said to be a _____ curve.
2. When a curve is skewed in a positive direction, it is skewed to the _____; when a curve is skewed in a negative direction, it is skewed to the _____.

3. The points on either side of a normal curve at which the curve begins to change direction are known as the points of _____.
4. In a normal distribution, the points of inflection are located _____ standard deviation(s) above and below the mean.
5. In a normal distribution, the mean, median, and mode _____.

Application Questions/Problems

1. In a normal distribution, and using the 1-2-3 Rule, approximately what percentage of the area under the curve is found between one standard deviation above and below the mean?
2. In a normal distribution, and using the 1-2-3 Rule, approximately what percentage of the area under the curve is found between two standard deviations above and below the mean?
3. In a normal distribution, and using the 1-2-3 Rule, approximately what percentage of the area under the curve is found between three standard deviations above and below the mean?
4. In a normal distribution, what percentage of the area under the curve is found above the mean? What percentage of the area under the curve is found below the mean?
5. Assume that the mean of a distribution of test scores is 62 and the standard deviation is 4. Your score on the test is 70. How many standard deviations above or below the mean is your test score?
6. Assume that the mean of a distribution of test scores is 73 and the standard deviation is 5. You've been told that your test score is one standard deviation above the mean. What is your test score?
7. Assume that the mean of a distribution of test scores is 70, with a standard deviation of 5. You've been told that your score is two standard deviations above the mean. What is your test score?
8. Assume that the mean of a distribution of test scores is 200, with a standard deviation of 30. What would be the value of the score that falls two standard deviations below the mean?
9. Assume that the mean of a distribution of scores is 1250, with a standard deviation of 300. What would be the value of a score that falls one standard deviation below the mean?