

DESCRIBING UNCERTAINTY AND POPULATIONS OF DATA WITH NUMBERS AND PICTURES

Summarizing the “Center,” “Wobble,” and “Shape” of Data

INTRODUCTION

Newspaper articles, TV stories, radio updates, and Internet Web pages are full of statistics: average salaries in technology jobs are up; average weight of American children is climbing; global average temperatures are rising; median home prices are falling; markets showed more volatility this week than in previous weeks; the Dow Jones Industrial Average is down; Alex Rodriguez’s batting average is up.

It seems statistics are everywhere, and wherever they are used, writers and commentators use them to speak authoritatively. But are we always using the right statistic for the job?

WHAT’S AHEAD

In this chapter, you’ll learn:
How to use only a few numbers or one graph to represent lots of data
Statistics to measure the central tendency, variability, and graphed appearance of a dataset

The difference between mean, median, and mode
How to describe the shape of a dataset
The difference between populations and samples of data

IN THE REAL WORLD

Your real estate agent is on the phone, and she's thrilled. For months, she's been helping you look for a house. You've seen a lot of nice homes, but all of them were located in bad neighborhoods. Several times, the agent has shown you a beautiful, new house next to small, rundown homes with overgrown yards. Now your agent says she's found a great house in a neighborhood with an average home price of \$250,000. As you drive into the neighborhood, you notice promising signs. You pass two mansions—huge, million-dollar homes. Yet when you reach the street with the house for sale, your heart sinks. You find yourself among small, dilapidated houses clearly worth no more than \$75,000.

Has your broker lied? Can the neighborhood's average home price actually be \$250,000?

THE KEY CONCEPTS

Statistical analysis means summarizing a lot of information with only a few shorthand quantities or *statistics*. For example, if someone asks you how your customers feel about your current product, you could answer by explaining how each and every customer likes or doesn't like your product—but that would be very difficult to do! Statistics offer ways to describe a large set of data in a short summary using pictures or numbers.

Using statistics, we can describe three important features of a set of information (or a *dataset*): the central tendency or “center” of the data, the variability or “wobble” of the data, and the geometry or “shape” of the data. Key measures that you may have seen or heard include the following:

The center or central tendency can be described by three statistics: the average (or *mean*), the *median*, and the *mode*.

The wobble or variability of a dataset can be described by the *variance* and the *standard deviation*.

The shape or geometry of a dataset can be described by the *skewness* or *kurtosis* of data that has been represented visually on a graph or histogram.

Now let’s define how to measure these features of a dataset by considering two sets of midterm scores from a statistics courses. The scores range from 0 to 100, and we have selected 13 scores from last year and 13 from this year. Let’s examine which group did better on the midterm—this year’s class or last year’s class—by looking at the average score.

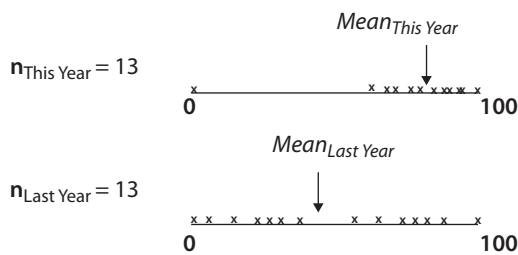
What Does *Mean* Really Mean?

A common way to summarize data is to find the average or *mean* score. The mean is the sum of all the scores divided by the number of scores. We can write this mathematically using the Greek symbol sigma (Σ) as the sum and the lowercase letter n to represent the number of scores in the dataset.

$$\text{Mean} = \frac{1}{n} \sum x_i$$

↙ Each score
↖ All
↗ Sum
↘ scores
↘ Number of
↘ scores

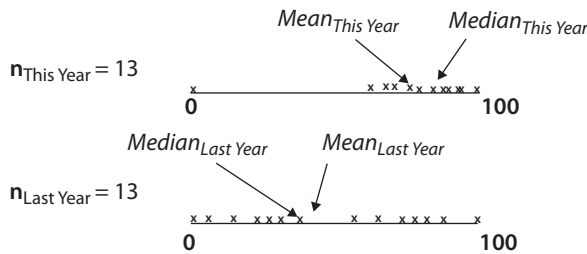
Notice that based on the mean midterm scores, this year’s group did better than last year’s group.



