Chapter 8: Statistics

Figure 1: Statistics can be used to decide on a fair salary for sports stars.

Before the 2021 WNBA season, professional basketball player Candace Parker signed a contract with the Chicago Sky, which entitled her to a salary of $190,000. This amount was the 23rd highest in the league at the time. How did the team's management decide on her salary? They likely considered some intangible qualities, like her leadership skills. However, much of their deliberations probably took into account her performance on the court. For example, Parker led the league in rebounds in the 2020 season (214 of them) and scored 14.7 points per game (which ranked her 18th among all WNBA players). Further, Parker brought 13 seasons of experience to the team. All of these factors played a role in deciding the terms of her contract.

Estimating the value of one variable (like salary) based on other, measurable variables (points per game, experience, rebounds, etc.) is among the most important applications of statistics, which is the mathematical field devoted to gathering, organizing, summarizing, and making decisions based on data.

8.1 Gathering and Organizing Data

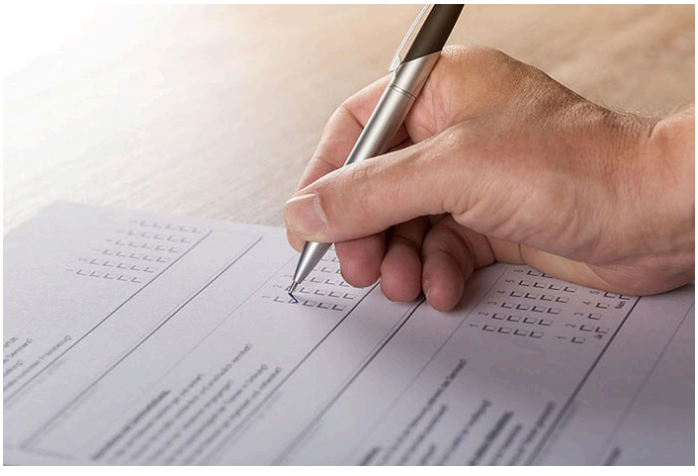


Figure 2: Surveys are commonly used to gather data.

Learning Objectives

After completing this section, you should be able to:

1. Distinguish among sampling techniques.
2. Organize data using an appropriate method.
3. Create frequency distributions.

When a polling organization wants to try to establish which candidate will win an upcoming election, the first steps are to write questions for the survey and to choose which people will be asked to respond to the survey. These can seem like simple steps, but they have far-reaching implications in the analysis the pollsters will later carry out. The process by which **samples** (or groups of **units** from which we collect data) are chosen can strongly affect the **data** that are collected. Units are anything that can be measured or surveyed (such as people, animals, objectives, or experiments) and data are observations made on units.

One of the most famous failures of good sampling occurred in the first half of the twentieth century. *The Literary Digest* was among the most respected magazines of the early twentieth century. Despite the name, the *Digest* was a weekly newsmagazine. Starting in 1916, the *Digest* conducted a poll to try to predict the winner of each US Presidential election. For the most part, their results were good; they correctly predicted the outcome of all five elections between 1916 and 1932. In 1936, the incumbent President Franklin Delano Roosevelt faced Kansas governor Alf Landon, and once again the *Digest* ran their famous poll, with results published the week before the election. Their conclusion? Landon would win in a landslide, 57% to 43%. Once the actual votes had been counted, though, Roosevelt ended up with 61% of the popular vote, 18% more than the poll predicted. What went wrong?

The short answer is that the people who were chosen to receive the survey (over ten million of them!) were not a good representation of the **population** of voting adults. The sample

was chosen using the *Digest's* own base of subscribers as well as publicly available lists of people that were likely adults (and therefore eligible to vote), mostly phone books and vehicle registration records. The pollsters then mailed every single person on these lists a survey. Around a quarter of those surveys were returned; this constituted the sample that was used to make the *Digest*'s disastrously incorrect prediction. However, the *Digest* made an error in failing to consider that the election was happening during the Great Depression, and only the wealthy had disposable income to spend on telephone lines, automobiles, and magazine subscriptions. Thus, only the wealthy were sent the *Digest*'s survey. Since Roosevelt was extremely popular among poorer voters, many of Roosevelt's supporters were excluded from the *Digest*'s sample.

Another more complicated factor was the low response rate; only around 25% of the surveys were returned. This created what's called a non-response bias.

Sampling and Gathering Data

The *Digest's* failure highlights the need for what is now considered the most important criterion for sampling: randomness. This randomness can be achieved in several ways. Here we cover some of the most common.

A **simple random sample** is chosen in a way that every unit in the population has an equal chance of being selected, and the chances of a unit being selected do not depend on the units already chosen. An example of this is choosing a group of people by drawing names out of a hat (assuming the names are well-mixed in the hat).

A **systematic random sample** is selected from an ordered list of the population (for example, names sorted alphabetically or students listed by student ID). First, we decide what proportion of the population will be in our sample. We want to express that proportion as a fraction with 1 in the numerator. Let's call that number D. Next, we'll choose a random number between one and D. The unit at that position will go into our sample. We'll find the rest of our sample by choosing every Dth unit in the list, starting with our random number.

To walk through an example, let's say we want to sample 2% of the population: $2\% = \frac{2}{100} = \frac{1}{50}$. (Note: If the number in the denominator isn't a whole number, we can just round it off. This part of the process doesn't have to be precise.) We can then use a random number generator to find a random number between 1 and 50; let's use 31. In our example, our sample would then be the units in the list at positions 31, 81 (31 + 50), 131 (81 + 50), and so forth.

A **stratified sample** is one chosen so that particular groups in the population are certain to be represented. Let's say you are studying the population of students in a large high school (where the grades run from 9th to 12th), and you want to choose a sample of 12 students. If you use a simple or systematic random sample, there's a pretty good chance that you'll miss one grade completely. In a stratified sample, you would first divide the population into groups (the *strata*), then take a random sample within each *stratum* (that's the singular form of "strata"). In the high school example, we could divide the population into grades, then take a random sample of three students within each grade. That would get us to the 12 students we need while ensuring coverage of each grade.

A **cluster sample** is a sample where clusters of units are chosen at random, instead of choosing individual units. For example, if we need a sample of college students, we may take a list of all the course sections being offered at the college, choose three of them at random (the sections are the clusters), and then survey all the students in those sections. A sample like this one has the advantage of convenience: If the survey needs to be administered in person, many of your sample units will be located in one place at the same time.

Example 1 **Random Sampling**

For each of the following situations, identify whether the sample is a simple random sample, a systematic random sample, a stratified random sample, a cluster random sample, or none of these.

1. A postal inspector wants to check on the performance of a new mail carrier, so she chooses four streets at random among those that the carrier serves. Each household on the selected streets receives a survey.
2. A hospital wants to survey past patients to see if they were satisfied with the care they received. The administrator sorts the patients into groups based on the department of the hospital where they were treated (ICU, pediatrics, or general), and selects patients at random from each of those groups.
3. A quality control engineer at a factory that makes smartphones wants to figure out the proportion of devices that are faulty before they are shipped out. The phones are currently packed in boxes for shipping, each of which holds 20 devices. The engineer wants to sample 100 phones, so he selects five crates at random and tests every phone in those five crates.
4. A newspaper reporter wants to write a story on public perceptions on a project that will widen a congested street. She stands on the side of the street in question and interviews the first five people she sees there.
5. An executive at a streaming video service wants to know if her subscribers would support a second season of a new show. She gets a list of all the subscribers who have watched at least one episode of the show, and uses a random number generator to select a sample of 50 people from the list.
6. An agent for a state's Department of Revenue is in charge of selecting 100 tax returns for audit. He has a list of all of the returns eligible for audit (about 12,000 in all), sorted by the taxpayer's ID number. He asks a computer to give him a random number between 1 and 120; it gives him 15. The agent chooses the 15th, 135th, 255th, 375th, and every 120th return after that to be audited.

Solution

To decide which type of random sample is being used in each of these, we need to focus on how the randomization is being incorporated.

1. The surveys are being given to households, so households are the units in this case. But households aren't being chosen randomly; instead, *streets* are being chosen at random. These form clusters of units, so this is a cluster random sample.
2. In this case, the administrator isn't selecting patients at random from the entire list of patients. Instead, she is choosing at random from the patients who were in each of the departments (ICU, pediatrics, general) separately. The departments form *strata*, so this is a stratified random sample.
3. The engineer is testing whether the phones are faulty, so those are the units. But the random process is being used to select the *crates* of phones. Those crates form clusters, so this is a cluster random sample.
4. The reporter isn't using a random process at all, so this sample doesn't belong to any of the types we have been talking about. A sample like this one is sometimes described as a *convenience sample*, and shouldn't be used in a statistical setting.
5. The executive is choosing her sample completely at random from the full population, so this is a simple random sample.
6. The agent is choosing from the full population, but is only choosing the first unit for the sample at random; the rest are chosen by skipping down the list systematically. Thus, this is a systematic random sample.

PEOPLE IN MATHEMATICS
George Gallup

Figure 3: George Gallup was a founder of survey sampling techniques, and his legacy lives on to this day.

George Gallup (1901–1984) rose to fame in 1936 when his prediction of the percentage of the vote going to each candidate in that year's U.S. Presidential election was more accurate than the one published in *Literary Digest*, and he did so using a sample that was much smaller than the *Digest*. He even took it one step farther, predicting with high accuracy the erroneous results of the poll that the *Literary Digest* would end up publishing! Gallup's theories on public opinion polling essentially created that field. In 1948, Gallup's reputation took a bit of a hit, when he famously, but incorrectly, predicted that Thomas Dewey would beat incumbent Harry Truman in that year's Presidential election. Over the following decades, however, public trust in Gallup's polls recovered and even steadily increased. The company Gallup founded continues to conduct daily public opinion polls, as well as provides consulting services for businesses.

Organizing Data

Once data have been collected, we turn our attention to analysis. Before we analyze, though, it's useful to reorganize the data into a format that makes the analysis easier. For example, if

our data were collected using a paper survey, our *raw data* are all broken down by respondent (represented by an individual response sheet). To perform an analysis on all the responses to an individual question, we need to first group all the responses to each question together. The way we organize the data depends on the type of data we've collected.

There are two broad types of data: categorical and quantitative. **Categorical data** classifies the unit into a group (or category). Examples of categorical data include a response to a yes-or-no question, or the color of a person's eyes. **Quantitative data** is a numerical measure of a property of a unit. Examples of quantitative data include the time it takes for a rat to run through a maze or a person's daily calorie intake. We'll look at each type of data in turn when considering how best to organize.

Categorical Data Organization

The best way to organize categorical data is using a **categorical frequency distribution**. A categorical frequency distribution is a table with two columns. The first contains all the categories present in the data, each listed once. The second contains the *frequencies* of each category, which are just a count of how often each category appears in the data.

Example 2 **Creating a Categorical Frequency Distribution**

A teacher records the responses of the class (28 students) on the first question of a multiple choice quiz, with five possible responses (A, B, C, D, and E):

A	A	C	A	B	B	A	E	A	C	A	A	A	C
E	A	B	A	A	C	A	B	E	E	A	A	C	C

Create a categorical frequency distribution that organizes the responses.

Solution

Step 1: For each possible response, count the number of times that response appears in the data. In the responses for this class, "A" appears 14 times, "B" 4 times, "C" 6 times, "D" 0 times, and "E" 4 times.

Step 2: Make a table with two columns. The first column should be labeled so that the reader knows what the responses mean, and the second should be labeled "Frequency."

Response to First Question	Frequency
A	14
B	4
C	6
D	0
E	4

Step 3: Check your work. If you add up your frequencies, you should get the same number as the total number of responses. Twenty-eight students answered that first question, and $14 + 4 + 6 + 0 + 4 = 28$.

Quantitative Data

We have a couple of options available for organizing quantitative data. If there are just a few possible responses, we can create a frequency distribution just like the ones we made for categorical data above. For example, if we're surveying a group of high school students and we ask for each student's age, we'll likely only get whole-number responses between 13 and 19. Since there are only around seven (and likely fewer) possible responses, we can treat the data as if they're categorical and create a frequency distribution as before.

Example 3 **Creating a Quantitative Frequency Distribution**

Attendees of a conflict resolution workshop are asked how many siblings they have. The responses are as follows:

1	0	1	1	2	0	3	1	1	4	1	2	0	1	3
1	2	1	2	4	1	0	1	3	0	1	2	2	1	5

Create a frequency distribution to organize the responses.

Solution

Step 1: Count the number of times you see each unique response: "0" appears 5 times, "1" appears 13 times, "2" appears 6 times, "3" appears 3 times, "4" appears twice, and "5" appears once.

Step 2: Make a table with two columns. The first column should be labeled so that the reader knows what the responses mean, and the second should be labeled "Frequency." Then fill in the results of our count.

Number of Siblings	Frequency	Number of Siblings	Frequency
0	5	3	3
1	13	4	2
2	6	5	1

Step 3: Check your work. If you add up your counts, you should get the same number as the total number of responses. Looking back at the raw data, there were 30 responses, and $5 + 13 + 6 + 3 + 2 + 1 = 30$.

If there are many possible responses, a frequency distribution table like the ones we've seen so far isn't really useful; there will likely be many responses with a frequency of one, which means the table will be no better than looking at the raw data. In these cases, we can create a **binned**

frequency distribution. A binned frequency distribution groups the data into ranges of values called *bins*, then records the number of responses in each bin.

For example, if we have height data for individuals measured in centimeters, we might create bins like 150–155 cm, 155–160 cm, and so forth (making sure that every data value falls into a bin). We must be careful, though; in this scenario, it's not clear which bin would contain a response of 155 cm. Usually, responses on the edge of a bin are placed in the higher bin, but it's good practice to make that clear. In cases where responses are rounded off, you can avoid this issue by leaving a gap between the bins that couldn't contain any responses. In our example, if the measurements were all rounded off to the nearest centimeter, we could make bins like 150–154 cm, 155–159 cm, etc. (since a response like 154.2 isn't possible). We'll use this method going forward. How do we decide what the boundaries of our bins should be? There's no one right way to do that, but there are some guidelines that can be helpful.

1. Every data value should fall into exactly one bin. For example, if the lowest value in our data is 42, the lowest bin should not be 45–49.
2. Every bin should have the same width. Note that if we shift the upper limits of our bins down a bit to avoid ambiguity (like described above), we can't simply subtract the lower limit from the upper limit to get the bin width; instead, we subtract the lower limit of the bin from the lower limit of the *next* bin. For example, if we're looking at GPAs rounded to the nearest hundredth, we might choose bins like 2.00–2.24, 2.25–2.49, 2.50–2.74, etc. These bins all have a width of 0.25.
3. If the minimum or maximum value of the data falls right on the boundary between two bins, then it's OK to bend the rule just a little in order to avoid having an additional bin containing just that one value. We'll see an example of this in just a moment.
4. If we have too many or too few bins, it can be difficult to get a good sense of the distribution. Seven or eight bins is ideal, but that's not a firm rule; anything between five and twelve is fine. We often choose the number of bins so that the widths are round numbers.

Example 4 **Creating a Binned Frequency Distribution**

The GPAs of students enrolled in an advanced sociology class are listed in the following table. At this institution, 4.00 is the maximum possible GPA.

3.93	3.43	2.87	2.51	2.70	1.91	2.32	2.85	3.06	3.03	3.49	1.84	3.72	2.56
1.99	3.40	3.74	3.23	1.98	3.05	1.43	2.90	1.20	3.72	3.56	3.07	2.58	4.00
2.79	3.81	2.60	3.69	2.88	3.34	1.51	3.63	3.45	1.89	2.30	2.98	3.04	2.70

Create a binned frequency distribution for the data.

Solution

Step 1: Identify the max and min values in your bins. Looking at the dataset, you can see that the lowest value is 1.20, and the highest is 4.00.

Step 2: Get a rough idea of bin widths. Aim for seven or eight bins, give or take a couple. For eight bins, the minimum width can be found by taking the difference between the largest and smallest data values and dividing by the number of bins:

$$\frac{\text{maximum} - \text{minimum}}{\#\text{ of bins}} = \frac{4.00 - 1.20}{8} = 0.35\,.$$

If we use 0.35 for our widths, starting at our minimum value of 1.20, we'll get bins with these boundaries: 1.20, 1.55, 1.90, 2.25, 2.60, 2.95, 3.30, 3.65, 4.00.

Step 3: Consider the context of the values. Because these are GPAs, there are natural breaks at 2.00 and 3.00 that are important. (People like whole numbers!) Since 0.35 is very close to $\frac{1}{3}$, let's use that for our bin width instead, and make sure that whole numbers fall on the boundaries. That means our first bin needs to start at 1.00 and go up to 1.33 to make sure our minimum value is included. The next bin will run from 1.34 to 1.66, and so forth.

Step 4: Create the distribution table. We start our distribution table by filling in the bins:

GPA Range	Frequency	GPA Range	Frequency	GPA Range	Frequency
1.00–1.33		2.00–2.33		3.00–3.33	
1.34–1.66		2.34–2.66		3.34–3.66	
1.67–1.99		2.67–2.99		3.67–4.00	

Notice that the last bin doesn't follow the pattern; since our maximum data value is right on the upper boundary of that last bin, this is a case where we can bend that rule just a little to avoid creating a bin for 4.00–4.33 (which wouldn't really make sense in the context of these GPAs anyway, since 4.00 is the maximum possible GPA).

Step 5: Complete the table with the frequencies. Finish the table by counting the number of data values that fall in each bin, and recording them in the frequency column:

GPA Range	Frequency	GPA Range	Frequency	GPA Range	Frequency
1.00–1.33	1	2.00–2.33	2	3.00–3.33	6
1.34–1.66	2	2.34–2.66	4	3.34–3.66	7
1.67–1.99	5	2.67–2.99	8	3.67–4.00	7

Step 6: Check your work. Add up the frequencies to make sure all the data values are included. We started with forty-two data values, and $1 + 2 + 5 + 2 + 4 + 8 + 6 + 7 + 7 = 42$.

Key Terms

- sample

- units
- data
- population
- simple random sample
- systematic random sample
- stratified sample
- cluster sample
- quantitative data
- categorical data
- categorical frequency Distribution
- binned frequency distribution

Key Concepts

- Categorical data places units into groups (categories), while quantitative data is a numerical measure of a property of a unit.
- The sampling method for a study depends on the way that randomization is used to select units for the sample.
- Frequency distributions help to summarize data by counting the number of units that fall into a particular category or range of quantitative values.

8.2 Visualizing Data

Figure 4: Data visualizations can help people quickly understand important features of a dataset.

Learning Objectives

After completing this section, you should be able to:

1. Create charts and graphs to appropriately represent data.
2. Interpret visual representations of data.
3. Determine misleading components in data displayed visually.

Summarizing raw data is the first step we must take when we want to communicate the results of a study or experiment to a broad audience. However, even organized data can be difficult to read; for example, if a frequency table is large, it can be tough to compare the first row to the last row. As the old saying goes: a picture is worth a thousand words (or, in this case, summary statistics)! Just as our techniques for organizing data depended on the type of data we were looking at, the methods we'll use for creating visualizations will vary. Let's start by considering categorical data.

Visualizing Categorical Data

If the data we're visualizing is categorical, then we want a quick way to represent graphically the relative numbers of units that fall in each category. When we created the frequency distributions in the last section, all we did was count the number of units in each category and record that number (this was the *frequency* of that category). Frequencies are nice when we're organizing and summarizing data; they're easy to compute, and they're always whole numbers. But they can be difficult to understand for an outsider who's being introduced to your data.

Let's consider a quick example. Suppose you surveyed some people and asked for their favorite color. You communicated your results using a frequency distribution. Jerry is interested in data on favorite colors, so he reads your frequency distribution. The first row shows that twelve people indicated green was their favorite color. However, Jerry has no way of knowing if that's a lot of people without knowing how many people total took your survey. Twelve is a pretty significant number if only twenty-five people took the survey, but it's next to nothing if you recorded a thousand responses. For that reason, we will often summarize categorical data not with frequencies, but with **proportions**. The proportion of data that fall into a particular category is computed by dividing the frequency for that category by the total number of units in the data.

$$\text{Proportion of a category} = \frac{\text{Category frequency}}{\text{Total number of data units}}$$

Proportions can be expressed as fractions, decimals, or percentages.

