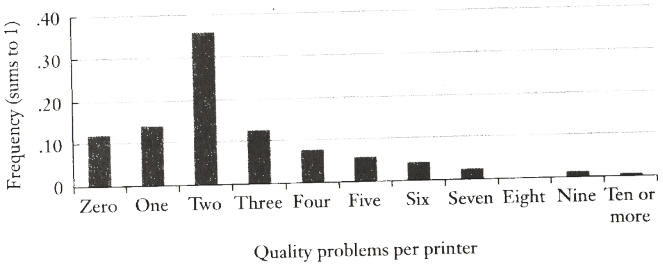
6

For this reason, we have another statistic that also signals the "middle" of a distribution, albeit differently: the median. The median is the point that divides a distribution in half, meaning that half of the observations lie above the median and half lie below. (If there is an even number of observations, the median is the midpoint between the two middle observations.) If we return to the bar stool example, the median annual income for the ten guys originally sitting in the bar is $35,000. When Bill Gates walks in with his parrot and perches on a stool, the median annual income for the eleven of them is still $35,000. If you literally envision lining up the bar patrons on stools in ascending order of their incomes, the income of the guy sitting on the sixth stool represents the median income for the group. If Warren Buffett comes in and sits down on the twelfth stool next to Bill Gates, the median still does not change.*

For distributions without serious outliers, the median and the mean will be similar. I've included a hypothetical summary of the quality data for the competitor's printers. In particular, I've laid out the data in what is known as a frequency distribution. The number of quality problems per printer is arrayed along the bottom; the height of each bar represents the percentages of printers sold with that number of quality problems. For example, 36 percent of the competitor's printers had two quality defects during the warranty period. Because the distribution includes all possible quality outcomes, including zero defects, the proportions must sum to 1 (or 100 percent).

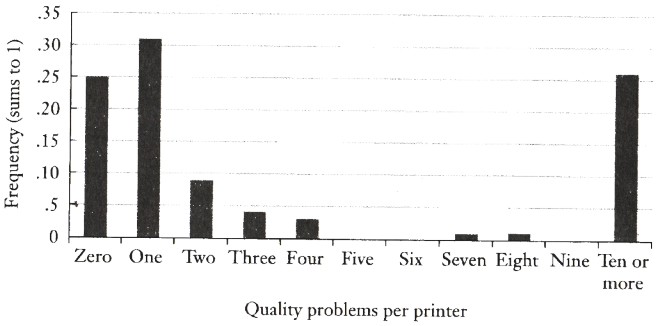
Frequency Distribution of Quality Complaints for Competitor's Printers



Because the distribution is nearly symmetrical, the mean and median are relatively close to one another. The distribution is slightly skewed to the right by the small number of printers with many reported quality defects. These outliers move the mean slightly rightward but have no impact on the median. Suppose that just before you dash off the quality report to your boss you decide to calculate the *median* number of quality problems for your firm's printers and the competition's. With a few keystrokes, you get the result. The median number of quality complaints for the competitor's printers is 2; the median number of quality complaints for your company's printers is 1.

Huh? Your firm's median number of quality complaints per printer is actually *lower* than your competitor's. Because the Kardashian marriage is getting monotonous, and because you are intrigued by this finding, you print a frequency distribution for your own quality problems.

Frequency Distribution of Quality Complaints at Your Company



What becomes clear is that your firm does not have a uniform quality problem; you have a "lemon" problem; a small number of printers have a huge number of quality complaints. These outliers inflate the mean but not the median. More important from a production standpoint, you do not need to retool the whole manufacturing process; you need only figure out where the egregiously low-quality printers are coming from and fix that.*

7

Neither the median nor the mean is hard to calculate; the key is determining which measure of the "middle" is more accurate in a particular situation (a phenomenon that is easily exploited). Meanwhile, the median has some useful relatives. As we've already discussed, the median divides a distribution in half. The distribution can be further divided into quarters, or quartiles. The first quartile consists of the bottom 25 percent of the observations; the second quartile consists of the next 25 percent of the observations; and so on. Or the distribution can be divided into deciles, each with 10 percent of the observations. (If your income is in the top decile of the American income distribution, you would be earning more than 90 percent of your fellow workers.) We can go even further and divide the distribution into hundredths, or percentiles. Each percentile represents 1 percent of the distribution, so that the 1st percentile represents the bottom 1 percent of the distribution and the 99th percentile represents the top 1 percent of the distribution.

The benefit of these kinds of descriptive statistics is that they describe where a particular observation lies compared with everyone else. If I tell you that your child scored in the 3rd percentile on a reading comprehension test, you should know immediately that the family should be logging more time at the library. You don't need to know anything about the test itself, or the number of questions that your child got correct. The percentile score provides a ranking of your child's score relative to that of all the other test takers. If the test was easy, then most test takers will have a high number of answers correct, but your child will have fewer correct than most of the others. If the test was extremely difficult, then all the test takers will have a low number of correct answers, but your child's score will be lower still.