But the mean is not the only way to represent the center of a dataset. We can also use the *median* or “middle” value. When the dataset has been sorted, the median is the actual middle value with an odd number of data values, or the mean of the two middle values in an even number of data points. For example, in a set of 11 test scores, the 6th score is the median; but in a set of 10 scores, the median is the average of the 5th and 6th scores.

Because it sits in the middle of the dataset, the median is also often referred to as the 50th percentile, the 5th decile, and/or the 2nd quartile.

So where is the median test score for each dataset? If we have 13 test scores in each year's dataset, then the median will be the 7th test score when the results are sorted in order. Although some people use the terms interchangeably, the median is not necessarily the same as the mean—in fact, they are often different:



When we look at medians, the difference between the two classes is even greater—this year's students performed even better than last year's.

A third statistic used to describe central tendency is the *mode* of a dataset, or the most frequently occurring value. Modes are most often helpful when trying to determine which response—or score, in this case—was most frequent or popular. In the charts above, because it appears that each test score (for each dataset) is unique—that is, no score appears more than once—these distributions have no modes.

What Is the Best Statistic to Summarize Central Tendency?

Thus, we have an interesting dilemma—which statistic best describes central tendency? Is it the average? The median? Maybe the mode? If we look at news reports, we see the average used most often in stories—average income, average age, average profit, and batting averages. So is the average (or mean) the best statistic since it seems to be so common?

No! In fact, we have to be very careful with the average statistic. Because the average is the sum of all the data points divided by the number of data points, any very large or very small data point could significantly affect the average.

For example, in the charts above, note that this year’s scores are mostly high except for one zero score. This single score, which sits far away from the rest of the data points, causes the mean to be lower than it would be without this extreme value in the dataset. In fact, if we removed this score, the mean of this year’s data would likely be very close to the median score.

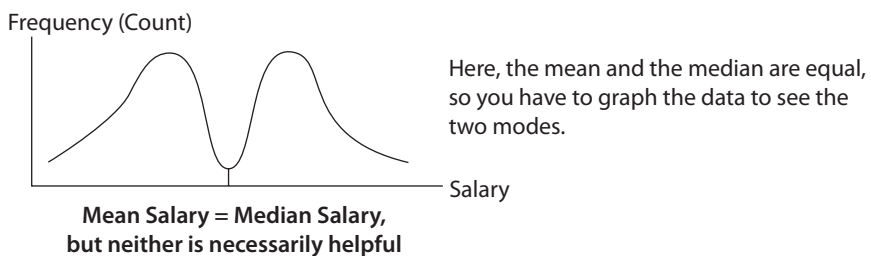
A value that is far from the rest of the dataset is called an *outlier*. Outliers, which are often determined subjectively, are values with small frequencies (small in number) that occur away from most of the other data. Thus, the presence of these outliers will have a significant effect on the mean statistic.

When it comes to describing the central tendency of a dataset, we have a choice; the average is not necessarily the best statistic.

TIP: *If a dataset has outliers, the median is often a better statistic to use than the mean. The median is less influenced by outlying values; therefore, it may be a better, more reliable descriptive statistic to use. However, if the dataset is a mirror image around its median, also called symmetric, the mean (which will equal the median) is the better statistic to use.*

But how can we determine if the mean is being heavily influenced by outliers? The simple answer is: Don’t just look at one statistic—look at several. If the mean and median are not close together, then the mean may be affected by outliers.

Is there a case where the mode might be better than the mean and median? Take a look at the example below. In this case, the horizontal x-axis is salaries of one company’s employees, and the vertical y-axis is how frequently these salaries are reported. Thus, a higher frequency means that more employees earn that particular salary amount.



On the graph for this dataset, you can see that the mean salary and median salary are very close. Thus, you might be led to think the average salary would be the better summary statistic to use.

But wait—look again. The mean and median salaries are some of the least frequently reported values. These salaries appear to be *bimodal*, which means there are two modes. (Perhaps in this case both staff and executive salaries have been collected.) Because there are two frequently occurring salaries, the mode salary values may be the best way to summarize the dataset.