Example 1 **Finding Proportions**

Recall Example 2 in Section 8.1, in which a teacher recorded the responses on the first question of a multiple choice quiz, with five possible responses (A, B, C, D, and E). The raw data was as follows:

A	A	C	A	B	B	A	E	A	C	A	A	A	C
E	A	B	A	A	C	A	B	E	E	A	A	C	C

We computed a frequency distribution that looked like this:

Response to First Question	Frequency
A	14
B	4
C	6
D	0
E	4

$$\text{Proportion of a category} = \frac{\text{Category frequency}}{\text{Total number of data units}}$$

Now, let's compute the proportions for each category.

Solution

Step 1: In order to compute a proportion, we need the frequency (which we have in the table above) and the total number of units that are represented in our data. We can find that by adding up the frequencies from all the categories: $14 + 4 + 6 + 0 + 4 = 28$.

Step 2: To find the proportions, we divide the frequency by the total. For the first category ("A"), the proportion is $\frac{14}{28} = \frac{1}{2} = 0.5 = 50\%$. We can compute the other proportions similarly, filling in the rest of the table:

Response to First Question	Frequency	Proportion
A	14	$\frac{14}{28} = 50\%$
B	4	$\frac{4}{28} = 14.3\%$
C	6	$\frac{6}{28} = 21.4\%$
D	0	$\frac{0}{28} = 0\%$
E	4	$\frac{4}{28} = 14.3\%$

Step 3: Check your work: If you add up your proportions, you should get 1 (if you're using fractions or decimals) or 100% (if you're using percentages). In this case, $50\% + 14.3\% + 21.4\% + 0\% + 14.3\% = 100\%$.

> **CHECKPOINT**
> *If you need to round off the results of the computations to get your percentages or decimals, then the sum might not be exactly equal to 1 or 100% in the end due to that rounding error.*

Now that we can compute proportions, let's turn to visualizations. There are two primary visualizations that we'll use for categorical data: bar charts and pie charts. Both of these data

representations work on the same principle: If proportions are represented as areas, then it's easy to compare two proportions by assessing the corresponding areas. Let's look at bar charts first.

Bar Charts

A **bar chart** is a visualization of categorical data that consists of a series of rectangles arranged side-by-side (but not touching). Each rectangle corresponds to one of the categories. All of the rectangles have the same width. The height of each rectangle corresponds to either the number of units in the corresponding category or the proportion of the total units that fall into the category.

Example 2 **Building a Bar Chart**

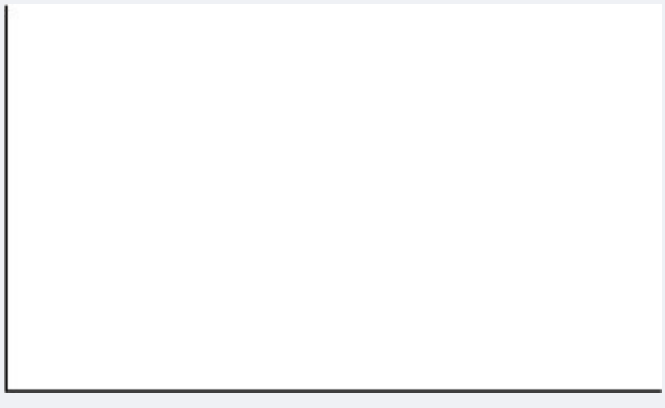
In Example 1, we computed the following proportions:

Response to First Question	Frequency	Proportion
A	14	50%
B	4	14.3%
C	6	21.4%
D	0	0%
E	4	14.3%

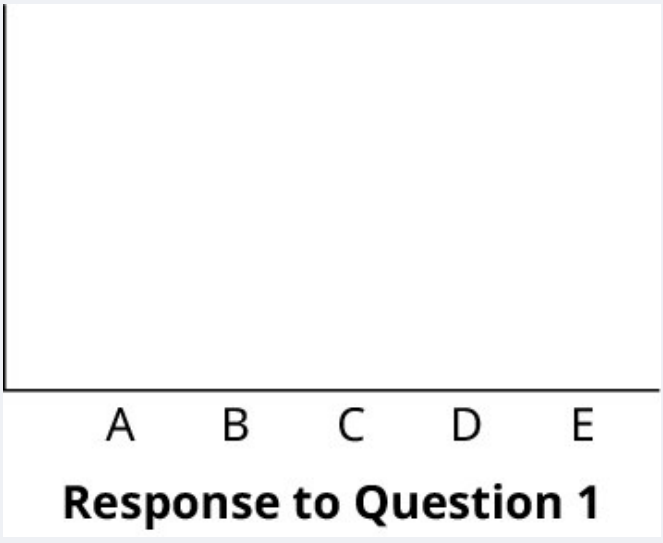
Draw a bar chart to visualize this frequency distribution.

Solution

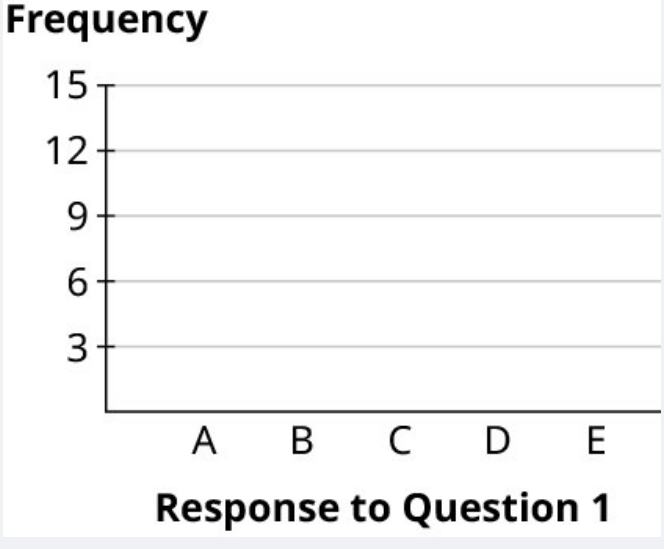
Step 1: To start, we'll draw axes with the origin (the point where the axes meet) at the bottom left:



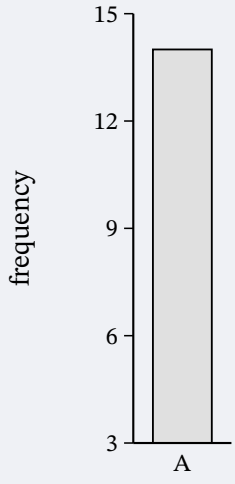
Step 2: Next, we'll place our categories evenly spaced along the bottom of the horizontal axis. The order doesn't really matter, but if the categories have some sort of natural order (like in this case, where the responses are labeled A to E), it's best to maintain that order. We'll also label the horizontal axis:



Step 3: Now, we have a decision to make: Will we use frequencies to define the height of our rectangles, or will we use proportions? Let's try it both ways. First, let's use frequencies. Notice that our frequencies run from zero to 14; this will correspond to the scale we put on the vertical axis. If we put a tick mark for every whole number between 0 and 14, the result will be pretty crowded; let's instead put a mark on the multiples of 3 or 5:

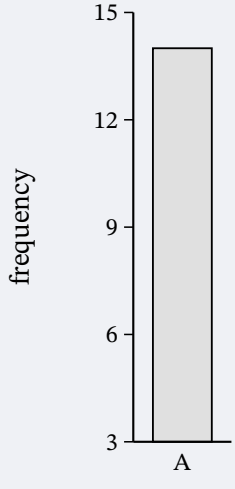


Step 4: Now, let's draw in the first rectangle. The frequency associated with "A" is 14. So we'll go to 14 on the vertical axis, and place a mark at that height above the "A" label:



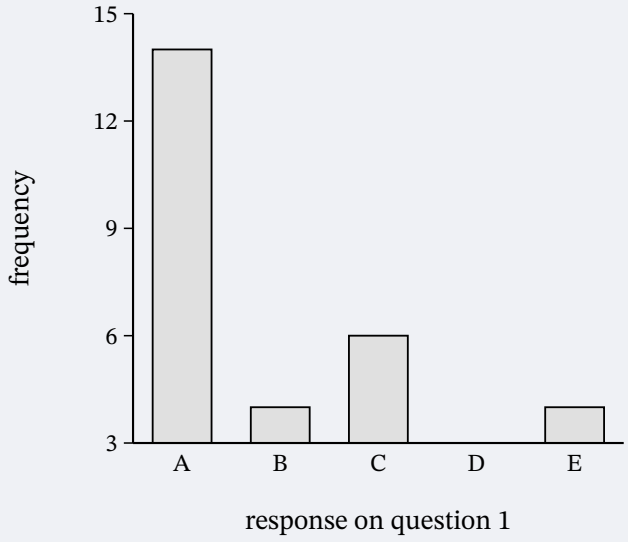
response on question 1

Step 5: Then, draw vertical lines straight down from the edges of your mark to make a rectangle:

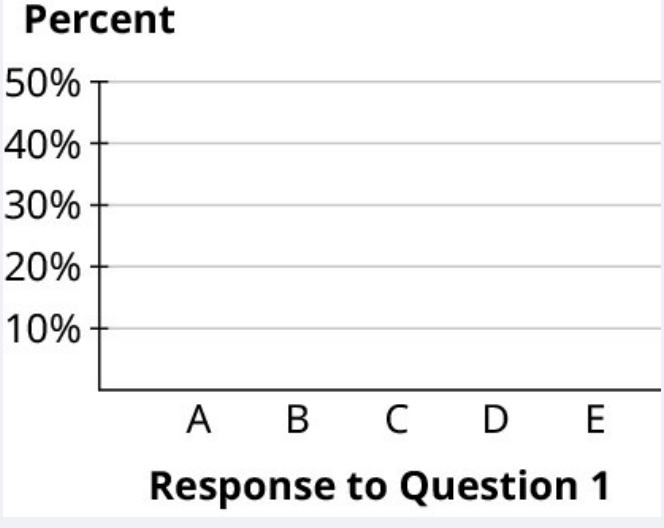


response on question 1

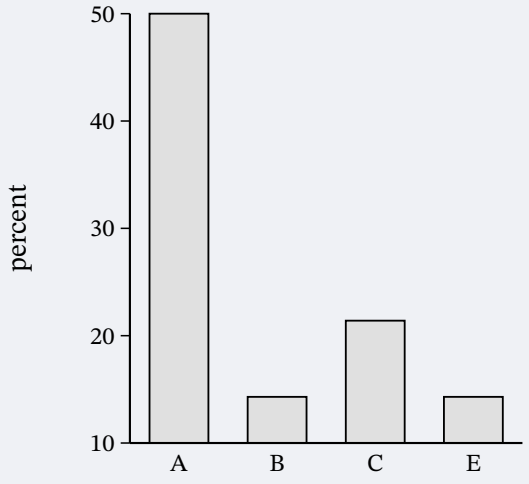
Step 6: Finally, we can build the rest of the rectangles, making sure that the bases all have the same length of the base = width of the rectangle, and the rectangles don't touch. Notice that, since the frequency for "D" is zero, that category has no rectangle (but we'll leave a space there so the reader can see that there is a category with frequency zero). Here's the result:



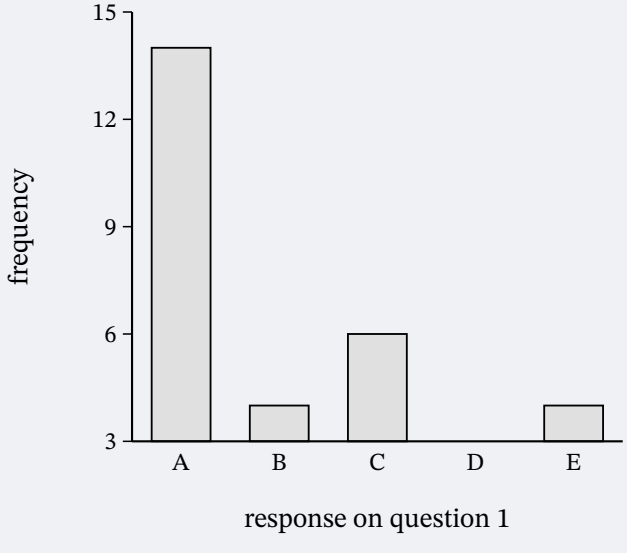
Step 7: That's it! Now, let's use proportions instead of frequencies. We'll label the vertical axis with evenly spaced numbers that run the full range of the percentages in our table: 0% to 50%. We can divide that into five equal parts (so that each has width 10%), and use that to label our vertical axis:



Step 8: Then, we can fill in the rectangles just as we did before. The height of the "A" rectangle is 50%, the "B" rectangle goes up to 14.3%, "C" goes to 21.4%, there is no rectangle for "D" (since its proportion is 0%), and the "E" rectangle also goes up to 14.3%:



Step 9: Notice that the rectangles are basically identical in our two final bar charts. That's no coincidence! Bar charts that use proportions and those that use frequencies will always look identical (which is why it doesn't really matter much which option you choose). Here's why: look at the bars for "B" and "C". The frequencies for these are 4 and 6 respectively. Notice that 6 is 50% bigger than 4 (since $6 = 1.5 \times 4$), which means that the "C" bar will be 50% higher than the "B" bar. Now look at the same bars using proportions: since $21.4\% = 1.5 \times 14.3\%$, the bar for "C" will be 50% higher than the bar for "B." The same relationships hold for the other bars, too.



In practice, most graphs are now made with computers. You can use Google Sheets, which is available for free from any web browser.

VIDEO
Make a Simple Bar Graph in Google Sheets

Now that we've explored how bar graphs are made, let's get some practice reading bar graphs.

Example 3 **Reading Bar Graphs**

The bar graph shown gives data on 2020 model year cars available in the United States. Analyze the graph to answer the following questions.

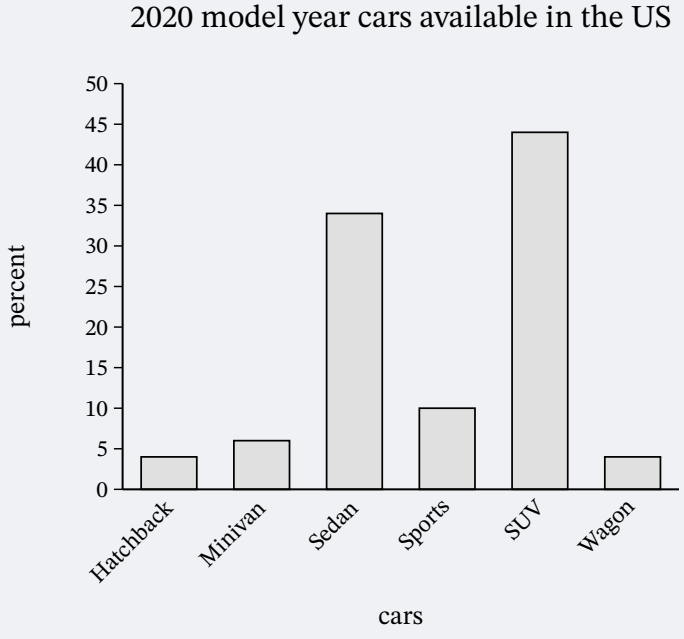


Figure 14: (data source: consumerreports.org/cars)

1. What proportion of available cars were sports cars?
2. What proportion of available cars were sedans?
3. Which categories of cars each made up less than 5% of the models available?

Solution

1. The bar for sports cars goes up to 10%, so the proportion of models that are considered sports cars is 10%.
2. The bar corresponding to sedan goes up past 30% but not quite to 35%. It looks like the proportion we want is between 33% and 34%.
3. We're looking for the bars that don't make it all the way to the 5% line. Those categories are hatchback and wagon.

WORK IT OUT

Candy Color: Frequency and Distribution

M&Ms, Skittles, and Reese's Pieces are all candies that have pieces that are uniformly shaped, but which have different colors. Do the colors in each bag appear with the same frequency? Get a bag of one of these candies and make a bar chart to visualize the color distribution.

Pie Charts

A **pie chart** consists of a circle divided into wedges, with each wedge corresponding to a category. The proportion of the area of the entire circle that each wedge represents corresponds to the proportion of the data in that category. Pie charts are difficult to make without technology because they require careful measurements of angles and precise circles, both of which are tasks better left to computers.

VIDEO

Create Pie Charts Using Google Sheets

Pie charts are sometimes embellished with features like labels in the slices (which might be the categories, the frequencies in each category, or the proportions in each category) or a legend that explains which colors correspond to which categories. When making your own pie chart, you can decide which of those to include. The only rule is that there has to be some way to connect the slices to the categories (either through labels or a legend).

Example 4 **Making Pie Charts**

Use the data that follows to generate a pie chart.

Type	Percent	Type	Percent
SUV	43.6%	Minivan	5.5%
Sedan	33.6%	Hatchback	3.6%
Sports	10.0%	Wagon	3.6%

Solution

First, enter the chart above into a new sheet in Google Sheets. Next, click and drag to select the full table (including the header row). Click on the "Insert" menu, then select "Chart." The result may be a pie chart by default; if it isn't, you can change it to a pie chart using the "Chart type" drop-down menu in the Chart Editor.

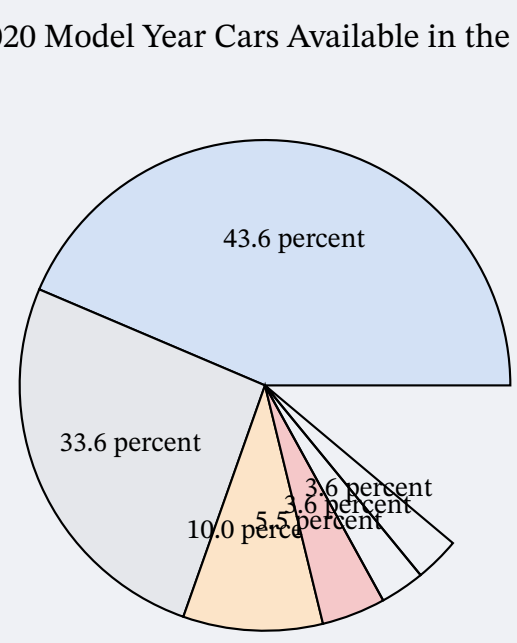


Figure 15: (data source: consumerreports.org/cars)

You can choose to use a legend to identify the categories, as well as label the slices with the relevant percentages.

PEOPLE IN MATHEMATICS

Florence Nightingale

Florence Nightingale (1820–1910) is best remembered today for her contributions in the medical field; after witnessing the horrors of field hospitals that tended to the wounded during the Crimean War, she championed reforms that encouraged sanitary conditions in hospitals. For those efforts, she is today considered the founder of modern nursing.

Figure 16: Florence Nightingale's significant contribution to the field of statistical graphics cannot be understated.

Nightingale is also remembered for her contributions in statistics, especially in the ways we visualize data. She developed a version of the pie chart that is today known as a *polar area diagram*, which she used to visualize the causes of death among the soldiers in the war, highlighting the number of preventable deaths the British Army suffered in that conflict.

In 1859, the Royal Statistical Society honored her for her contributions to the discipline by electing her to join the organization. She was the first woman to be so honored. She was later named an honorary member of the American Statistical Association. Nightingale's status as a revered pioneer in both nursing and statistics is a complex one, because some of her writings and opinions demonstrate a colonialist mindset and disregard for those who lost their lives and lands at the hands of the British. Her core statistical writings indicated that she felt superior to the Indigenous people she was treating. Members of both fields continue to debate her near-iconic role.

Visualizing Quantitative Data

There are several good ways to visualize quantitative data. In this section, we'll talk about two types: stem-and-leaf plots and histograms.

Stem-and-Leaf Plots

Stem-and-leaf plots are visualization tools that fall somewhere between a list of all the raw data and a graph. A stem-and-leaf plot consists of a list of stems on the left and the corresponding leaves on the right, separated by a line. The stems are the numbers that make up the data only up to the next-to-last digit, and the leaves are the final digits. There is one leaf for every data value (which means that leaves may be repeated), and the leaves should be evenly spaced across all stems. These plots are really nothing more than a fancy way of listing out all the raw data; as a result, they shouldn't be used to visualize large datasets.

This concept can be difficult to understand without referencing an example, so let's first look at how to read a stem-and-leaf plot.