Here is a good point to introduce some useful terminology. An "absolute" score, number, or figure has some intrinsic meaning. If I shoot 83 for eighteen holes of golf, that is an absolute figure. I may do that on a day that is 58 degrees, which is also an absolute figure. Absolute figures can usually be interpreted without any context or additional information. When I tell you that I shot 83, you don't need to know what other golfers shot that day in order to evaluate my performance. (The exception might be if the conditions are particularly awful, or if the course is especially difficult or easy.) If I place ninth in the golf tournament, that is a relative statistic. A "relative" value or figure has meaning only in comparison to something else, or in some broader context, such as compared with the eight golfers who shot better than I did. Most standardized tests produce results that have meaning only as a relative statistic. If I tell you that a

Excerpts from Wheelan, C. (2013). *Naked Statistics*. New York, NY. W. W. Norton and Company.

third grader in an Illinois elementary school scored 43 out of 60 on the mathematics portion of the Illinois State Achievement Test, that absolute score doesn't have much meaning. But when I convert it to a percentile—meaning that I put that raw score into a distribution with the math scores for all other Illinois third graders—then it acquires a great deal of meaning. If 43 correct answers falls into the 83rd percentile, then this student is doing better than most of his peers statewide. If he's in the 8th percentile, then he's really struggling. In this case, the percentile (the relative score) is more meaningful than the number of correct answers (the absolute score).

8

Another statistic that can help us describe what might otherwise be a jumble of numbers is the standard deviation, which is a measure of how dispersed the data are from their mean. In other words, how spread out are the observations? Suppose I collected data on the weights of 250 people on an airplane headed for Boston, and I also collected the weights of a sample of 250 qualifiers for the Boston Marathon. Now assume that the mean weight for both groups is roughly the same, say 155 pounds. Anyone who has been squeezed into a row on a crowded flight, fighting for the armrest, knows that many people on a typical commercial flight weigh more than 155 pounds. But you may recall from those same unpleasant, overcrowded flights that there were lots of crying babies and poorly behaved children, all of whom have enormous lung capacity but not much mass. When it comes to calculating the average weight on the flight, the heft of the 320-pound football players on either side of your middle seat is likely offset by the tiny screaming infant across the row and the six-year-old kicking the back of your seat from the row behind.

On the basis of the descriptive tools introduced so far, the weights of the airline passengers and the marathoners are nearly identical. *But they're not.* Yes, the weights of the two groups have roughly the same "middle," but the airline passengers have far more dispersion around that midpoint, meaning that their weights are spread farther from the midpoint. My eight-year-old son might point out that the marathon runners look like they all weigh the same amount, while the airline passengers have some tiny people and some bizarrely large people. The weights of the airline passengers are "more spread out," which is an important attribute when it comes to describing the weights of these two groups. The standard deviation is the descriptive statistic that allows us to assign a single number to this dispersion around the mean. The formulas for calculating the standard deviation and the variance (another common measure of dispersion from which the standard deviation is derived) are included in an appendix at the end of the chapter. For now, let's think about why the measuring of dispersion matters.

9

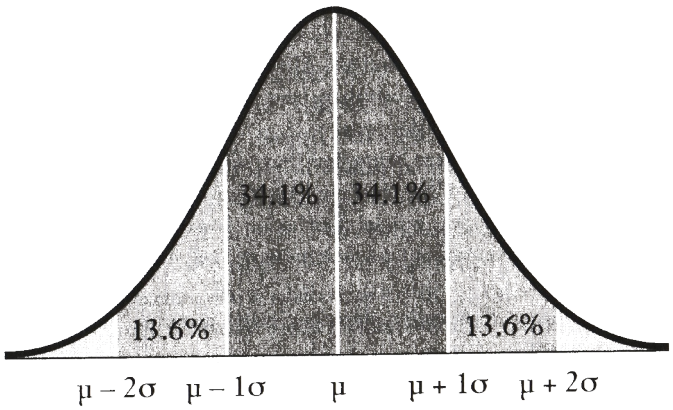
Some distributions are more dispersed than others. Hence, the standard deviation of the weights of the 250 airline passengers will be higher than the standard deviation of the weights of the 250 marathon runners. A frequency distribution with the weights of the airline passengers would literally be fatter (more spread out) than a frequency distribution of the weights of the marathon runners. Once we know the mean and standard deviation for any collection of data, we have some serious intellectual traction. For example, suppose I tell you that the mean score on the SAT math test is 500 with a standard deviation of 100. As with height, the bulk of students taking the test will be within one standard deviation of the mean, or between 400 and 600. How many students do you think score 720 or higher? Probably not very many, since that is more than two standard deviations above the mean.

In fact, we can do even better than "not very many." This is a good time to introduce one of the most important, helpful, and common distributions in statistics: the normal distribution. Data that are distributed normally are symmetrical around their mean in a bell shape that will look familiar to you.

The normal distribution describes many common phenomena. Imagine a frequency distribution describing popcorn popping on a stove top. Some kernels start to pop early, maybe one or two pops per second; after ten or fifteen seconds, the kernels are exploding frenetically. Then gradually the number of kernels popping per second fades away at roughly the same rate at which the popping began. The heights of American men are distributed more or less normally, meaning that they are roughly symmetrical around the mean of 5 feet 10 inches. Each SAT test is specifically designed to produce a normal distribution of scores with mean 500 and standard deviation of 100. According to the *Wall Street Journal*, Americans even tend to park in a normal distribution at shopping malls; most cars park directly opposite the mall entrance—the "peak" of the normal curve—with "tails" of cars going off to the right and left of the entrance.

The beauty of the normal distribution—its Michael Jordan power, finesse, and elegance—comes from the fact that we know by definition exactly what proportion of the observations in a normal distribution lie within one standard deviation of the mean (68.2 percent), within two standard deviations of the mean (95.4 percent), within three standard deviations (99.7 percent), and so on. This may sound like trivia. In fact, it is the foundation on which much of statistics is built. We will come back to this point in much great depth later in the book.

The Normal Distribution



The mean is the middle line which is often represented by the Greek letter μ. The standard deviation is often represented by the Greek letter σ. Each band represents one standard deviation.

Excerpts from Wheelan, C. (2013). *Naked Statistics*. New York, NY. W. W. Norton and Company.