We could not have determined the best statistic to report central tendency in this dataset until we graphed the data. This is a critical point!

TIP: *When deciding what statistic is most appropriate to measure central tendency of a dataset, always graph the data! Numbers and statistics are not enough; we must actually “look” at the dataset to see the complete story. “Running the numbers” to get mean, median, and mode is simply not sufficient. Always look at numbers and pictures before deciding how best to summarize a dataset.*

Rule #1 of Statistics: Graph the Data!

How do we draw a picture of a dataset? The answer is to create a distribution like the one above. A chart shows the variable we are measuring (test scores, for example) on the x-axis and the frequency of the test scores on the y-axis. Such a chart is called a *frequency distribution* or *histogram*. Such charts can be easily created in Microsoft Excel (with the use of the included Data Analysis Toolkit add-ins) and with many other popular statistical packages/programs.

HISTOGRAM

A histogram is also referred to as a *frequency distribution*.

A histogram graphs quantitative data in contiguous ranges, often called *bins* or *buckets*, along the x-axis

You can order a histogram in order of frequency; this is called a *Pareto distribution* if you order the frequency bars from highest (most frequent) to lowest.

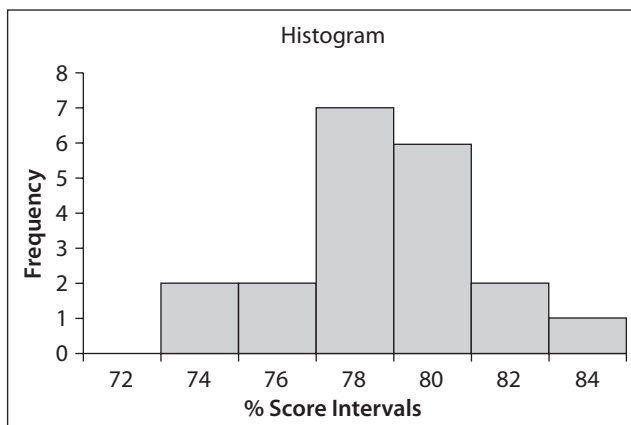
A histogram will let you “see” the descriptive statistics such as mean and median.

A histogram is *not* the same as a bar chart; histograms have continuous values along the x-axis. Bar charts do not necessarily (see example below). In a bar chart

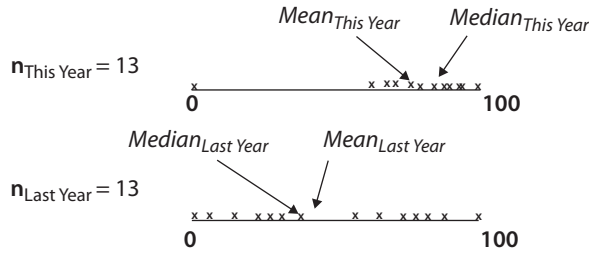
the data are not related (States not related; FL not “greater” than MO),

bin size is not a factor, and

there is no quantifiable relationship among bins.



We are trying to measure the spread or “wobble” of the dataset. Let’s return to our test scores example (see charts below). Which group—this year’s or last year’s—had more variability in its test scores? Which group had less consistent scores or less predictable scores? Let’s consider the simplest way to measure how datasets can be spread out (or narrow).



The easiest way to think about spread or volatility of a dataset is literally to consider how “wide” it is. This is called the *range* of the dataset, or the distance between the minimum and maximum values. Note that in the two datasets above, the ranges are identical; that is, the distance from the minimum score (0) to the maximum score (100) is the same range, or 100 points.

Thus, the range doesn’t help us much here. We need to look inside these end points of the data distribution to figure out which dataset is more or less variable.

Maybe one way is to find out how far away each test score is from the mean score. If a set of scores is further away from the mean, the dataset with the higher average distance from the mean should be more spread out or variable.