Example 5 **Reading a Stem-and-Leaf Plot**

A collector of trading cards records the sale prices (in dollars) of a particular card on an online auction site, and puts the results in a stem-and-leaf plot:

0	5 8 9
1	0 0 0 3 4 4 5 5 5 5 6 9 9
2	0 0 0 0 5 5 9 9
3	0 0 0 5 5
4	0 0 5
5	
6	0

Answer the following questions about the data:

1. How many prices are represented?
2. What prices represent the five most expensive cards? The five least expensive?
3. What is the full set of data?

Solution

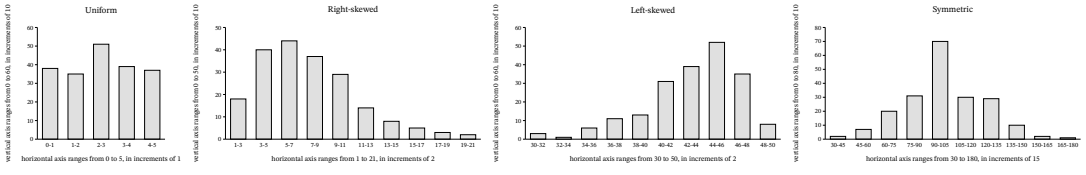
1. Each leaf (the numbers on the right side of the bar) represents one data value. So, on the first row (which looks like 0 | 5 8 9), there are three data values (one for each leaf: 5, 8, and 9). The next row has thirteen leaves, then eight, five, three, zero, and one. Adding those up, we get $3 + 13 + 8 + 5 + 3 + 0 + 1 = 33$ data points or prices.
2. The most expensive card is the last one listed. Its stem is 6 and its leaf is 0, so the price is \$60. There are no leaves associated with the 5 stem, so there were no cards sold for \$50 to \$59. The next most expensive cards are then on the 4 stem: \$45, \$40, and

\$40 (remember, repeated leaves mean repeated values in the dataset). So, we have our four most expensive cards. The fifth would be on the next stem up. The biggest leaf on the 3 stem is a 5, so the fifth-most expensive card sold for \$35.
As for the five least-expensive cards, the smallest stem is 0, with leaves 5, 8, and 9. So, the three least expensive cards sold for \$5, \$8, and \$9 (notice that we don't write down that leading 0 from the stem in the tens place). The next two least-expensive cards will be the two smallest leaves on the next stem: \$10 and \$10.

3. The full list of data is: 5, 8, 9, 10, 10, 10, 13, 14, 14, 15, 15, 15, 15, 16, 19, 19, 20, 20, 20, 24, 25, 25, 29, 29, 30, 30, 30, 35, 35, 40, 40, 45, 60.

Stem-and-leaf plots are useful in that they give us a sense of the shape of the data. Are the data evenly spread out over the stems, or are some stems "heavier" with leaves? Are the heavy stems on the low side, the high side, or somewhere in the middle? These are questions about the **distribution** of the data, or how the data are spread out over the range of possible values.

Some words we use to describe distributions are *uniform* (data are equally distributed across the range), *symmetric* (data are bunched up in the middle, then taper off in the same way above and below the middle), *left-skewed* (data are bunched up at the high end or larger values, and taper off toward the low end or smaller values), and *right-skewed* (data are bunched up at the low end, and taper off toward the high end). See below figures.



Looking back at the stem-and-leaf plot in the previous example, we can see that the data are bunched up at the low end and taper off toward the high end; that set of data is right-skewed. Knowing the distribution of a set of data gives us useful information about the property that the data are measuring.

Now that we have a better idea of how to read a stem-and-leaf plot, we're ready to create our own.

Example 6 **Constructing a Stem-and-Leaf Plot**

An entomologist studying crickets recorded the number of times different crickets (of differing species, genders, etc.) chirped in a one-minute span. The raw data are as follows:

89	97	82	102	84	99	93	103	120	91
115	105	89	109	107	89	104	82	106	92
101	109	116	103	100	91	85	104	104	106

Construct a stem-and-leaf plot to visualize these results.

Solution

Step 1: Before we can create the plot, we need to sort the data in order from smallest to largest:

82	82	84	85	89	89	89	91	91	92
93	97	99	100	101	102	103	103	104	104
104	105	106	106	107	109	109	115	116	120

Step 2: Next, we identify the stems. To do that, we cut off the final digit of each number, which leaves us with stems of 8, 9, 10, 11, and 12. Arrange the stems vertically, and add the bar to separate these from the leaves:

8
9
10
11
12

Step 3: Write down the leaves on the right side of the bar, giving just the final digit (that we cut off to make the stems) of each data value. List these in order, and make sure they line up vertically:

8	2 2 4 5 9 9 9
9	1 1 2 3 7 9
10	0 1 3 3 4 4 4 5 6 6 7 9 9
11	5 6
12	0

As we mentioned above, stem-and-leaf plots aren't always going to be useful. For example, if all the data in your dataset are between 20 and 29, then you'll just have one stem, which isn't terribly useful. (Although there are methods like *stem splitting* for addressing that particular problem, we won't go into those at this time.) On the other end of the spectrum, the data may be so spread out that every stem has only one leaf. (This problem can sometimes be addressed by rounding off the data values to the tens, hundreds, or some other place value, then using that place for the leaves.) Finally, if you have dozens or hundreds (or more) of data values, then a stem-and-leaf plot becomes too unwieldy to be useful. Fortunately, we have other tools we can use.

Histograms

Histograms are visualizations that can be used for any set of quantitative data, no matter how big or spread out. They differ from a categorical bar chart in that the horizontal axis is labeled

with numbers (not ranges of numbers), and the bars are drawn so that they touch each other. The heights of the bars reflect the frequencies in each bin. Unlike with stem-and-leaf plots, we cannot recreate the original dataset from a histogram. However, histograms are easy to make with technology and are great for identifying the distribution of our data. Let's first create one histogram without technology to help us better understand how histograms work.

Example 7 **Constructing a Histogram**

In Example 6, we built a stem-and-leaf plot for the number of chirps made by crickets in one minute. Here are the raw data that we used then:

89	97	82	102	84	99
115	105	89	109	107	89
101	109	116	103	100	91
93	103	120	91	85	104
104	82	106	92	104	106

Construct a histogram to visualize these results.

Solution

Step 1: Add data to bins. Histograms are built on binned frequency distributions, so we'll make that first. Luckily, the stem-and-leaf plot we made earlier can help us do this much more quickly:

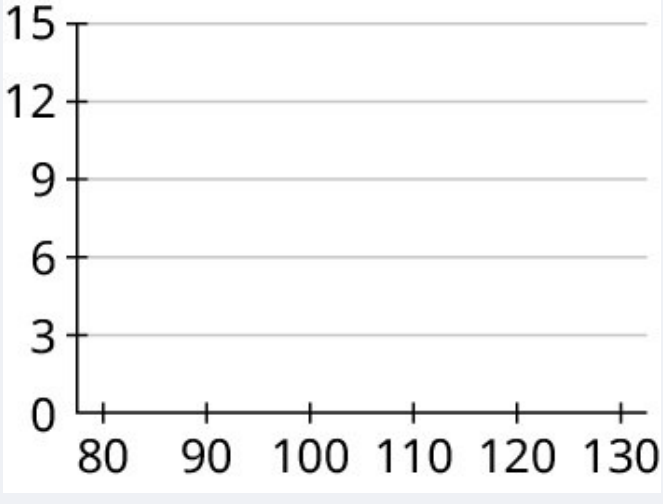
8	2 2 4 5 9 9 9
9	1 1 2 3 7 9
10	0 1 2 3 3 4 4 4 5 6 6 7 9 9
11	5 6
12	0

If we're using bins of width 10, we can compute the frequencies by counting the numbers of leaves associated with the corresponding stem:

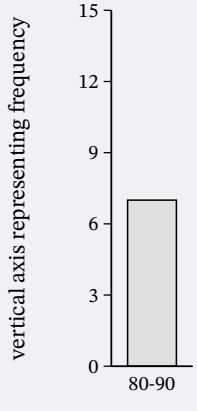
Bin	Frequency
80-89	7
90-99	6
100-109	14
110-119	2
120-129	1

(Note that, when we made binned frequency diagrams in the last module, we noted that if the biggest data value was right on the border between two bins, it was OK to lump it in with the lower bin. That's not recommended when building histograms, so the data value 120 is all alone in the 120-129 bin.)

Step 2: Create the axes. On the horizontal axis, start labeling with the lower end of the first bin (in this case, 80), and go up to the higher end of the last bin (120). Mark off the other bin boundaries, making sure they're all evenly spaced. On the vertical axis, start with zero and go up at least to the greatest frequency you see in your bins (14 in this example), making sure that the labels you make are evenly spaced and that the difference between those numbers is the same. Let's count off our vertical axis by threes:

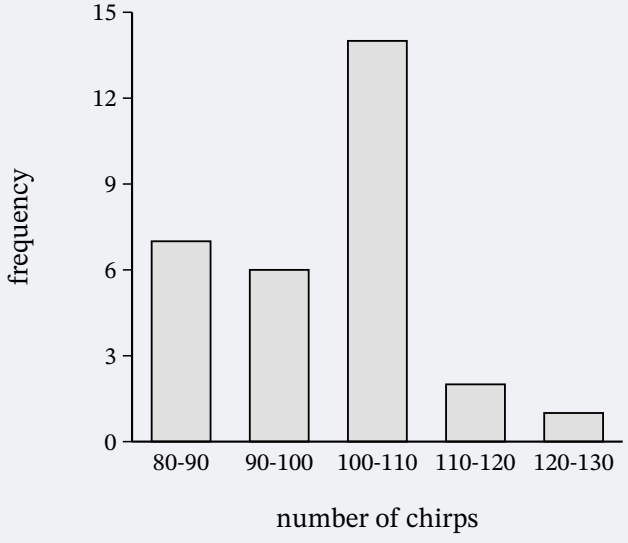


Step 3: Draw in the bars. Remember that the bars of a histogram touch, and that the heights are determined by the frequency. So, the first bar will cover 80 to 90 on the horizontal axis, and have a height of 7:

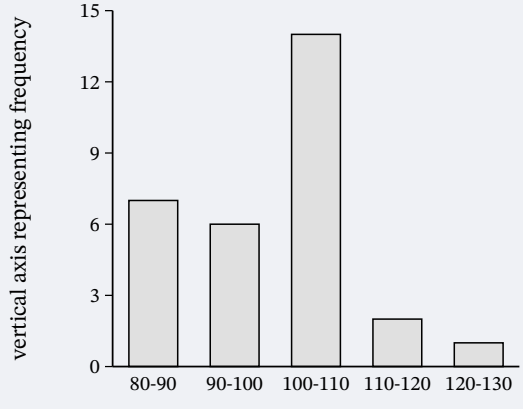


horizontal axis representing values from 80 to 130 in increments of 10

Now, we can fill in the others:



Step 4: Let's compare the histogram we just created to the stem-and-leaf plot we made earlier:



horizontal axis representing values from 80 to 130 in increments of 10

Notice that the leaves on the rotated stem-and-leaf plot match the bars on our histogram! We can view stem-and-leaf plots as sideways histograms. But, as we'll see soon, we can do much more with histograms.

Now that we've seen the connection between stem-and-leaf plots and histograms, we are ready to look at how we can use Google Sheets to build histograms.

VIDEO
Make a Histogram Using Google Sheets

Let's use Google Sheets to create a histogram for a large dataset.

Example 8 **Creating a Histogram in Google Sheets**

The data in "AvgSAT" contains the average SAT score for students attending every institution of higher learning in the US for which data is available. Create a histogram in Google Sheets of the average SAT scores. Use bins of width 50. Are the data uniformly distributed, symmetric, left-skewed, or right-skewed?

Solution

Using the procedure described in the video above, we get this:

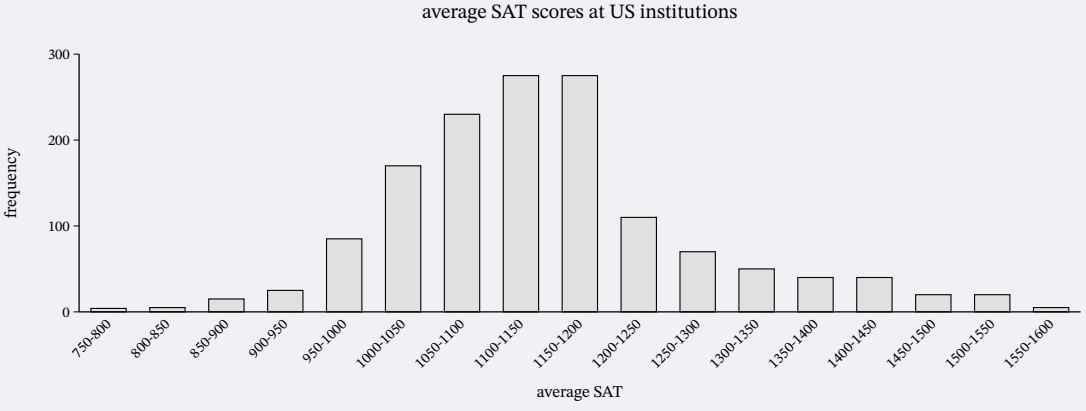


Figure 22: (data source: https://data.ed.gov)

The data are fairly symmetric, but slightly right-skewed.

Bar Charts for Labeled Data

Sometimes we have quantitative data where each value is labeled according to the source of the data. For example, in the Your Turn above, you looked at in-state tuition data. Every value you used to create that histogram was associated with a school; the schools are the labels. In YOUR TURN 8.11, you found a histogram of the wins of every Major League Baseball team in 2019. Each of those win totals had a label: the team. If we're interested in visualizing differences among the different teams, or schools, or whatever the labels are, we create a different version of the bar graph known as a **bar chart for labeled data**.

These graphs are made in Google Sheets in exactly the same way as regular bar graphs. The only change is that the vertical axis will be labeled with the units for your quantitative data instead of just "Frequency."

Example 9 **Building a Bar Chart for Labeled Data**

The following table shows the gross domestic product (GDP) for the United States for the years 2010 to 2019:

Year	GDP (in $ trillions)	Year	GDP (in $ trillions)
2010	14.992	2015	18.225
2011	15.543	2016	18.715
2012	16.197	2017	19.519
2013	16.785	2018	20.580
2014	17.527	2019	21.433

Construct a histogram that represents these data.

Solution

In this case, the years are the labels, and the data we are interested in are the GDP numbers. Once you have the table above (including the labels) entered into a spreadsheet, click and drag to select the full table. Then, in the "Insert" menu, click "Chart." The result may not be a bar chart; if it's not, select "Column chart" in the drop-down menu "Chart type" in the Chart Editor. If you want, you can edit things like the chart title in the "Customize" tab in the Chart Editor.

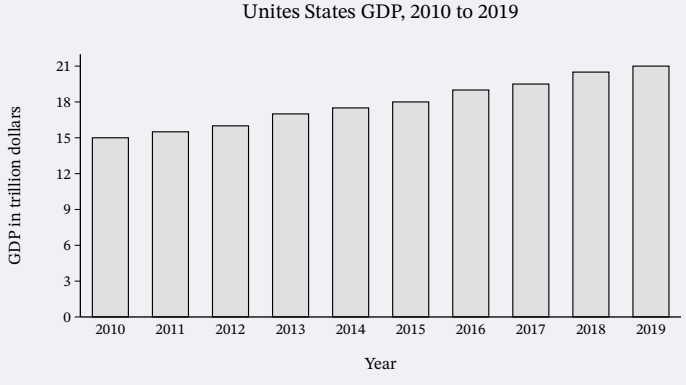


Figure 23: (data source: https://data.worldbank.org)

Misleading Graphs

Graphical representations of data can be manipulated in ways that intentionally mislead the reader. There are two primary ways this can be done: by manipulating the scales on the axes and by manipulating or misrepresenting areas of bars. Let's look at some examples of these.

Example 10 **Misleading Graphs**

The table below shows the teams, and their payrolls, in the English Premier League, the top soccer organization in the United Kingdom.

Team	Salary (£1,000,000s)	Team	Salary (£1,000,000s)
Manchester United F.C.	175.7	Newcastle United F.C.	56.9
Manchester City F.C.	136.5	Aston Villa F.C.	52.3
Chelsea F.C.	132.8	Fulham F.C.	52.1
Arsenal F.C.	130.7	Southampton F.C.	49.6
Tottenham Hotspur F.C.	129.2	Wolverhampton Wanderers F.C.	49.5
Liverpool F.C.	118.6	Brighton & Hove Albion	43.7
Crystal Palace	85.0	Burnley F.C.	35.5
Everton F.C.	82.5	West Bromwich Albion F.C.	23.8
Leicester City	73.7	Leeds United F.C.	22.5
West Ham United F.C.	69.2	Sheffield United F.C.	19.7

How might someone present this data in a misleading way?

Solution

Step 1: Let's focus on the top five teams. Here's a bar chart of their payrolls:

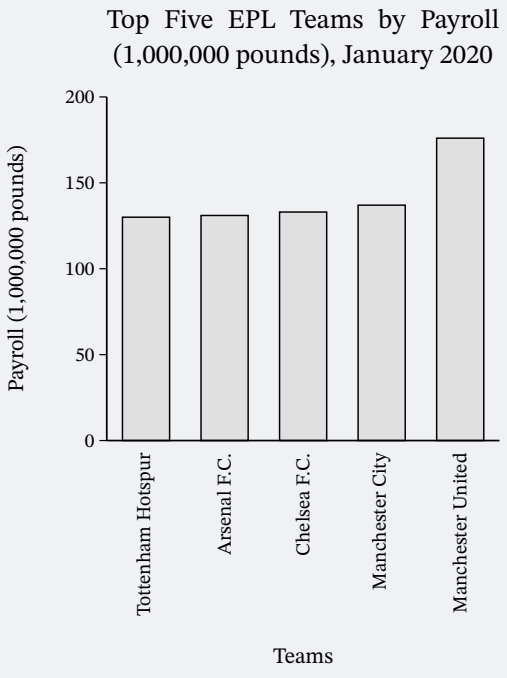


Figure 24: (data source: www.spotrac.com)

Step 2: Now, here's another bar chart visualizing exactly the same data:

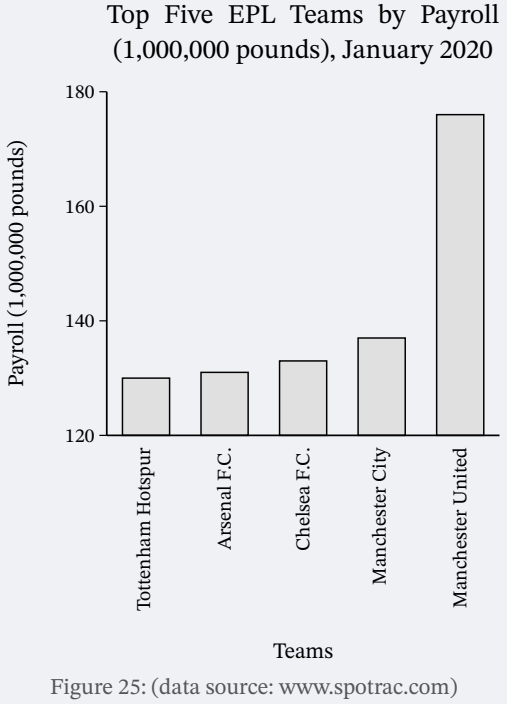


Figure 25: (data source: www.spotrac.com)

Step 3: You should notice that despite using the same data, these two graphs look strikingly different. In the second graph, the gap between Manchester United and the other four teams looks significantly larger than in the first graph. The scale on the vertical axis has been manipulated here. The first graph's axis starts at zero, while the lowest value on the second graph's axis is 120. This trick has a strong impact on the viewer's perception of the data.

CHECKPOINT

Beware of vertical axes that don't start at zero! They overemphasize differences in heights.

Step 4: To further emphasize the difference this creates in our perception, let's look at that data again, but this time using graphics instead of colored areas on our bar graph.