Let’s write this in a simple equation as we did before when we created the formula in “sigma” notation for the mean of a dataset. To express the average (signed) distance (or *deviation*) of one test score from the mean score, we must first find that score’s distance from the mean. Then we do this for all the test scores, adding up all of these deviations (some positive, some negative). Now, find the average of all these individual deviations from the mean. We can express this using the following formula:

$$\frac{1}{n} \sum (X_i - \bar{X}) \quad \leftarrow \text{average deviation from the mean}$$

This equation represents the average distance that a value, x_i , is from the mean of the dataset, μ . We add up all the deviations from the mean and then average them (or divide by the total number of scores in our dataset).

Of course, this formula always equals zero! But why? The average (signed) distance from the average is zero—on average, we are at the average. We must improve this formula slightly so that the deviations on either side of the mean don't offset each other in the aggregate. To get rid of the offsets, we could remove the signs of the deviations by taking the absolute value of the distance between a test score and the mean of all scores. But an even easier way to get rid of the signs in front of the deviations is to square them.

To do so, we can simply *square* the deviations to create the *mean squared deviation* from the mean. The further the data points are from the mean, the larger this quantity becomes. And the larger the dispersion or “wobble” of the dataset being studied! We call the mean squared deviation from the mean the statistical *variance*.

$$\frac{1}{n} \sum (x_i - \mu)^2 \quad \leftarrow \text{mean squared deviation} = \text{variance (in units squared)}$$

But along comes a problem with variance. This quantity is in units of the data, squared. Wouldn't it be better to use a spread or wobble statistic that is expressed in the same units of the data we are studying, just like the mean, median, and mode?

The answer is yes. So how do we get variance into a form that is in units of the data? Simple—just take the square root of the variance. This will give the root mean squared deviation from the mean, or *standard deviation* of the dataset.

$$\sqrt{\frac{1}{n} \sum (x_i - \mu)^2} \quad \leftarrow \begin{aligned} &\text{root mean squared deviation} \\ &= \text{square root of variance} \\ &= \text{standard deviation (units of the data)} \end{aligned}$$

This quantity is in units of the data now (or in units of points when considering the midterm scores problem, above). This standard deviation quantity is the preferred measure of the variability of a dataset, because it's in the units of the data and we don't have to worry about potential mathematical issues with the absolute value function.

Using our two examples above, although the two datasets have the same range, this year’s test scores are *less variable* or *more consistent* than last year’s scores. Note that this year’s scores cluster more closely around the mean—save for one outlier—while last year’s scores are somewhat all over the place. The standard deviation, then, of this year’s test scores will be lower—and will indicate less wobble or variability in the data—than the standard deviation of last year’s test scores.

How Can We Describe the Shape of a Dataset?

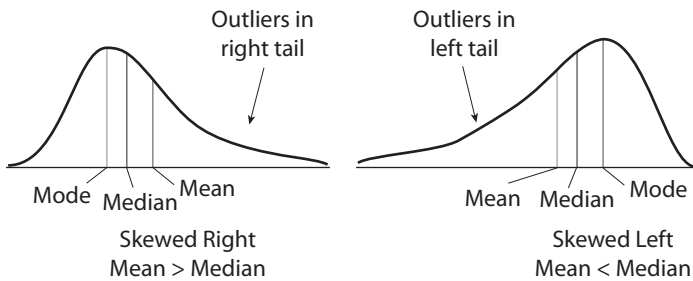
Because we now have a way to graph the data—using the histogram—we can use some additional statistics to describe the “shape” of the data distribution. Recall the effect that outliers can have on the mean. If a high-value outlier exists—like one or two perfect test scores—and these scores are well above most others, the mean test score will get pulled or *skewed* to the right (see chart at left, below). On the other hand, if most scores are high but a few low scores exist, the effect is to skew the mean to the left (see chart below).

However, if no significant outliers or long “tails” exist at either end of the data distribution, the graph is said to be *not skewed*. Note in the middle picture, below, that in a symmetric distribution, the mean and median are equal to each other. In both the right- and left-skew cases, note that the skewness or *tail* of the distribution is in the direction of the outliers. Right-skewed distributions are *right-tailed* while left-skewed distributions are *left-tailed*.

SKEWNESS		
Skewness is the measurement of the deviation of the distribution from symmetry. It is defined by the location of the outliers—a symmetric distribution has no skewness. Skewness can be positive or negative.		
Rightward Skewness	Symmetry	Leftward Skewness
Skewness > 0	Skewness = 0	Skewness < 0

Thus, the *skewness statistic* will be positive when the distribution is skewed to the right (and the mean is greater than the median) and

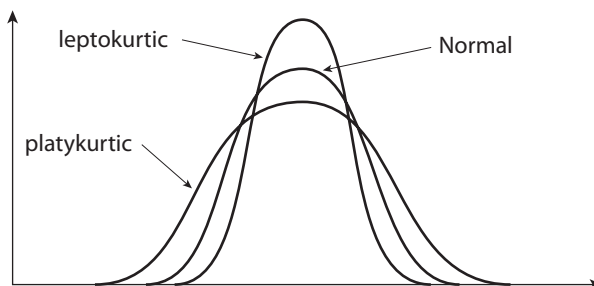
negative when the distribution is skewed to the left (and the mean is less than the median). Note that a symmetric distribution has no or zero skewness. Therefore, if we know the skewness, we can immediately know whether the mean is to the left or right of the median. And as we've learned, in skewed distributions, the median is often the preferred statistic because it is less susceptible to skewing effects.



But what about the “hump” in the distribution? Is there a way to measure how pointed or flat the peak of the distribution is?

The answer lies in the *kurtosis statistic*. Kurtosis measures a distribution’s “peakedness,” the height of the hump. The narrower the hump, the more compact are the data points in relation to the mean. A highly kurtotic curve is also said to have *fat tails* due to more outlier behavior further from the mean (caused by the higher, thinner peak in the middle).

In most statistical programs, a descriptive statistics tool is available that can produce several standard descriptive statistics for a dataset. From these tools, a kurtosis of three is the dividing line between a tall, peaked distribution (*leptokurtic*) and a flatter distribution (*platykurtic*). A kurtosis of exactly three describes a bell-shaped or *normal* distribution. The bell curve is said to have no kurtosis or is *mesokurtic* even though formulaically the kurtosis is equal to three.



Note that in some tools, like Microsoft Excel, three is subtracted from the kurtosis formula so that a bell-shaped distribution has zero kurtosis. In general, however, a kurtosis of three is the bell curve and the dividing point between leptokurtic (high) and platykurtic (low) distributions. See the following table for a comparison of kurtosis results from a popular tool, STATA, and from Microsoft Excel.

STATA	Description	Excel
Greater than 3	Peaked, leptokurtic	Greater than 0
Less than 3	Flatter, platykurtic	Less than 0
Equal to 3	Bell-shaped or normal, mesokurtic	Equal to 0

What Is the Difference between Populations and Samples of Data?

A *population* of data is essentially the entire set of individuals or objects of a particular group, such as all students in MBA programs worldwide. A *sample*, on the other hand, is just a part of a population, such as African-American females over 25 years of age in this year’s entering MBA class at a particular school.

In general, getting populations of data—and summary descriptive statistics—is difficult, if not impossible. In some cases, population data are possible, such as in a drug study where only 100 people in the world have tried a new cancer treatment. This is not usually the case in business. We cannot survey every customer, audit every transaction, or inspect every product we produce. Thus, we perform *sampling*, and samples are used to infer population behaviors. The average overdue amount for credit cards held by a sample of cardholders age 25 to 35 may be used to infer how the population—all cardholders in this age range—behaves in terms of average overdue amount. So why do we sample?

Cost: Only need enough data to be “sufficiently” accurate.

Accuracy: Maintain better control of data collection errors.

Timeliness: Business decisions are made in real time; we cannot wait for perfect information.

Amount of information: Process and model more detailed information using samples to get to population inferences more quickly and efficiently.

Destructive testing: If we need to destroy a product we are inspecting to test its quality, we do not want to destroy the entire population for one analysis.

Therefore, it becomes important to differentiate population statistics (often referred to as *parameters*) from sample statistics (measurements or descriptions made of a sample). When we write statistical formulas, we generally use Greek symbols for population parameters and Roman characters for sample statistics as in the Symbol Summary below.