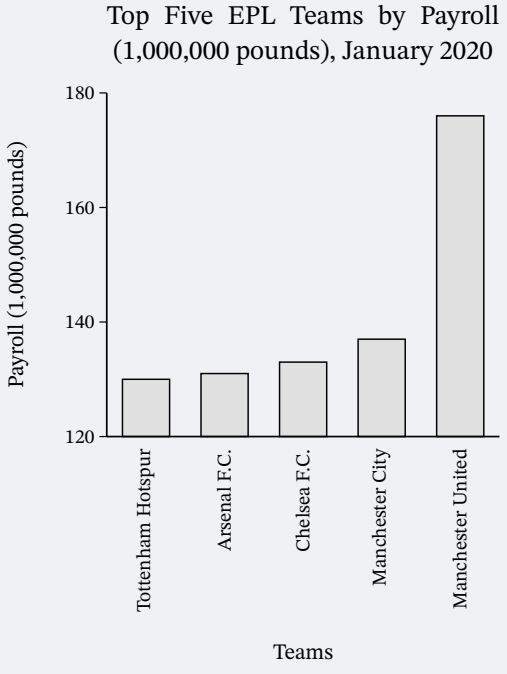


Figure 26: (data source: www.spotrac.com)

This graph uses an image of a £10 banknote in place of the bars. Using an image that evokes the context of the data in place of a standard, "boring" bar is a common tool that people use when creating *infographics*. However, this is generally not a good practice because it distorts the data. Notice that our "bars" (the banknotes) are just as tall here as they were in the previous figure. But, to maintain the right proportions, the *widths* had to be adjusted as well, which changes the area (height × width) of each bar. A key point is that when looking at rectangles, the human eye tends to process areas more easily than heights.

CHECKPOINT

Beware of infographics! Areas overemphasize a difference that should be measured with a height!

Step 5: Now, let's look at all 20 teams. This histogram indicates that the data are right-skewed, with the highest number of teams having a payroll between £40 million and £80 million:

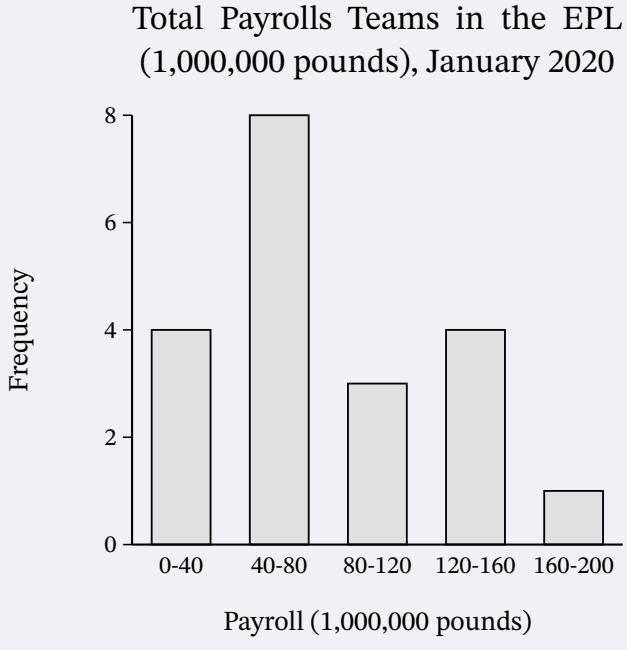


Figure 27: (data source: www.spotrac.com)

Step 6: Now let's view this same data in another chart:

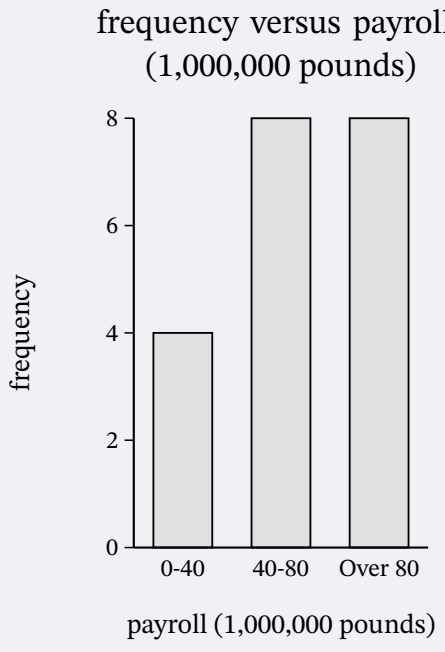


Figure 28: (data source: www.spotrac.com)

Step 7: Even though this chart uses the same data, the skew seems to be reversed. Why? Well, even though this graph *looks* like a histogram, it isn't. Look closely at the labels on the horizontal axis; they don't correspond to spots on the axis, but instead provide a range, meaning this is a bar graph based on a binned frequency distribution.
When we review these ranges, we can see that the last range is misleading as it consists of all data "over 80." If the bins all had the same width, that last bin would run from 80 to 120. However, we can see from the histogram that the maximum value for this data is between 160 and 200. If the last bin in this bar graph were labeled honestly, it would read "80–200," which would drive home the fact that the width of that bar is misleading.

CHECKPOINT
Always check the horizontal axis on histograms! The widths of all the bars should be equal.

VIDEO
How to Spot a Misleading Graph

WHO KNEW?
Napoleon's Failed Invasion

One of the most famous data visualizations ever created is the cartographic depiction by Charles Joseph Minard of Napoleon's disastrous attempted invasion of Russia.

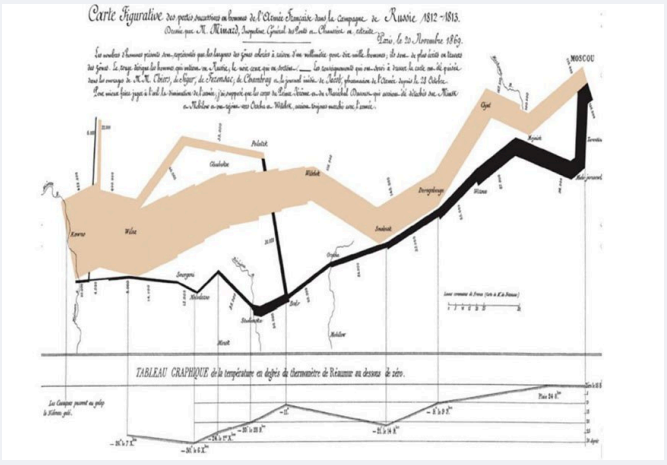


Figure 29: Minard's Napoleon Map

Minard's chart is remarkable in that it shows not just how the size of Napoleon's army shrank drastically over time, but also the location on the map, the direction the army was traveling at the time, and the temperature during the retreat.

Key Terms

- proportion
- bar chart
- pie chart
- stem-and-leaf plot
- distribution (of quantitative data)
- histogram
- bar chart for labeled data

Key Concepts

- Categorical data can be visualized using pie charts or bar charts; quantitative data can be visualized using stem-and-leaf plots or histograms.
- Areas in pie charts and bar charts represent proportions of the data falling into a particular category, while areas in histograms represent proportions of the data that fall into a given range of data values (or "bins"). Stem-and-leaf plots are visual representations of entire datasets.
- By manipulating the axes, changing widths of bars, or making bad choices for bins, we can create data visualizations that misrepresent the distribution of data.

Videos

- Make a Simple Bar Graph in Google Sheets

- Create Pie Charts Using Google Sheets
- Make a Histogram Using Google Sheets
- How to Spot a Misleading Graph

8.3 Mean, Median and Mode

Figure 30: What does it mean to say someone has average height?

Learning Objectives

After completing this section, you should be able to:

1. Calculate the mode of a dataset.
2. Calculate the median of a dataset.
3. Calculate the mean of a dataset.
4. Contrast measures of central tendency to identify the most representative average.
5. Solve application problems involving mean, median, and mode.

   What exactly do we mean when we describe something as "average"? Is the height of an average person the height that more people share than any other? What if we line up every person in the world, in order from shortest to tallest, and find the person right in the middle: Is that person's height the average? Or maybe it's something more complicated.

   Imagine a game where you and a friend are trying to guess the typical person's height. Once the guesses are made, you bring in every person and measure their height. You and your friend figure out how far off each of your guesses were from the actual value, then square that number. The result is the number of points you earn for that person. After we check every height and award points accordingly, the person with the lower score wins (because a lower score means that person's guess was, overall, closer to the actual values). Could we define the average height to be the number that you should guess to give you the smallest possible score?

Each of these three methods of determining the "average" is commonly used. They are all methods of measuring *centrality* (or *central tendency*). Centrality is just a word that describes the middle of a set of data. All give potentially different results, and all are useful for different reasons. In this section, we'll explore each of these methods of finding the "average."

The Mode

In our discussion of average heights, the first possible definition we offered was the height that more people share than any other. This is the **mode**, or the value that appears most often. If there are two modes, the data are **bimodal**.

Let's look at some examples.

Example 1 **Finding the Mode Using a Stem-and-Leaf Plot**

In Example 5 in Section 8.2, we looked at a stem-and-leaf plot of the sale prices (in dollars) of a particular collectible trading card:

0	5 8 9
1	0 0 0 3 4 4 5 5 5 5 6 9 9
2	0 0 0 0 5 5 9 9
3	0 0 0 5 5
4	0 0 5
5	
6	0

What is the mode price?

Solution

The mode is the price that appears most often. Both 15 and 20 appear 4 times, more than any other values. So, they are the modes (and we can conclude that this set of data is bimodal).

When we have a complete list of the data or a stem-and-leaf plot, it's pretty straightforward to find the mode; we just need to find the number that appears most often. If we're given a frequency distribution instead, the technique is different (but just as straightforward): we're looking for the number with the highest frequency.

Example 2 **Finding the Mode Using a Frequency Distribution**

In Example 3 in Section 8.1, we created a frequency distribution of the number of siblings of conflict resolution class attendees.

Number of Siblings	Frequency
0	5
1	13
2	6
3	3
4	2
5	1

What is the mode of the number of siblings?

Solution

The mode is the value that appears the most often, which means it has the greatest frequency. Thirteen of the respondents have one sibling, more than any other number. So, the mode is 1.

What happens if there is no number in the data that appears more than once? In that case, by our definition, every data value is a mode. But according to some other definitions, the data would have no mode. In practice, though, it doesn't really matter; if no data value appears more than once, then the mode is not helpful at all as a measure of centrality.

The Median

Let's revisit our example of trying to identify the height of the "average" person. If we lined everyone up in order by height and found the person right in the middle, that person's height is called the **median**, or the value that is greater than no more than half and less than no more than half of the values.

Let's look at a really simple example. Consider the following list of numbers: 11, 12, 13, 13, 14. Is the first number on the list, 11, the median? There are no values less than 11 (that's 0%), and there are four values greater than 11 (that's 80%). Since more than 50% of the data are greater than 11, the definition is violated; it's not the median. Here's a chart with the rest of the data, with red shading to show where the definition is violated:

Data Value	Number of Values Below	Percentage of Values Below	Number of Values Above	Percentage of Values Above
11	0	0%	4	**80%**
12	1	20%	3	**60%**
13	2	40%	1	20%
14	4	**80%**	0	0%

Only 13 has no violations, so it's the median according to the definition. In practice, we find the median just like we described in the average height example: by lining up all the data values in

order from smallest to largest and picking the value in the middle. For our easy example (with data values 11, 12, 13, 13, 14), that first 13 is right in the middle; there are two values to the left and two values to the right. If there's not one value right in the middle, we pick the two closest, then choose the number exactly between them. For example, let's say we have the data 41, 44, 46, 53. Since there are an even number of data values in our list, we can't pick the one right in the middle. The two closest to the middle are 44 and 46, so we'll choose the number halfway between those to be the median: 45. As this example shows, the median (unlike the mode) doesn't have to be a number in our original set of data.

In the examples we've looked at so far, it's been pretty easy to identify which number is right in the middle. If we had a very large dataset, though, it might be harder. Fortunately, we have some formulas to help us with that.

FORMULA

Suppose we have a set of data with n values, ordered from smallest to largest. If n is odd, then the median is the data value at position $\frac{n+1}{2}$. If n is even, then we find the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$. If those values are named a and b, then the median is defined to be $\frac{a+b}{2}$.

Let's put those formulas to work in an example.

Example 3 **Finding the Median Using a Stem-and-Leaf Plot**

In Example 5 in Section 8.2, we looked at a stem-and-leaf plot that contained 33 sale prices (in dollars) of a particular collectible trading card:

0	5 8 9
1	0 0 0 3 4 4 5 5 5 5 6 9 9
2	0 0 0 0 5 5 9 9
3	0 0 0 5 5
4	0 0 5
5	
6	0

What is the median price?

Solution

Step 1: Since 33 is odd, the median is the data value at position $\frac{n+1}{2}$, where n is the number of values in the dataset. There are 33 total values, so our formula becomes $\frac{33+1}{2} = 17$. That means we want to look for the 17th number in the dataset.

Step 2: We'll want to count from the lowest value to the 17th number. We can use our stem and leaf plot to do this.

	1 2 3
0	5 8 9
	4 5 6 7 8 9 10 11 12 13 14 15 16
1	0 0 0 3 4 4 5 5 5 5 6 9 9
	17
2	0 0 0 0 5 5 9 9
3	0 0 0 5 5
4	0 0 5
5	
6	0

The seventeenth number is 20, so the median is 20.

Now, let's tackle an example with an even number of values.

Example 4 **Finding the Median**

In Example 6 in Section 8.2, we looked at the number of times different crickets (of differing species, genders, etc.) chirped in a one-minute span. That data is again provided below:

89	97	82	102	84	99
115	105	89	109	107	89
101	109	116	103	100	91
93	103	120	91	85	104
104	82	106	92	104	106

Find the median.

Solution

Step 1: In order to find the median, we first need to sort the data so that they're in order, smallest to largest:

82	82	84	85	89	89
89	91	91	92	93	97
99	100	101	102	103	103
104	104	104	105	106	106
107	109	109	115	116	120

Step 2: Next, we figure out how many data values we have. Counting them up, we see there are 30, which is even.

Step 3: Since we have an even number of data values, we need to find the values in positions $\frac{30}{2} = 15$ and $\frac{30}{2} + 1 = 16$. These are 101 and 102.

Step 4: We use the formula to compute the median: $\frac{101+102}{2} = 101.5$.

Example 5 **Finding the Median Using a Frequency Distribution**

In Example 3 in Section 8.1, we created a frequency distribution of the number of siblings of the people who attended a conflict resolution class. Let's review that data again:

Number of Siblings	Frequency
0	5
1	13
2	6
3	3
4	2
5	1

What is the median of the number of siblings?

Solution

There are 30 data values total, so the median is between the 15th and 16th values in the ordered list. There are five 0s and thirteen 1s according to the frequency distribution, so items one through five are all 0s and items six through eighteen are all 1s. Since both items fifteen and sixteen are 1s, the median is 1.

The Mean

Recall our example of ways we could identify the "average" height of an individual. The last method we discussed was also the most complicated. It involved a game where the player guesses a height, then figures out how far off that guess is from every single person's height. Those differences get squared and added together to get a score. Our next measure of centrality gives the lowest possible score: No other guess would beat it in the game. Given a dataset containing n total values, the **mean** of the dataset is the sum of all the data values, divided by n.

This is a computation you have likely done before. In many places, including spreadsheet programs like Microsoft Excel and Google Sheets, this number is called the *average*. For statisticians, though, the word *average* has too many possible meanings, so they prefer the one we'll use: mean.

Example 6 **Finding the Mean**

Compute the mean of the numbers 12, 15, 17, 18, 18, and 19.

Solution

The mean is the sum of the values, divided by the number of values on the list. So, we get:

$$\frac{12+15+17+18+18+19}{6} = \frac{99}{6} = 16.5.$$

Example 7 **Finding the Mean Using a Frequency Distribution**

Refer again to the frequency distribution of the number of siblings people who attended a conflict resolution class reported:

Number of Siblings	Frequency
0	5
1	13
2	6
3	3
4	2
5	1

What is the mean of the number of siblings?

Solution

Step 1: We compute the mean by adding up all the data values and then dividing by the number of data values on the list.

Step 2: Adding up the frequencies, we get $5 + 13 + 6 + 3 + 2 + 1 = 30$ data values in our list.

Step 3: Now, to find the sum of all the data values, we could simply reconstruct the raw data and add up all the numbers there. But, there's an easier way: Remember that repeatedly adding a number to itself is the definition of multiplication. So, for example, since there are six 2s in our data, the sum of all those 2s must be $6 \times 2 = 12$.

Step 4: Let's add a column to our distribution for these products:

Number of Siblings	Frequency	(Number of Siblings) × (Frequency)
0	5	0
1	13	13
2	6	12
3	3	9
4	2	8
5	1	5

Step 5: So, the sum of all our data values is $0 + 13 + 12 + 9 + 8 + 5 = 47$. The mean is $\frac{47}{30} \approx 1.567$.

As the number of data values we are considering grows, the computation for the mean gets more and more complicated. That's why people generally trust technology to perform that computation.

VIDEO
Compute Measures of Centrality Using Google Sheets

Note that a recent update to Google Sheets introduced a new function called "MODE.MULT," which will find every mode (not just the first one on the list).

Example 8 **Using Google Sheets to Compute Measures of Centrality**

The dataset "InState" contains the in-state tuitions of every college and university in the country that reported that data to the Department of Education. Find the mode, median, and mean of those in-state tuition values.

Solution

Step 1: To find the mode, we select an empty cell type "=MODE(", click on the header of the column to insert a reference to the column into our formula, and then close the parentheses. When we hit the enter key, our formula is replaced with the mode: $13,380.

Step 2: We can find the median and mean using the same process, except using the functions "MEDIAN" and "AVERAGE" in place of "MODE". We find that the median is $11,207 and the mean is $15,476.79.

Which Is Better: Mode, Median, or Mean?

If the mode, median, and mean all purport to measure the same thing (centrality), why do we need all three? The answer is complicated, as each measure has its own strengths and weaknesses. The mode is simple to compute, but there may be more than one. Further, if no data value appears more than once, the mode is entirely unhelpful. As for the mean and median, the main difference between these two measures is how each is affected by extreme values.

Consider this example: the mean and median of 1, 2, 3, 4, 5 are both 3. But what if the dataset is instead 1, 2, 3, 4, 10? The median is still 3, but the mean is now 4. What this example shows is that the mean is sensitive to extreme values, while the median isn't. This knowledge can help us decide which of the two is more relevant for a given dataset. If it is important that the really high or really low values are reflected in the measure of centrality, then the mean is the better option. If very high or low values are not important, however, then we should stick with the median.

The decision between mean and median only really matters if the data are skewed. If the data are symmetric, then the mean and median are going to be approximately equal, and the distinction between them is irrelevant. If the data are skewed, the mean gets pulled in the direction of the skew (i.e., if the data are right-skewed, then the mean will be bigger than the median; if the data are left-skewed, the opposite relation is true).

Example 9 **Choosing Which Measure of Centrality to Use**

For the following situations, decide which measure(s) of centrality would be best:

1. You found a used car that you like, and you want to know if the price is too high or too low. You find a list of sale prices for that make and model, and you see that the distribution of those prices is skewed to the right. Some of the prices at the high end are close to the original sale price of the car, so you guess that those cars might have really low mileage, or have other enhancements added on that increased the value (but which don't apply to the car you found).
2. You are asked to analyze the responses to a survey. One of the questions asked, "How strongly do you agree with the statement, 'I believe my elected representatives have my best interests in mind when they vote'?" Responses are a number between 1 and 5, with 1 representing "strongly disagree" and 5 representing "strongly agree."
3. You are asked to find the "average" household income for a zip code. Those values are skewed right.
4. Thinking back to the situation at the beginning of this section: you want to find "average" height. The data you've collected seem to be distributed symmetrically.

Solution

1. In this situation, the high values are not comparable to the value of the car you found and we don't want them to affect the results. Also, we're unlikely to find many repeated values, so the mode is probably not useful. Median is best.
2. Here, we want to know what a typical result is. The mean doesn't really make sense; it involves adding the numbers together, so it would treat two "strongly disagree" and two "strongly agree" responses (those add to $1 + 1 + 5 + 5 = 12$) as exactly the same as four "neutral" responses ($3 + 3 + 3 + 3 = 12$). But those are really different situations; the first shows a strong polarization in the responses, while the second represents strong indifference. The mode is probably the best choice (because the data are actually categorical), but the median would be good too.

3. The mode isn't going to be useful in this situation because it's unlikely you will find many households that have exactly the same income. The mean and median will be different because of the skew, so the choice comes down to the extreme values. Remember that the data are skewed right, so high values are prevalent. Because these high values are important for our analysis, we want them to be reflected in the results. Thus, the mean is best. That being said, the median is also useful; it allows us to say something like "50% of the households surveyed make more than" the median.
4. Because we aren't likely to find many people with exactly the same height, the mode won't be useful. Since the data are symmetric, the mean and the median will be about the same. So, it doesn't really matter which of those two we choose.

Key Terms

- mode
- bimodal
- median
- mean

Key Concepts

- The mode of a dataset is the value that appears the most frequently. The median is a value that is greater than or equal to no more than 50% of the data and less than or equal to no more than 50% of the data. The mean is the sum of all the data values, divided by the number of units in the dataset.
- The median of a dataset is not affected by outliers, but the mean will be biased toward outliers. This distinction might affect which measure of centrality is used to summarize a dataset.

Formulas

Suppose we have a set of data with n values, ordered from smallest to largest. If n is odd, then the median is the data value at position $\frac{n+1}{2}$. If n is even, then we find the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$. If those values are named a and b, then the median is defined to be $\frac{a+b}{2}$.

Videos

- Compute Measures of Centrality Using Google Sheets

8.4 Range and Standard Deviation

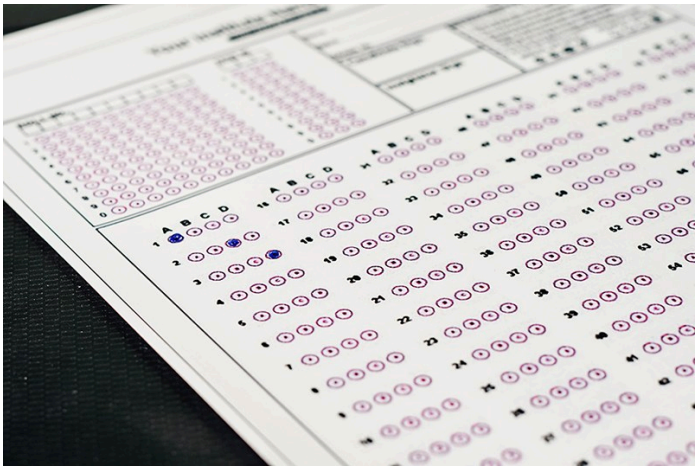


Figure 31: Measures of spread help us get a better understanding of test scores.

Learning Objectives

After completing this section, you should be able to:

1. Calculate the range of a dataset
2. Calculate the standard deviation of a dataset

Measures of centrality like the mean can give us only part of the picture that a dataset paints. For example, let's say you've just gotten the results of a standardized test back, and your score was 138. The mean score on the test is 120. So, your score is above average! But how good is it *really*? If all the scores were between 100 and 140, then you know your score must be among the best. But if the scores ranged from 0 to 200, then maybe 140 is good, but not great (though still above average). Knowing information about how the data are spread out can help us put a particular data value in better context. In this section, we'll look at two numbers that help us describe the spread in the data: the range and the standard deviation. These numbers are called measures of dispersion.

The Range

Our first measure of dispersion is the **range**, or the difference between the maximum and minimum values in the set. It's the measure we used in the standardized test example above.

Let's look at a couple of examples.

Example 1 **Finding the Range**

You survey some of your friends to find out how many hours they work each week. Their responses are: 5, 20, 8, 10, 35, 12. What is the range?

Solution

The maximum value in the set is 35 and the minimum is 5, so the range is $35 - 5 = 30$.

For large datasets, finding the maximum and minimum values can be daunting. There are two ways to do it in a spreadsheet. First, you can ask the spreadsheet program to sort the data from smallest to largest, then find the first and last numbers on the sorted list. The second method uses built-in functions to find the minimum and maximum.

VIDEO
Find the Minimum and Maximum Using Google Sheets

In either method, once you've found the maximum and minimum, all you have to do is subtract to find the range.

Example 2 **Finding the Range with Google Sheets**

The data in "AvgSAT" contains the average SAT score for students attending every institution of higher learning in the US for which data is available. What is the range of these average SAT scores?

Solution

Step 1: To find the maximum, click on an empty cell in the spreadsheet, type "=MAX(", and then click on the letter that marks the top of the column containing the AvgSAT data. That inserts a reference to the column into our function. Then we close the parentheses and hit the enter key. The formula is replaced with the maximum value in our data: 1566.

Step 2: Using the same process (but with "MIN" instead of "MAX"), we find the minimum value is 785.

Step 3: So, the range is $1566 - 785 = 781$.

The range is very easy to compute, but it depends only on two of the data values in the entire set. If there happens to be just one unusually high or low data value, then the range might give a distorted measure of dispersion. Our next measure takes every single data value into account, making it more reliable.

The Standard Deviation

The standard deviation is a measure of dispersion that can be interpreted as approximately the average distance of every data value from the mean. (This distance from the mean is the "deviation" in "standard deviation.")

FORMULA

The standard deviation is computed as follows:

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$$

Here, x represents each data value, $\overline{x}$ is the mean of the data values, n is the number of data values, and the capital sigma (Σ) indicates that we take a sum.

To compute the standard deviation using the formula, we follow the steps below:

1. Compute the mean of all the data values.
2. Subtract the mean from each data value.
3. Square those differences.
4. Add up the results in step 3.
5. Divide the result in step 4 by $n - 1$
6. Take the square root of the result in step 5.

Let's see that process in action.

Example 3 **Computing the Standard Deviation**

You surveyed some of your friends to find out how many hours they work each week. Their responses were: 5, 20, 8, 10, 35, 12. What is the standard deviation?

Solution

Let's follow the six steps mentioned previously to compute the standard deviation.

Step 1: Find the mean: $\overline{x} = \frac{5+20+8+10+35+12}{6} = 15$.

Step 2: Subtract the mean from each data value. To help keep track, let's do this in a table. In the first row, we'll list each of our data values (and we'll label the row x); in the second, we'll subtract $\overline{x} = 15$ from each data value.

x	5	20	8	10	35	12
$x - \overline{x}$	-10	5	-7	-5	20	-3

Step 3: Square the differences. Let's add a row to our table for those values:

x	5	20	8	10	35	12
$x - \overline{x}$	-10	5	-7	-5	20	-3
$(x - \overline{x})^2$	100	25	49	25	400	9

Step 4: Add up those squares: $100 + 25 + 49 + 25 + 400 + 9 = 608$.

Step 5: Divide the sum by $n - 1$. Since we have 6 data values, that gives us $\frac{608}{6-1} = 121.6$.

Step 6: Take the square root of the result: $\sqrt{121.6} \approx 11.027$.

Thus, the standard deviation is $s \approx 11.027$.

The computation for the standard deviation is complicated, even for just a small dataset. We'd never want to compute it without technology for a large dataset! Luckily, technology makes this calculation easy.

VIDEO

Find the Standard Deviation Using Google Sheets

Example 4 **Finding the Standard Deviation with Google Sheets**

The data in "AvgSAT" contains the average SAT score for students attending every institution of higher learning in the US for which data is available. What is the standard deviation of these average SAT scores?

Solution

To find the standard deviation, we click in an empty cell in our spreadsheet and then type "=STDEV(". Next, click on the letter at the top of the column containing our data; this will put a reference to that column into our formula. Then close the parentheses with and hit the enter key. The formula is replaced with the result: 125.517.

Key Terms

- range
- standard deviation

Key Concepts

- The range of a dataset is the difference between its largest and smallest values. The standard deviation is approximately the mean difference (in absolute value) that individual units fall from the mean of the dataset.

Formulas

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$$

Here, s is the standard deviation, x represents each data value, $\overline{x}$ is the mean of the data values, n is the number of data values, and the capital sigma (Σ) indicates that we take a sum.

Videos

- Find the Minimum and Maximum Using Google Sheets
- Find the Standard Deviation Using Google Sheets

8.5 Percentiles

Figure 32: Two students graduating with the same class rank could be in different percentiles depending on the school population.

Learning Objectives

After completing this section, you should be able to:

1. Compute percentiles.
2. Solve application problems involving percentiles.

A college admissions officer is comparing two students. The first, Anna, finished 12th in her class of 235 people. The second, Brian, finished 10th in his class of 170 people. Which of these outcomes is better? Certainly 10 is less than 12, which favors Brian, but Anna's class was much bigger. In fact, Anna beat out 223 of her classmates, which is $\frac{223}{235} \approx 95\%$ of her classmates, while Brian bested 160 out of 170 people, or 94%. Comparing the proportions of the data values that are below a given number can help us evaluate differences between individuals in separate populations. These proportions are called **percentiles**. If $p\%$ of the values in a dataset are less than a number n, then we say that n is at the pth percentile.

Finding Percentiles

There are some other terms that are related to "percentile" with meanings you may infer from their roots. Remember that the word *percent* means "per hundred." This reflects that percentiles divide our data into 100 pieces. The word **quartile** has a root that means "four." So, if a data value is at the first quantile of a dataset, that means that if you break the data into four parts (because of the *quart-*), this data value comes after the first of those four parts. In other words, it's greater than 25% of the data, placing it at the 25th percentile. **Quintile** has a root meaning "five," so a data value at the third quintile is greater than three-fifths of the data in the set. That

would put it at the 60th percentile. The general term for these is **quantiles** (the root *quant–* means "number").

In Mean, Median, and Mode, we defined the *median* as a number that is greater than no more than half of the data in a dataset and is less than no more than half of the data in the dataset. With our new term, we can more easily define it: The median is the value at the 50th percentile (or second quartile).

Let's look at some examples.

Example 1 **Finding Percentiles**

Consider the dataset 5, 8, 12, 1, 2, 16, 2, 15, 20, 22.

1. At what percentile is the value 5?
2. What value is at the 60th percentile?

Solution

Before we can answer these two questions, we must put the data in increasing order: 1, 2, 2, 5, 8, 12, 15, 16, 20, 22.

1. There are three values (1, 2, and 2) in the set that are less than 5, and there are ten values in the set. Thus, 5 is at the $\frac{3}{10} \times 100 = 30$ th percentile.
2. To find the value at the 60th percentile, we note that there are ten data values, and 60% of ten is six. Thus, the number we want is greater than exactly six of the data values. Thus, the 60th percentile is 15.

In each of the examples above, the computations were made easier by the fact that the we were looking for percentiles that "came out evenly" with respect to the number of values in our dataset. Things don't always work out so cleanly. Further, different sources will define the term *percentile* in different ways. In fact, Google Sheets has *three* built-in functions for finding percentiles, none of which uses our definition. Some of the definitions you'll see differ in the inequality that is used. Ours uses "less than or equal to," while others use "less than" (these correspond roughly to Google Sheets' 'PERCENTILE.INC' and 'PERCENTILE.EXC'). Some of them use different methods for *interpolating* values. (This is what we did when we first computed medians without technology; if there were an even number of data values in our dataset, found the mean of the two values in the middle. This is an example of interpolation. Most computerized methods use this technique.) Other definitions don't interpolate at all, but instead choose the closest actual data value to the theoretical value. Fortunately, for large datasets, the differences among the different techniques become very small.

So, with all these different possible definitions in play, what will we use? For small datasets, if you're asked to compute something involving percentiles *without technology*, use the technique we used in the previous example. In all other cases, we'll keep things simple by using the built-in 'PERCENTILE' and 'PERCENTRANK' functions in Google Sheets (which do the same thing as the 'PERCENTILE.INC' and 'PERCENTRANK.INC' functions; they're "inclusive, interpolating" definitions).

VIDEO
Using RANK, PERCENTRANK, and PERCENTILE in Google Sheets

Example 2 **Using Google Sheets to Compute Percentiles: Average SAT Scores**

The data in "AvgSAT" contains the average SAT score for students attending every institution of higher learning in the US for which data is available.

1. What score is at the 3rd quartile?
2. What score is at the 40th percentile?
3. At what percentile is Albion College in Michigan (average SAT: 1132)?
4. At what percentile is Oregon State University (average SAT: 1205)?

Solution

1. The 3rd quartile is the 75th percentile, so we'll use the PERCENTILE function. Click on an empty cell, and type "=PERCENTILE(". Next, enter the data: click on the letter at the top of the column containing the average SAT scores. Then, tell the function which percentile we want; it needs to be entered as a decimal. So, type a comma (to separate the two pieces of information we're giving this function), then type "0.75" and close the parentheses with ")". The result should look like this (assuming the data are in column C): "=PERCENTILE(C:C, 0.75)". When you hit the enter key, the formula will be replaced with the average SAT score at the 75th percentile: 1199.
2. Using the PERCENTILE function, we find that an average SAT of 1100 is at the 40th percentile.
3. Since we want to know the percentile for a particular score, we'll use the PERCENTRANK function. Like the PERCENTILE function, we need to give PERCENTRANK two pieces of information: the data, and the value we care about. So, click on an empty cell, type "=PERCENTRANK(", and then click on the letter at the top of the column containing the data. Next, type a comma and then the value we want to find the percentile for; in this case, we'll type ", 1132". Finally, close the parentheses with ")" and hit the enter key. The formula will be replaced with the information we want: an average SAT of 1132 is at the 54th percentile.
4. Using the PERCENTRANK function, an average SAT of 1205 is at the 76.1th percentile.

Example 3 **Using Google Sheets to Compute Percentiles: In-State Tuition**

The dataset "InState" contains the in-state tuitions of every college and university in the country that reported that data to the Department of Education. Use that dataset to answer these questions.

1. What tuition is at the second quintile?

2. What tuition is at the 95th percentile?
3. At what percentile is Walla Walla University in Washington (in-state tuition: $28,035)?
4. At what percentile is the College of Saint Mary in Nebraska (in-state tuition: $20,350)?
   Solution
   1. The second quintile is the 40th percentile; using PERCENTILE in Google Sheets, we get $8,400.
   2. Using PERCENTILE again, we get $44,866.
   3. Using PERCENTRANK, we get the 81.6th percentile.
   4. Using PERCENTRANK, we get the 74.8th percentile.

Key Terms

- percentile
- quartile
- quintile
- quantile

Key Concepts

- The percentile rank of a data value is the percentage of all values in the dataset that are less than or equal to the given value.

Videos

- Using RANK, PERCENTRANK, and PERCENTILE in Google Sheets

8.6 The Normal Distribution

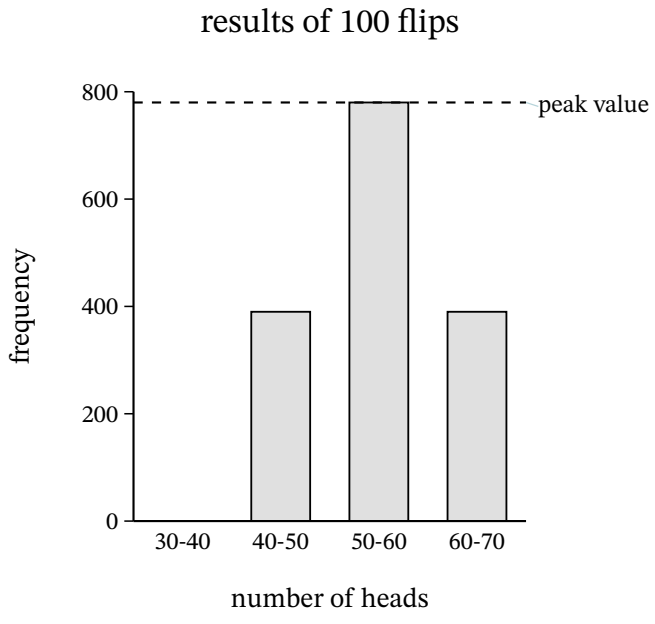


Figure 33: Symmetric, bell-shaped distributions arise naturally in many different situations, including coin flips.

Learning Objectives

After completing this section, you should be able to:

1. Describe the characteristics of the normal distribution.
2. Apply the 68-95-99.7 percent groups to normal distribution datasets.
3. Use the normal distribution to calculate a z-score.
4. Find and interpret percentiles and quartiles.

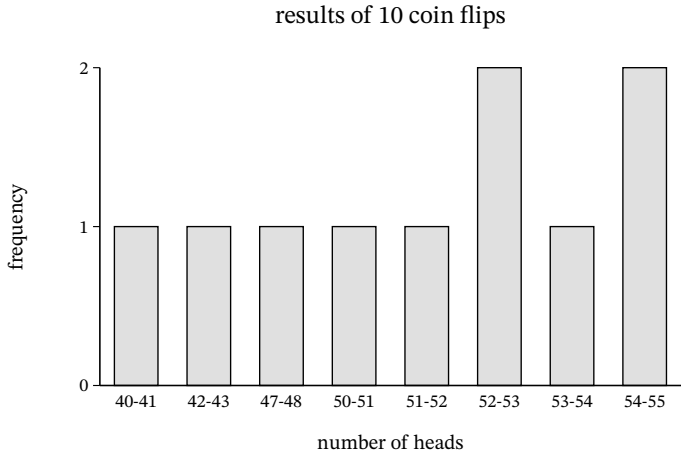
Many datasets that result from natural phenomena tend to have histograms that are symmetric and bell-shaped. Imagine finding yourself with a whole lot of time on your hands, and nothing to keep you entertained but a coin, a pencil, and paper. You decide to flip that coin 100 times and record the number of heads. With nothing else to do, you repeat the experiment ten times total. Using a computer to simulate this series of experiments, here's a sample for the number of heads in each trial:

54, 51, 40, 42, 53, 50, 52, 52, 47, 54

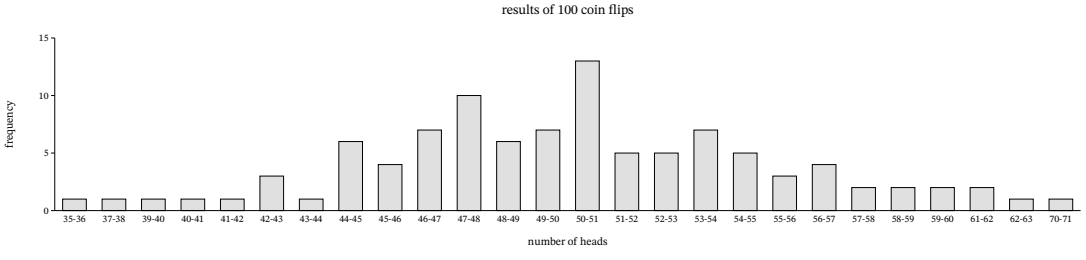
It makes sense that we'd get somewhere around 50 heads when we flip the coin 100 times, and it makes sense that the result won't *always* be exactly 50 heads. In our results, we can see numbers that were generally near 50 and not always 50, like we thought.

Moving Toward Normality

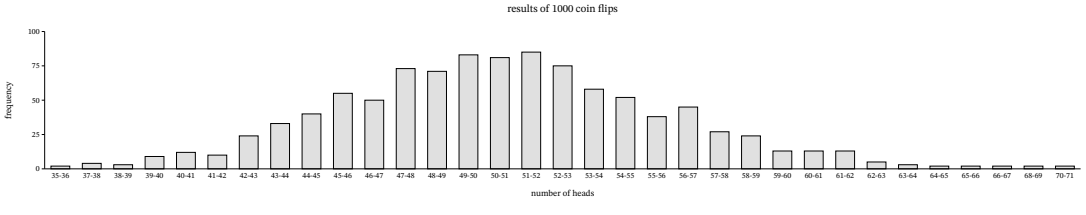
Let's take a look at a histogram for the dataset in our section opener:



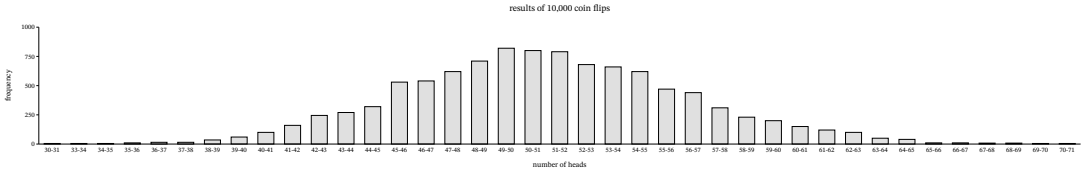
This is interesting, but the data seem pretty sparse. There were no trials where you saw between 43 and 47 heads, for example. Those results don't seem impossible; we just didn't flip enough times to give them a chance to pop up. So, let's do it again, but this time we'll perform 100 coin flips 100 times. Rather than review all 100 results, which could be overwhelming, let's instead visualize the resulting histogram.



From the histogram, we see that most of the trials resulted in between, say, 44 and 56 heads. There were some more unusual results: one trial resulted in 70 heads, which seems really unlikely (though still possible!). But we're starting to maybe get a sense of the distribution. More data would help, though. Let's simulate another 900 trials and add them to the histogram!