Symbol Summary		
Descriptive Statistic	Symbol for Sample (A piece of the population)	Symbol for Population (All data over all time, categories, etc.)
Mean	\bar{x}	μ
Variance	s^2	σ^2
Standard Deviation	s	σ
Median	$x_{0.5}$	$\mu_{0.5}$

STATISTICS IN ACTION

Look at the two investment funds below:

<u>MBA Student Fund</u>	<u>Faculty Fund</u>
Average return over 10 years = 5%	Average return over 10 years = 5%
Median return = 7%	Median return = 2%
Standard Deviation = 10%	Standard Deviation = 1%

In which fund would you invest? Note that the average returns over the past ten years are the same; does that mean that the funds are equally good investments? Not at all—don't consider just the average. Notice that

the median return is higher for the student fund. Perhaps a bad year or two has pulled down the average return while the median return has stayed relatively high.

The opposite is true for the faculty fund. The mean is higher than the median. This may mean that a very good year or two has pulled up the average but left the median (a lowly 2 percent return) relatively unchanged.

It's still not clear which is the better investment. Maybe we should look at the volatility of the returns or “risk” of each fund (as measured by volatility or variability from the average return). This may help us decide: the faculty fund has only a 1 percent standard deviation (or risk) as compared to 10 percent for the student fund.

If we like steady, predictable returns, we want to pick the less volatile faculty fund (even though a good year may have affected the average return quite a bit). But if we are looking for higher returns—and we are willing to assume the higher risk that comes with them—perhaps the student fund is a better choice.

TEST YOURSELF

Look at the data below on student expenditures. These data are student expenditures per capita for every state in the United States.

State	Expenditure Per Student	State	Expenditure Per Student
NJ	\$12,428.71	SC	\$7,988.55
NY	\$12,087.10	WA	\$7,825.26
CT	\$11,530.56	TX	\$7,799.66
AK	\$10,395.06	IA	\$7,799.49
DE	\$10,242.85	MO	\$7,601.18
VT	\$9,791.73	CO	\$7,561.11
MA	\$9,758.17	NV	\$7,540.11
MI	\$9,706.44	FL	\$7,539.73
RI	\$9,604.36	NC	\$7,480.77
PA	\$9,436.55	HI	\$7,421.80

Continued

Continued

State	Expenditure Per Student	State	Expenditure Per Student
WI	\$9,343.26	KS	\$7,412.11
MD	\$9,248.83	SD	\$7,333.20
MN	\$9,138.28	NM	\$7,309.57
ME	\$9,004.43	MT	\$7,182.06
IL	\$8,989.23	AL	\$6,744.91
OH	\$8,732.32	TN	\$6,711.61
WY	\$8,608.02	ND	\$6,700.84
IN	\$8,544.66	LA	\$6,614.54
OR	\$8,328.51	AZ	\$6,595.01
VA	\$8,298.51	OK	\$6,491.74
CA	\$8,167.40	KY	\$6,403.51
WV	\$8,162.72	ID	\$6,257.05
NE	\$8,157.35	AR	\$6,117.58
NH	\$8,107.08	MS	\$5,669.75
GA	\$8,086.15	UT	\$5,571.30

1. Why are the data presented as expenditures per student and not as total, statewide numbers?
2. Describe the data in terms of the basic descriptive statistics (mean, median, and so on). Are the data normally distributed? Why or why not?
3. Create a histogram of the data. Describe the histogram. Are there any outliers? Which states/expenditures appear to be outliers?
4. What bin size did you choose for your histogram and why? Can you compare histograms of different bin sizes?

KEY POINTS TO REMEMBER

In this chapter, we have examined some basic descriptive statistics—in both number and chart form—that are commonly used to represent a great deal of data succinctly:

Central tendency

Mean: The sum of all the scores divided by the number of scores

Median: The actual middle value in an odd number of data values or the mean of the two middle values in an even number of data points

Mode: The most frequently occurring value

Variability

Variance: The measure of statistical dispersion in a dataset in units of data squared

Standard deviation: The measure of statistical dispersion in a dataset in units of the data

Geometry

Skewness: The direction and relative magnitude the mean is pulled and/or the direction the tail of a graphed dataset is pulled

Kurtosis: The measure of how peaked or “pointy” a data distribution is in a graph. Also a measure of how “fat” the tails of the distribution are.