We can still see that 70 is a really unusual observation, though we came close in another trial (one that had 68 heads). Now, the distribution is coming more into focus: It looks quite symmetric and bell-shaped. Let's just go ahead and take this thought experiment to an extreme conclusion: 10,000 trials.



The distribution is pretty clear now. Distributions that are symmetric and bell-shaped like this pop up in all sorts of natural phenomena, such as the heights of people in a population, the circumferences of eggs of a particular bird species, and the numbers of leaves on mature trees of a particular species. All of these have bell-shaped distributions. Additionally, the results of many types of repeated experiments generally follow this same pattern, as we saw with the coin-flipping example; this fact is the basis for much of the work done by statisticians. It's a fact that's important enough to have its own name: the Central Limit Theorem.

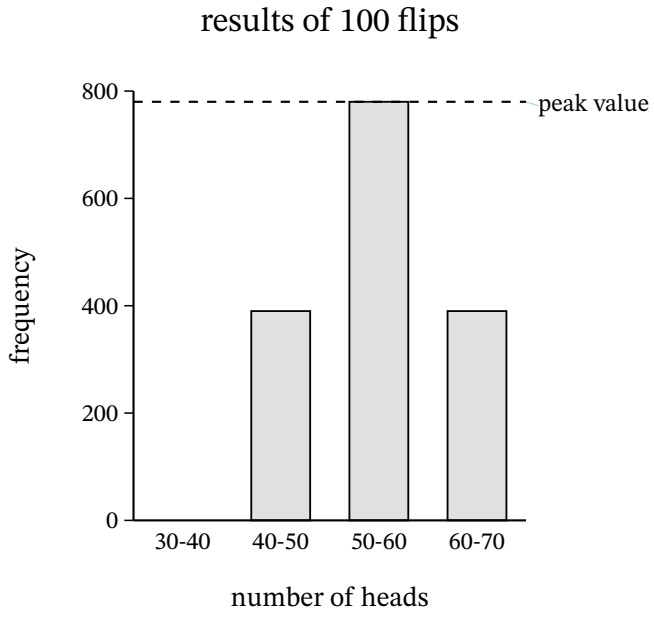
PEOPLE IN MATHEMATICS

John Kerrich

Having enough time on your hands to actually perform this coin-flipping experiment may sound far-fetched, but the English mathematician John Kerrich found himself in just such a situation. While he was studying abroad in Denmark in 1940, that country was invaded by the Germans. Kerrich was captured and placed in an internment camp, where he remained for the duration of the war. Kerrich knew that he had all kinds of time on his hands. He also studied statistics, so he knew what should happen *theoretically* if he flipped a coin many, many times. He also knew of nobody who had ever *tested* that theory with an actual, large-scale experiment. So, he did it: While he was incarcerated, Kerrich flipped a regular coin 10,000 times and recorded the results. Sure enough, the theory held up!

The Normal Distribution

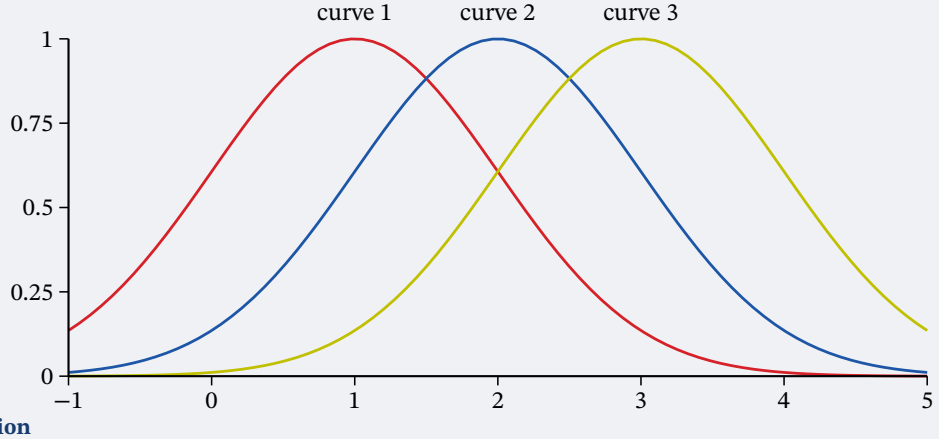
In the coin flipping example above, the distribution of the number of heads for 10,000 trials was close to perfectly symmetric and bell-shaped:



Because distributions with this shape appear so often, we have a special name for them: **normal distributions**. Normal distributions can be completely described using two numbers we've seen before: the *mean* of the data and the *standard deviation* of the data. You may remember that we described the mean as a measure of centrality; for a normal distribution, the mean tells us exactly where the center of the distribution falls. The peak of the distribution happens at the mean (and, because the distribution is symmetric, it's also the median). The standard deviation is a measure of dispersion; for a normal distribution, it tells us how spread out the histogram looks. To illustrate these points, let's look at some examples.

Example 1 **Identifying the Mean of a Normal Distribution**

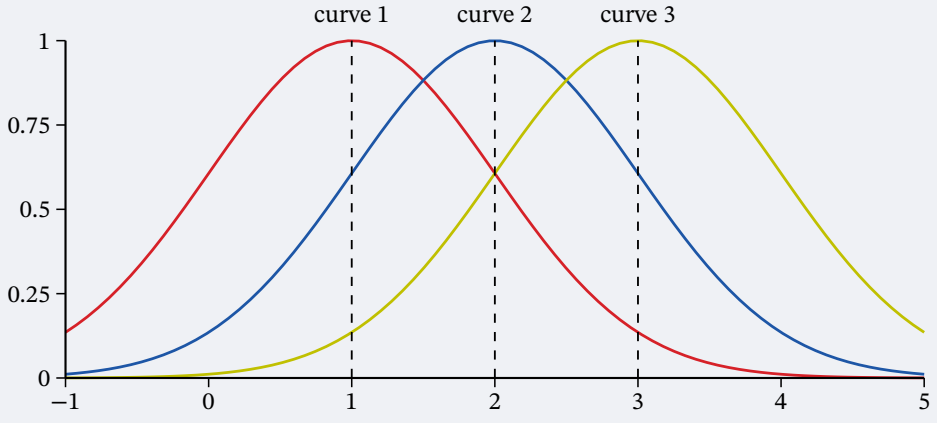
This graph shows three normal distributions. What are their means?



Solution

Step 1: Take a look at the three curves on the graph. Since the mean of a normal distribution

occurs at the peak, we should look for the highest point on each distribution. Let's draw a line from each curve's peak down to the axis, so we can see where these peaks occur:



Step 2: The peak of the red (leftmost) distribution occurs over the number 1 on the horizontal axis. Thus, the mean of the red distribution is 1. Similarly, the mean of the blue (middle) distribution is 2, and the mean of the yellow (rightmost) distribution is 3.

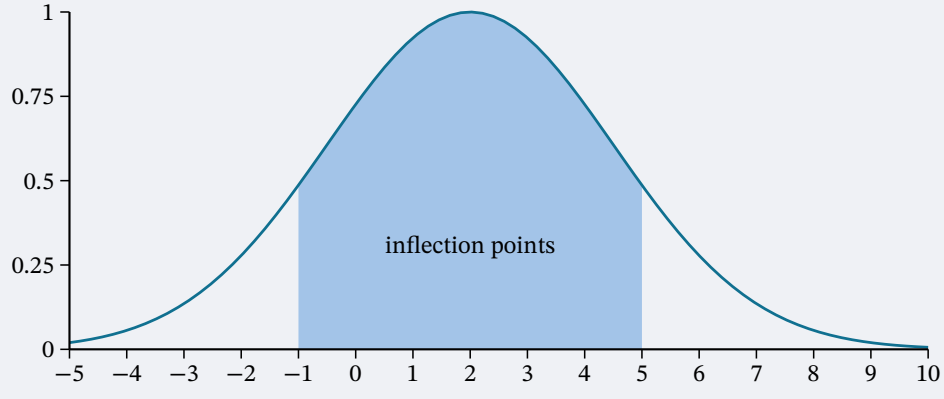
Example 2 **Identifying the Standard Deviation of a Normal Distribution**

This graph shows three distributions, all with mean 2. What are their standard deviations?

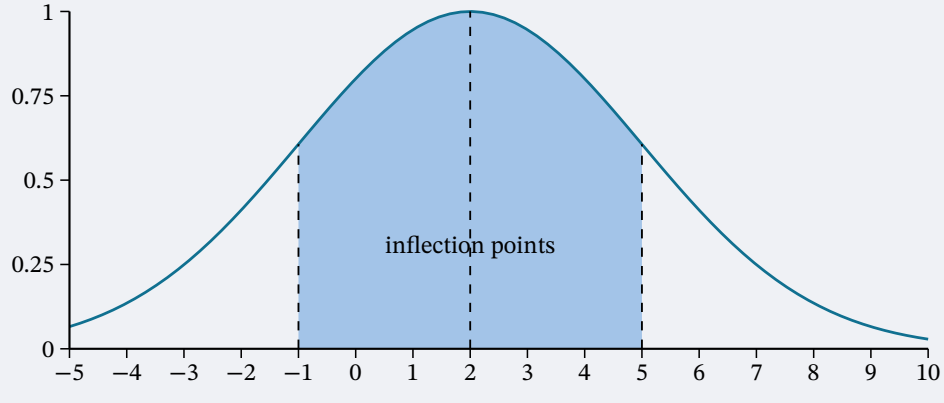
Solution

Step 1: Identifying the standard deviation from a graph can be a little bit tricky. Let's focus in on the yellow (lowest peaked) curve:

Step 2: Notice that the graph curves downward in the middle, and curves upwards on the ends. Highlighted in red is the part that curves downward and in green, the part that curves upward:

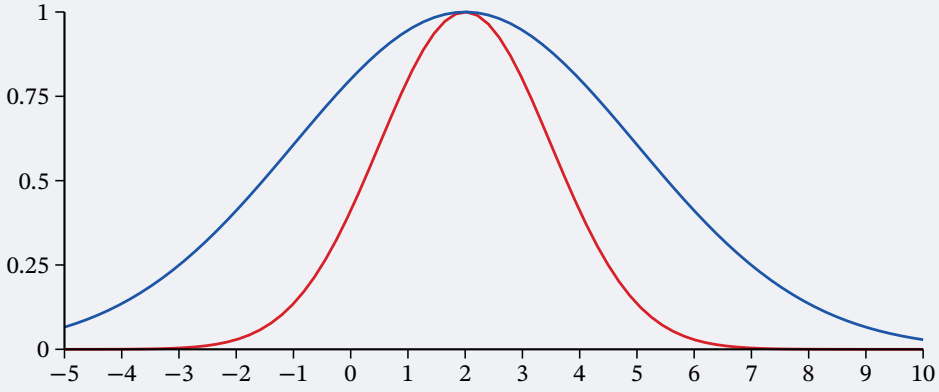


Step 3: The places where the graph changes from curving up to curving down (or vice versa) are called *inflection points*. Let's identify where those occur by dropping a line straight down from each:



Step 4: We can estimate that the inflection points occur at $x = -1$ and $x = 5$ the mean is at $x = 2$ (as shown by the middle dotted line). The difference between the mean and the location of either inflection point is the standard deviation; since $5 - 2 = 2 - (-1) = 3$, we conclude that the standard deviation of the green distribution is 3.

Step 5: Now, looking at the other two graphs, let's first identify the inflection points:

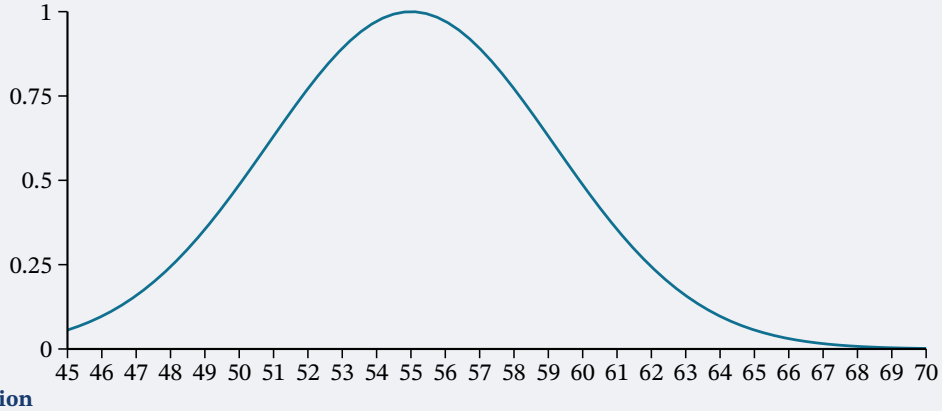


Step 6: The red (tallest peaked) distribution has inflection points at 1 and 3, and the mean is 2. Thus, the standard deviation of the red distribution is $3 - 2 = 2 - 1 = 1$. The blue (lower peaked) distribution has inflection points at 0 and 4, and its mean is also 2. So, the standard deviation of the blue distribution is 2.

Let's put it all together to identify a completely unknown normal distribution.

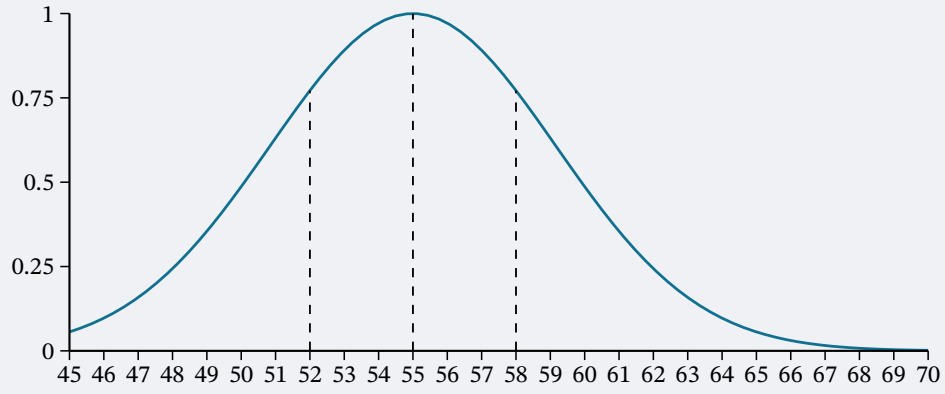
Example 3 **Identifying the Mean and Standard Deviation of a Normal Distribution**

Using the graph, identify the mean and standard deviation of the normal distribution.



Solution

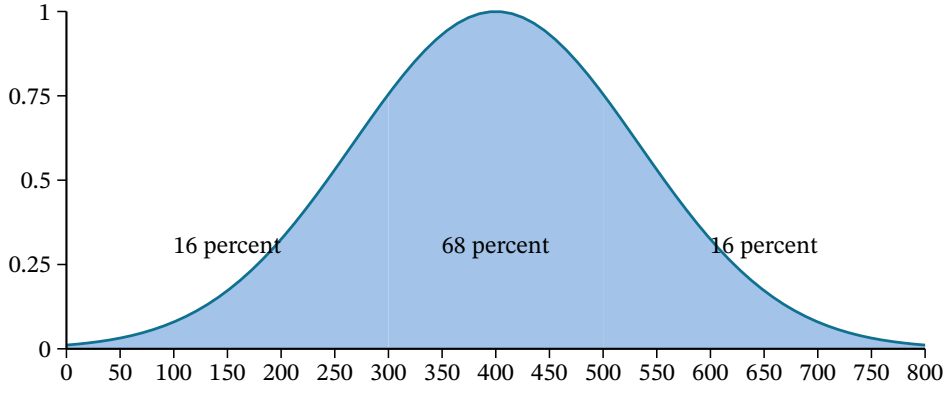
Step 1: Let's start by putting dots on the graph at the peak and at the inflection points, then drop lines from those points straight down to the axis:



Step 2: From the red (middle) line, we can see that the mean of this distribution is 55. The blue (outermost) lines are each 3 units away from the mean (at 52 and 58), so the standard deviation is 3.

Properties of Normal Distributions: The 68-95-99.7 Rule

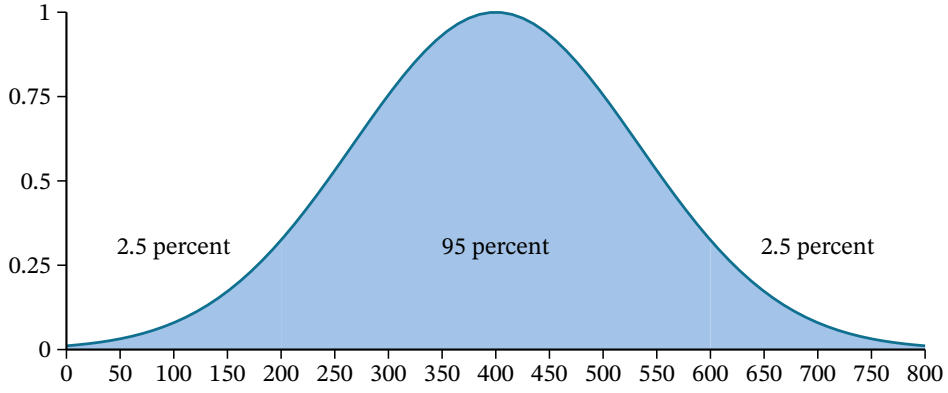
The most important property of normal distributions is tied to its standard deviation. If a dataset is perfectly normally distributed, then 68% of the data values will fall within one standard deviation of the mean. For example, suppose we have a set of data that follows the normal distribution with mean 400 and standard deviation 100. This means 68% of the data would fall between the values of 300 (one standard deviation *below* the mean: $400 - 100 = 300$) and 500 (one standard deviation *above* the mean: $400 + 100 = 500$). Looking at the histogram below, the shaded area represents 68% of the total area under the graph and above the axis:



Since 68% of the area is in the shaded region, this means that $100\% - 68\% = 32\%$ of the area is found in the unshaded regions. We know that the distribution is symmetric, so that 32% must be divided evenly into the two unshaded tails: 16% in each.

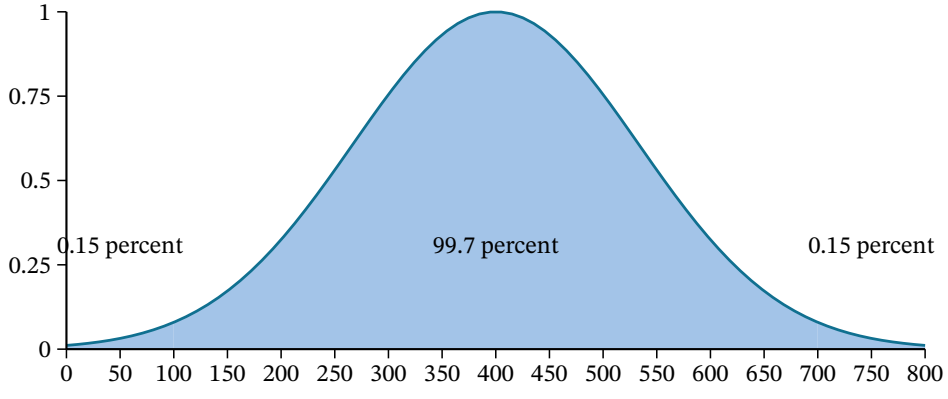
Of course, datasets in the real world are never perfect; when dealing with actual data that seem to follow a symmetric, bell-shaped distribution, we'll give ourselves a little bit of wiggle room and say that *approximately* 68% of the data fall within one standard deviation of the mean.

The rule for one standard deviation can be extended to two standard deviations. Approximately 95% of a normally distributed dataset will fall within 2 standard deviations of the mean. If the mean is 400 and the standard deviation is 100, that means 95% calculation describes the way we compute standardized scores. (two standard deviations *below* the mean: $400 - 2 \times 100 = 200$) and 600 (two standard deviations *above* the mean: $400 + 2 \times 100 = 600$). We can visualize this in the following histogram:



As before, since 95% of the data are in the shaded area, that leaves 5% of the data to go into the unshaded tails. Since the histogram is symmetric, half of the 5% (that's 2.5%) is in each.

We can even take this one step further: 99.7% of normally distributed data fall within 3 standard deviations of the mean. In this example, we'd see 99.7% of the data between 100 (calculated as $400 - 3 \times 100 = 100$) and 700 (calculated as $400 + 3 \times 100 = 700$). We can see this in the histogram below, although you may need to squint to find the unshaded bits in the tails!



This observation is formally known as the **68-95-99.7 Rule**.

Example 4 **Using the 68-95-99.7 Rule to Find Percentages**

1. If data are normally distributed with mean 8 and standard deviation 2, what percent of the data falls between 4 and 12?

2. If data are normally distributed with mean 25 and standard deviation 5, what percent of the data falls between 20 and 30?
3. If data are normally distributed with mean 200 and standard deviation 15, what percent of the data falls between 155 and 245?

Solution

1. Let's look at a table that sets out the data values that are even multiples of the standard deviation (SD) above and below the mean:

mean − 3 × SD	mean − 2 × SD	mean − 1 × SD	Mean	mean + 1 × SD	mean + 2 × SD	mean + 3 × SD
2	4	6	8	10	12	14

Since 4 and 12 represent two standard deviations above and below the mean, we conclude that 95% of the data will fall between them.

1. Let's build another table:

mean − 3 × SD	mean − 2 × SD	mean − 1 × SD	Mean	mean + 1 × SD	mean + 2 × SD	mean + 3 × SD
10	15	20	25	30	35	40

We can see that 20 and 30 represent one standard deviation above and below the mean, so 68% of the data fall in that range.

1. Let's make one more table:

mean − 3 × SD	mean − 2 × SD	mean − 1 × SD	Mean	mean + 1 × SD	mean + 2 × SD	mean + 3 × SD
155	170	185	200	215	230	245

Since 155 and 245 are three standard deviations above and below the mean, we know that 99.7% of the data will fall between them.

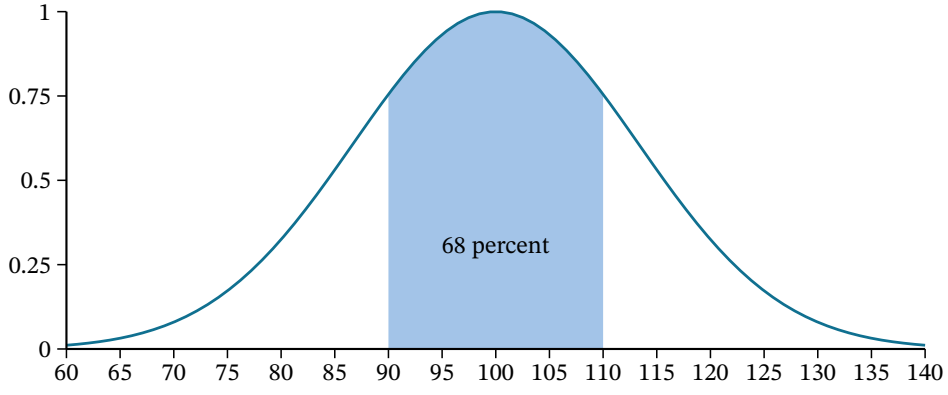
Example 5 **Using the 68-95-99.7 Rule to Find Data Values**

1. If data are distributed normally with mean 100 and standard deviation 20, between what two values will 68% of the data fall?
2. If data are distributed normally with mean 0 and standard deviation 15, between what two values will 95% of the data fall?
3. If data are distributed normally with mean 14 and standard deviation 2, between what two values will 99.7% of the data fall?

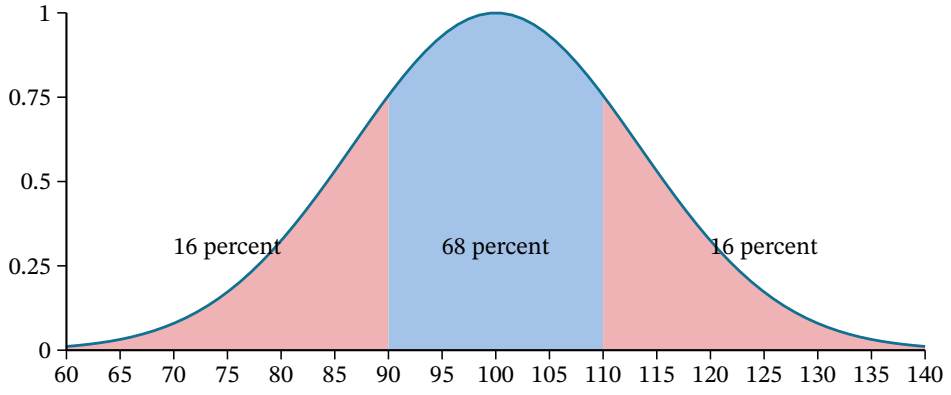
Solution

1. The 68-95-99.7 Rule tells us that 68% of the data will fall within one standard deviation of the mean. So, to find the values we seek, we'll add and subtract one standard deviation from the mean: $100 - 1 \times 20 = 80$ and $100 + 1 \times 20 = 120$. Thus, we know that 68% of the data fall between 80 and 120.
2. Using the 68-95-99.7 Rule again, we know that 95% of the data will fall within 2 standard deviations of the mean. Let's add and subtract two standard deviations from that mean: $0 - 2 \times 15 = -30$ and $0 + 2 \times 15 = 30$. So, 95% of the data will fall between -30 and 30.
3. Once again, the 68-95-99.7 Rule tells us that 99.7% of the data will fall within three standard deviations of the mean. So, let's add and subtract three standard deviations from the mean: and $14 + 3 \times 2 = 20$. Thus, we conclude that 99.7% of the data will fall between 8 and 20.

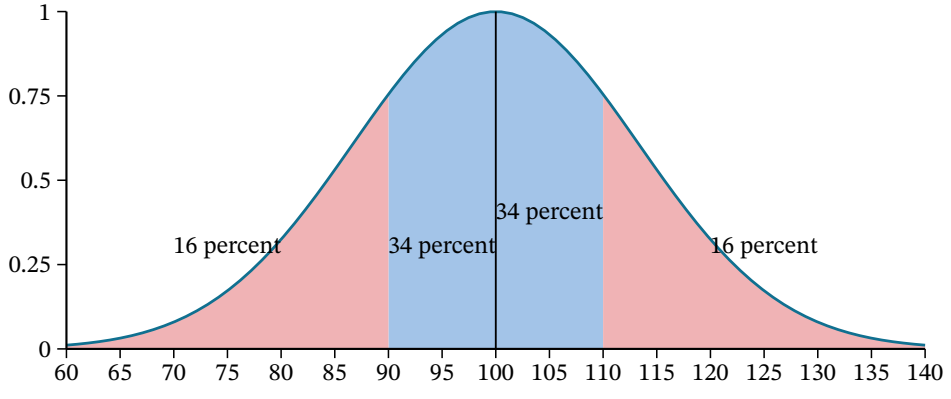
There are more problems we can solve using the 68-95-99.7 Rule. but first we must understand what the rule implies. Remember, the rule says that 68% of the data falls within one standard deviation of the mean. Thus, with normally distributed data with mean 100 and standard deviation 10, we have this distribution:



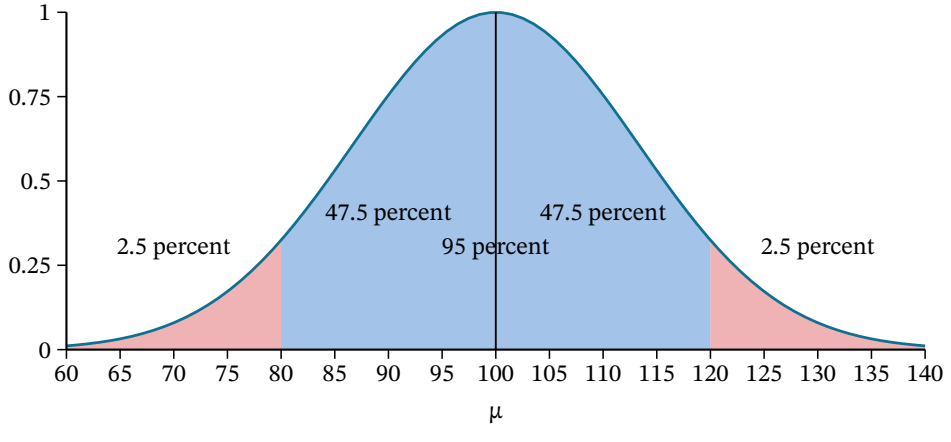
Since we know that 68% of the data lie within one standard deviation of the mean, the implication is that 32% of the data must fall *beyond* one standard deviation away from the mean. Since the histogram is symmetric, we can conclude that half of the 32% (or 16%) is more than one standard deviation *above* the mean and the other half is more than one standard deviation *below* the mean:

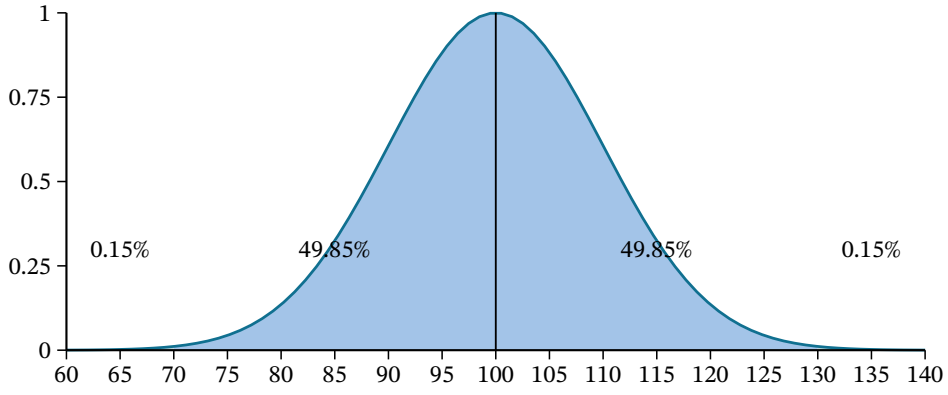


Further, we know that the middle 68% can be split in half at the peak of the histogram, leaving 34% on either side:

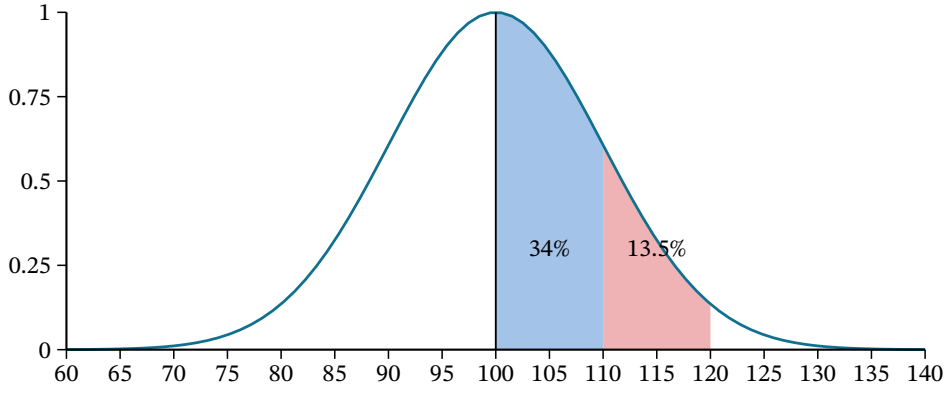


So, just the "68" part of the 68-95-99.7 Rule gives us four other proportions in addition to the 68% in the rule. Similarly, the "95" and "99.7" parts each give us four more proportions:





We can put all these together to find even more complicated proportions. For example, since the proportion between 100 and 120 is 47.5% and the proportion between 100 and 110 is 34%, we can subtract to find that the proportion between 110 and 120 is $47.5 - 34 = 13.5\%$:



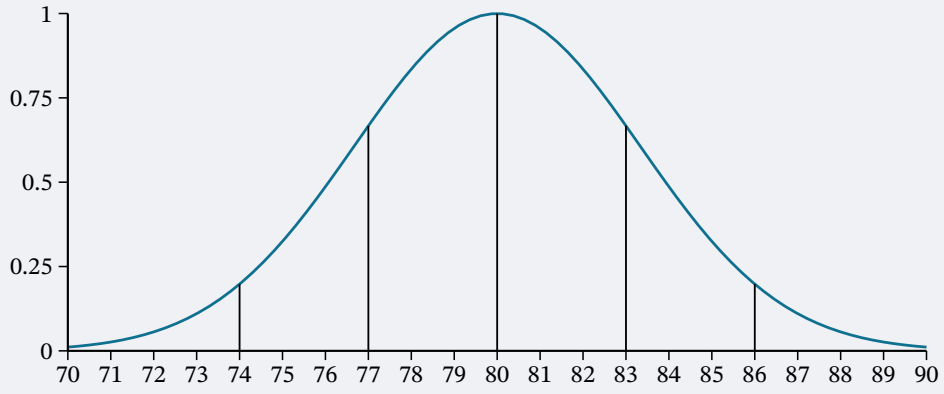
Example 6 **Finding Other Proportions Using the 68-95-99.7 Rule**

Assume that we have data that are normally distributed with mean 80 and standard deviation 3.

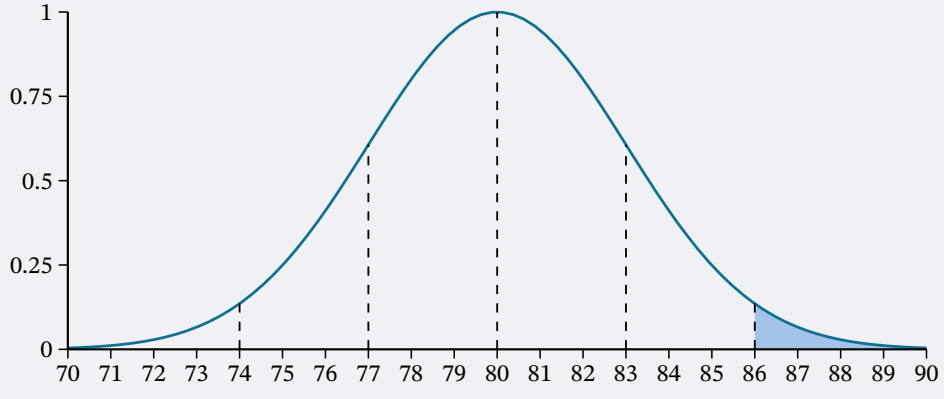
1. What proportion of the data will be greater than 86?
2. What proportion of the data will be between 74 and 77?
3. What proportion of the data will be between 74 and 83?

Solution

Before we can answer these questions, we must mark off sections that are multiples of the standard deviation away from the mean:

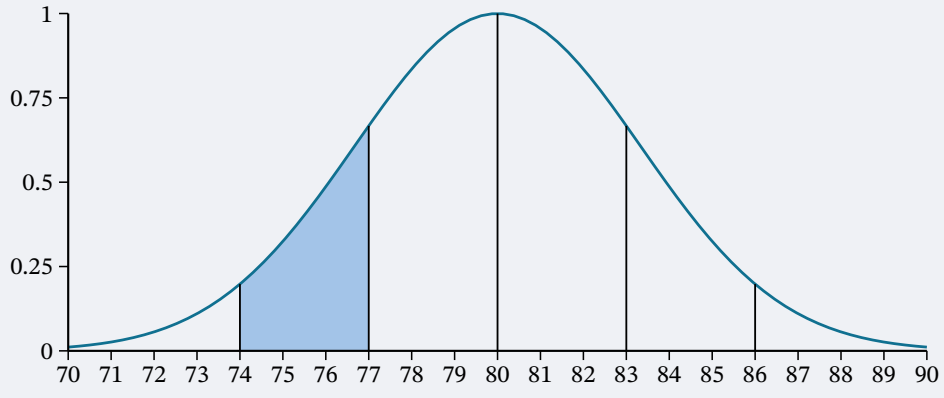


1. To figure out what proportion of the data will be greater than 86, let's start by shading in the area of data that are above 86 in our figure, or the data more than two standard deviations above the mean.

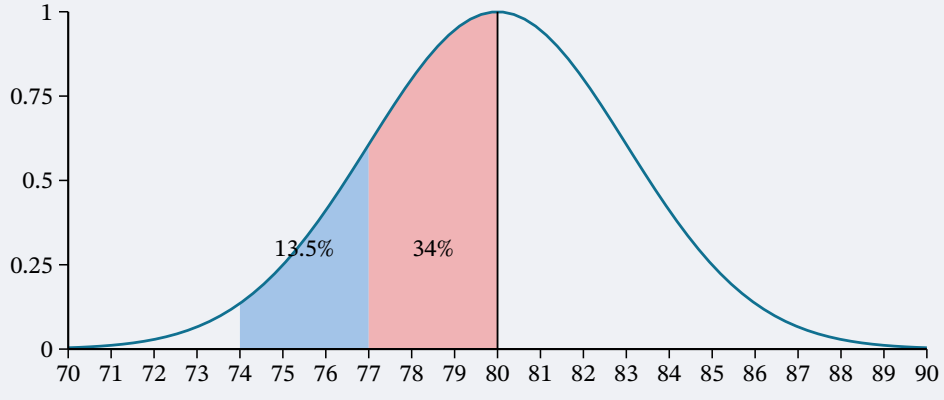


We saw in that this proportion is 2.5%.

1. To figure out what proportion of the data will be between 74 and 77, let's start by shading in that area of data. These are data that are more than one but less than two standard deviations below the mean.

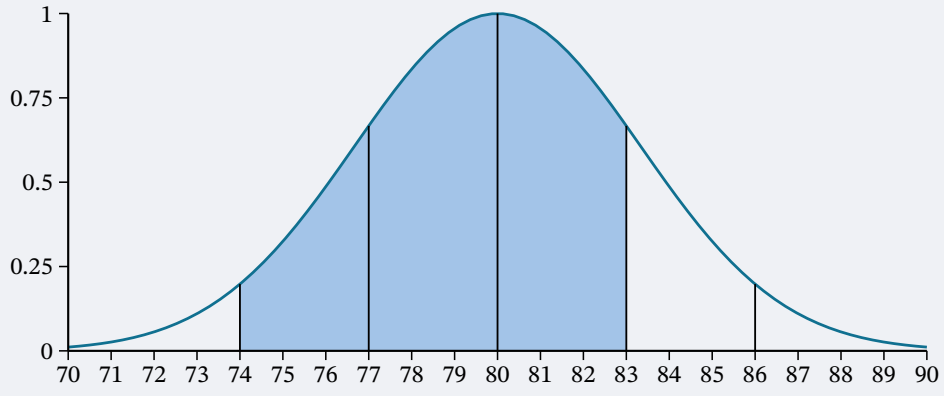


From , we know that the proportion of data less than two standard deviations below the mean is 47.5%. And, from YOUR TURN 8.33, we know that 34% of the data is less than one standard deviation below the mean:

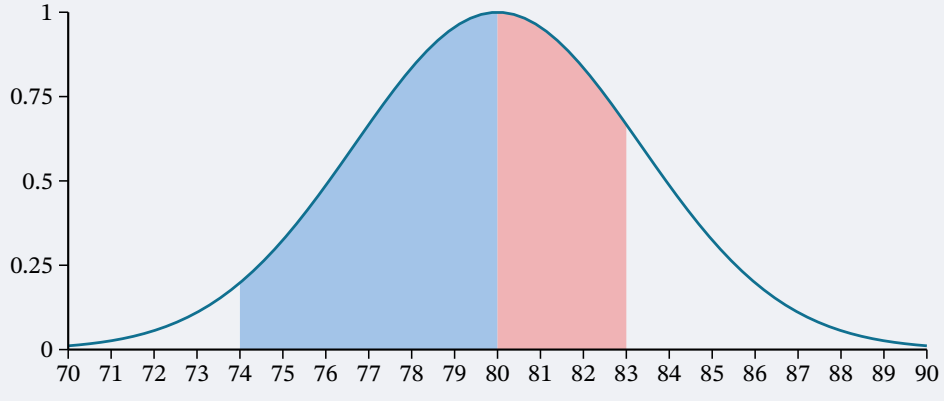


Subtracting, we see that the proportion of data between 74 and 77 is 13.5%.

2. To figure out what proportion of the data will be between 74 and 83, let's start by shading in that area of data in our figure.



Next, we'll break this region into two pieces at the mean:



From , we know the blue (leftmost) region represents 47.5% of the data. And, using YOUR TURN 8.33, we get that the red (rightmost) region covers 34% of the data. Adding those together, the proportion we want is 81.5%.

Standardized Scores

When we want to apply the 68-95-99.7 Rule, we must first figure out how many standard deviations above or below the mean our data fall. This calculation is common enough that it has its own name: the **standardized score**. Values above the mean have positive standardized scores, while those below the mean have negative standardized scores. Since it's common to use the letter z to represent a standard score, this value is also often referred to as a ***z*-score**.

So far, we've only really considered z-scores that are whole numbers, but in general they can be any number at all. For example, if we have data that are normally distributed with mean 80 and standard deviation 6, the value 85 is five units above the mean, which is less than one standard deviation. Dividing by the standard deviation, we get $\frac{5}{6}$. Since 85 is $\frac{5}{6}$ of one standard deviation

above the mean, we'd say that the standardized score for 85 is $z = \frac{5}{6}$ (which is positive, since $85 > 80$). This calculation describes the way we compute standardized scores.

FORMULA

If x is a member of a normally distributed dataset with mean μ and standard deviation σ, then the standardized score for x is

$$z = \frac{x - \mu}{\sigma}.$$

If you know a z-score but not the original data value x, you can find it by solving the previous equation for x:

$$x = \mu + z \times \sigma.$$

CHECKPOINT

The symbols μ and σ are the Greek letters *mu* and *sigma*. They are the analogues of the English letters *m* and *s*, which stand for *mean* and *standard deviation*.

If you convert every data value in a dataset into its z-score, the resulting set of data will have mean 0 and standard deviation 1. This is why we call these *standardized scores*: the normal distribution with mean 0 and standard deviation 1 is often called the *standard normal distribution*.

Example 7 **Standardizing Data**

Suppose we have data that are normally distributed with mean 50 and standard deviation 6. Compute the standardized scores (rounded to three decimal places) for these data values:

1. 52
2. 40
3. 68

Solution

For each of these, we'll plug the given values into the formula. Remember, the mean is $\mu = 50$ and the standard deviation is $\sigma = 6$:

1. $z = \frac{x-\mu}{\sigma} = \frac{52-50}{6} = 0.333$
2. $z = \frac{x-\mu}{\sigma} = \frac{40-50}{6} = -1.667$
3. $z = \frac{x-\mu}{\sigma} = \frac{68-50}{6} = 3$

Example 8 **Converting Standardized Scores to Original Values**

Suppose we have data that are normally distributed with mean 10 and standard deviation 2. Convert the following standardized scores into data values.

1. 1.4

2. -0.9
3. 3.5

Solution

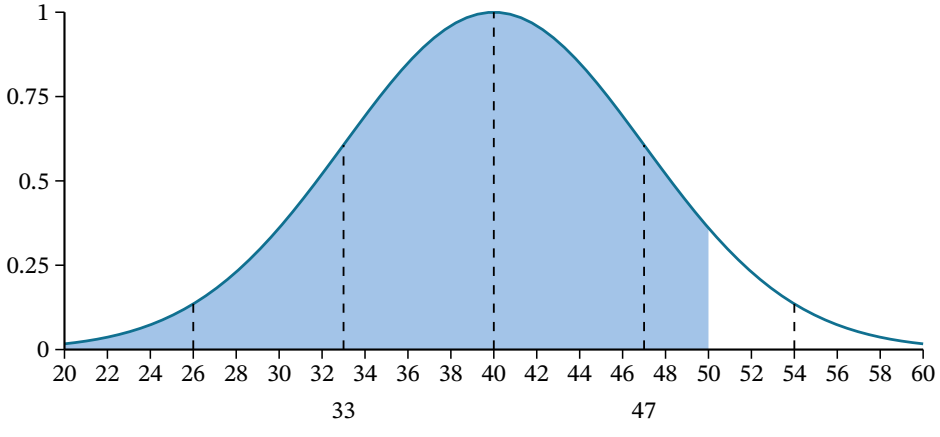
We'll use the formula previously introduced to convert z-scores into x-values. In this case, the mean is $\mu = 10$ and the standard deviation is $\sigma = 2$:

1. $x = \mu + z \times \sigma = 10 + 1.4 \times 2 = 12.8$
2. $x = \mu + z \times \sigma = 10 + (-0.9) \times 2 = 8.2$
3. $x = \mu + z \times \sigma = 10 + 3.5 \times 2 = 17$

Using Google Sheets to Find Normal Percentiles

The 68-95-99.7 Rule is great when we're dealing with whole-number z-scores. However, if the z -score is not a whole number, the Rule isn't going to help us. Luckily, we can use technology to help us out. We'll talk here about the built-in functions in Google Sheets, but other tools work similarly.

Let's say we're working with normally distributed data with mean 40 and standard deviation 7, and we want to know at what percentile a data value of 50 would fall. That corresponds to finding the proportion of the data that are less than 50. If we create our histogram and mark off whole-number multiples of the standard deviation like we did before, we'll see why the 68-95-99.7 Rule isn't going to help:



Since 50 doesn't line up with one of our lines, the 68-95-99.7 Rule fails us. Looking back at and , the best we can say is that 50 is between the 84th and 99.5th percentiles, but that's a pretty wide range. Google Sheets has a function that can help; it's called NORM.DIST. Here's how to use it:

1. Click in an empty cell in your worksheet.
2. Type "=NORM.DIST("
3. Inside the parentheses, we must enter a list of four things, separated by commas: the data value, the mean, the standard deviation, and the word "TRUE". These have to be entered in this order!

4. Close the parentheses, and hit Enter. The result is then displayed in the cell; convert it to a percent to get the percentile.

   So, for our example, we should type "=NORM.DIST(50, 40, 7, TRUE)" into an empty cell, and hit Enter. The result is 0.9234362745; converting to a percent and rounding, we can conclude that 50 is at the 92nd percentile. Let's walk through a few more examples.

Example 9 **Using Google Sheets to Find Percentiles**

Suppose we have data that are normally distributed with mean 28 and standard deviation 4. At what percentile do each of the following data values fall?

1. 30
2. 23
3. 35

Solution

1. By entering "=NORM.DIST(30, 28, 4, TRUE)" we find that 30 is at the 69th percentile.
2. By entering "=NORM.DIST(23, 28, 4, TRUE)" we find that 23 is at the 11th percentile.
3. By entering "=NORM.DIST(35, 28, 4, TRUE)" we find that 35 is at the 96th percentile.

Google Sheets can also help us go the other direction: If we want to find the data value that corresponds to a given percentile, we can use the NORM.INV function. For example, if we have normally distributed data with mean 150 and standard deviation 25, we can find the data value at the 30th percentile as follows:

1. Click on an empty cell in your worksheet.
5. Type "=NORM.INV("
6. Inside the parentheses, we'll enter a list of three numbers, separated by commas: the percentile in question expressed as a decimal, the mean, and the standard deviation. These must be entered in this order!
7. Close the parentheses and hit Enter. The desired data value will be in the cell!

   In our example, we want the 30th percentile; converting 30% to a decimal gives us 0.3. So, we'll type "=NORM.INV(0.3, 150, 25)" to get 136.8899872; let's round that off to 137.

Example 10 **Using Google Sheets to Find the Data Value Corresponding to a Percentile**

Suppose we have data that are normally distributed with mean 47 and standard deviation 9. Find the data values (rounded to the nearest tenth) corresponding to these percentiles:

1. 75th (that's the third quartile)

2. 12th
3. 90th

Solution

1. By entering "=NORM.INV(0.75, 47, 9)" we find that 53.1 is at the 75th percentile.
2. By entering "=NORM.INV(0.12, 47, 9)" we find that 36.4 is at the 12th percentile.
3. By entering "=NORM.INV(0.9, 47, 9)" we find that 58.5 is at the 90th percentile.

- normal distribution
- 68-95-99.7 Rule
- standardized score (z-score)

Key Concepts

- Normally distributed data follow a bell-shaped, symmetrical distribution.
- The mean of normally distributed data falls at the peak of the distribution. The standard deviation of normally-distributed data is the distance from the peak to either of the inflection points.
- Data that are normally distributed follow the 68-95-99.7 Rule, which says that approximately 68% of the data fall within one standard deviation of the mean, 95% fall within two standard deviations, and 99.7% fall within three standard deviations.
- The z-score for a data value is the number of standard deviations that value falls above (or below, if the z-score is negative) the mean.
- We can use the normal distribution to estimate percentiles.

Formulas

If x is a member of a normally distributed dataset with mean μ and standard deviation σ, then the standardized score for x is

$$z = \frac{x - \mu}{\sigma}.$$

If you know a z-score but not the original data value x, you can find it by solving the previous equation for x:

$$x = \mu + z \times \sigma.$$

8.7 Applications of the Normal Distribution

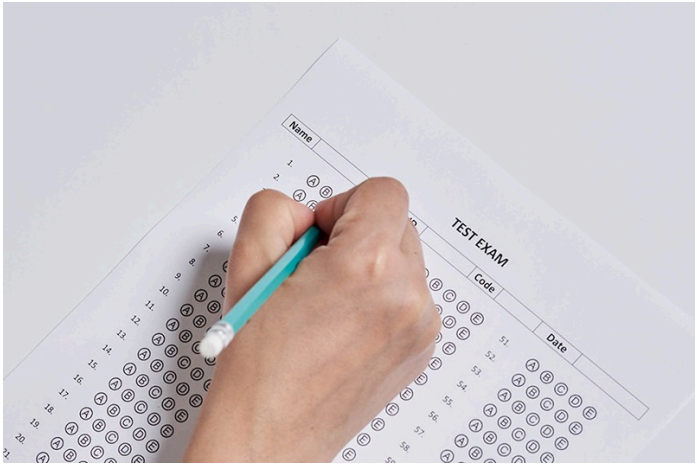


Figure 64: Standardized test results generally adhere to the normal distribution.

Learning Objectives

After completing this section, you should be able to:

1. Apply the normal distribution to real-world scenarios.

As we saw in The Normal Distribution, the word "standardized" is closely associated with the normal distribution. This is why tests like college entrance exams, state achievement tests for K–12 students, and Advanced Placement tests are often called "standardized tests": scores are assigned in a way that forces them to follow a normal distribution, with a mean and standard deviation that are consistent from year to year. Standardization also allows people like college admissions officers to directly compare an applicant who took the ACT (a college entrance exam) to an applicant who instead chose to take the SAT (a different college entrance exam). Standardization allows us to compare individuals from different groups; this is among the most important applications of the normal distribution. We'll explore this and other real-world uses of the normal distribution in this section.

College Entrance Exams

There are two good ways to compare two data values from different groups: using z-scores and using percentiles. The two methods will always give consistent results (meaning that we won't find, for example, that the first value is better using z-scores but the second value is better using percentiles), so use whichever method is more comfortable for you.

Example 1 **Evaluating College Entrance Exam Scores**

According to the Digest of Education Statistics, composite scores on the SAT have mean 1060 and standard deviation 195, while composite scores on the ACT have mean 21 and standard deviation 5.

1. At what percentile would an SAT score of 990 fall?
2. What is the z-score of an ACT score of 27?
3. Which is better: a score of 1450 on the SAT or 29 on the ACT?

Solution

1. Using Google Sheets, we can answer this question with the formula "=NORM.DIST(990, 1060, 195, TRUE)". A score of 990 would fall at the 36th percentile.
2. Using the formula $z = \frac{x-\mu}{s}$, we get $z = \frac{27-21}{5} = 1.2$.
3. Let's compare the values using both percentiles and z-values:

Percentiles: Using "=NORM.DIST(1450, 1060, 195, TRUE)" we find that an SAT score of 1450 is at the 98th percentile. Meanwhile, by entering "=NORM.DIST(29, 21, 5, TRUE)" we see that an ACT score of 29 is around the 95th percentile. Since it's at a higher percentile, we can conclude that an SAT score of 1450 is better than an ACT score of 29.

***z*-scores:** Using the formula, we see that the z-score for an SAT score of 1450 is $z = \frac{1450-1060}{195} = 2$, while the z-score for an ACT score of 29 is $z = \frac{29-21}{5} = 1.6$. Since it has a higher z-score, an SAT score of 1450 is better than an ACT score of 29.

Coin flipping

In the opening of The Normal Distribution, we saw that the number of heads we get when we flip a coin 100 times is distributed normally. It can be shown that if n is the number of flips, then the mean of that distribution is $\frac{n}{2}$ and the standard deviation is $\frac{\sqrt{n}}{2}$ (as long as $n \geq 20$). So, for 100 flips, the mean of the distribution is 50 and the standard deviation is 5. In that opening example, one of our early runs gave us 70 heads in 100 flips, which we noted seemed unusual. Using the normal distribution, we can identify exactly how unusual that really is. Using Google Sheets, the formula "=NORM.DIST(70, 50, 5, TRUE)" gives us 0.999968, which is the 99.997th percentile! How is that useful? Suppose you need to test whether a coin is fair, and so you flip it 100 times. While we might be suspicious if we get 70 heads out of the 100 flips, we now have a numerical measure for how unusual that is: If the coin were fair, we would expect to see 70 heads (or more) only $100 - 99.9968 = 0.0032\%$ of the time. That's really unlikely! Analysis like this is related to *hypothesis testing*, an important application of statistics in the sciences and social sciences.

Example 2 **Flipping a Coin**

Let's say we flip a coin 64 times and count the number of heads.

1. What would be the mean of the corresponding distribution?
2. What would be the standard deviation of the corresponding distribution?
3. Suppose we got 25 heads, which seems a little low. At what percentile would 25 heads fall?

Solution

1. Since $n = 64$, the mean is $\frac{64}{2} = 32$.
2. Again using $n = 64$, we get a standard deviation of $\frac{\sqrt{64}}{2} = 4$.
3. Using "=NORM.DIST(25, 32, 4, TRUE)", we see that 25 heads is at the 4th percentile.
   Reading and Interpreting Scatter Plots

Analyzing Data That Are Normally Distributed

Whenever we're working with a dataset that has a distribution that looks symmetric and bell-shaped, we can use techniques associated with the normal distribution to analyze the data.

Example 3 **Using Normal Techniques to Analyze Data**

The data in "AvgSAT" contains the average SAT score for students attending every institution of higher learning in the United States for which data is available. In Example 8 in Section 8.2, we created a histogram for these data:

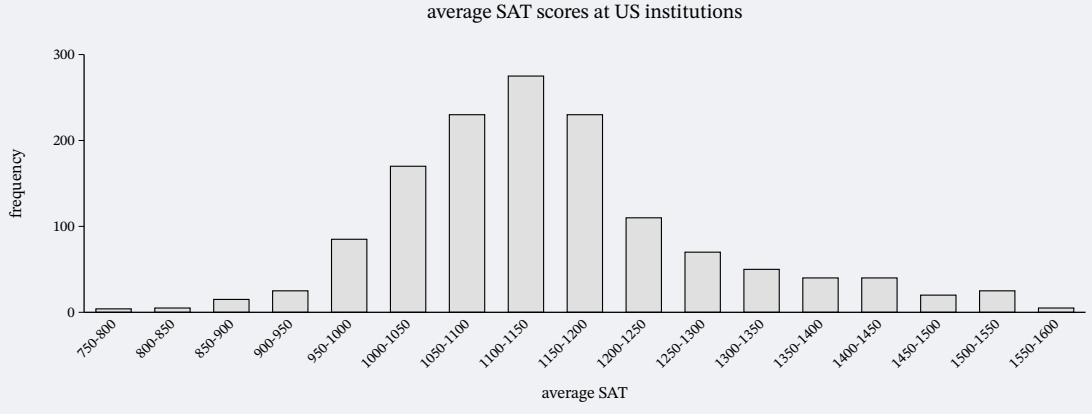


Figure 65: (data source: https://data.ed.gov)

This distribution is fairly symmetric (it's just a little right-skewed) and bell-shaped, so we can use normal distribution techniques to analyze the data.

1. What is the mean of these average SAT scores?
2. What is the standard deviation of these SAT scores?
3. Using the answers to the previous two questions, use NORM.DIST in Google Sheets to estimate at what percentile the University at Buffalo in New York (average SAT: 1250) falls.
4. Use PERCENTRANK to find the actual percentile of the University at Buffalo, and see how close the estimate in the previous question came.

Solution

1. Using the AVERAGE function in Google Sheets, we find that the mean is 1141.174.

2. Using the STDEV function, we get that the standard deviation is 125.517.
3. Entering "=NORM.DIST(1250, 1141, 125.517, TRUE)" into Google Sheets, we estimate that the University at Buffalo is at the 81st percentile.
4. Using PERCENTRANK, we find that the actual percentile is the 84th. These are close!

WHO KNEW?

Political Meddling Exposed

The normal distribution pops up in some unusual places. Recently, a team at Duke University has been using statistics to help identify partisan gerrymandering, where electoral districts have been carefully drawn in a way that benefits one political party over another. In their analysis, they found that hypothetical election results in randomly drawn districts are normally distributed. By using techniques similar to the ones we used above, they can quantify precisely how biased a particular electoral map is by finding the percentile rank of the actual election result on the normal distribution of the hypothetical results. You can find out more about their work at the "Quantifying Gerrymandering" site here.

Key Concepts

- We can use z-scores to compare data values from different datasets.

8.8 Scatter Plots, Correlation, and Regression Lines

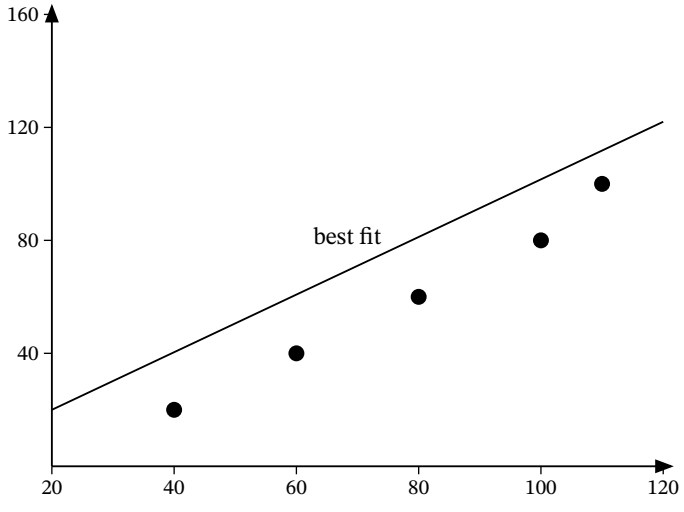


Figure 66: A scatter plot is a visualization of the relationship between quantitative dataset.

Learning Objectives

After completing this section, you should be able to:

1. Construct a scatter plot for a dataset.

2. Interpret a scatter plot.
3. Distinguish among positive, negative and no correlation.
4. Compute the correlation coefficient.
5. Estimate and interpret regression lines.

One of the most powerful tools statistics gives us is the ability to explore relationships between two datasets containing quantitative values, and then use that relationship to make predictions. For example, a student who wants to know how well they can expect to score on an upcoming final exam may consider reviewing the data on midterm and final exam scores for students who have previously taken the class. It seems reasonable to expect that there is a relationship between those two datasets: If a student did well on the midterm, they were probably more likely to do well on the final than the average student. Similarly, if a student did poorly on the midterm, they probably also did poorly on the final exam.

Of course, that relationship isn't set in stone; a student's performance on a midterm exam doesn't cement their performance on the final! A student might use a poor result on the midterm as motivation to study more for the final. A student with a really good grade on the midterm might be overconfident going into the final, and as a result doesn't prepare adequately.

The statistical method of *regression* can find a formula that does the best job of predicting a score on the final exam based on the student's score on the midterm, as well as give a measure of the confidence of that prediction! In this section, we'll discover how to use regression to make these predictions. First, though, we need to lay some graphical groundwork.

Relationships Between Quantitative Datasets

Before we can evaluate a relationship between two datasets, we must first decide if we feel that one might depend on the other. In our exam example, it is appropriate to say that the score on the final depends on the score on the midterm, rather than the other way around: if the midterm depended on the final, then we'd need to know the final score *first*, which doesn't make sense.

Here's another example: if we collected data on home purchases in a certain area, and noted both the sale price of the house and the annual household income of the purchaser, we might expect a relationship between those two. Which depends on the other? In this case, sale price depends on income: people who have a higher income can afford a more expensive house. If it were the other way around, people could buy a new, more expensive house and then expect a raise! (This is very bad advice.)

It's worth noting that not every pair of related datasets has clear dependence. For example, consider the percent of a country's budget devoted to the military and the percent earmarked for public health. These datasets are generally related: as one goes up, the other goes down. However, in this case, there's not a preferred choice for dependence, as each could be seen as depending on the other. When exploring the relationship between two datasets, if one set seems to depend on the other, we'll say that dataset contains values of the **response variable** (or **dependent variable**). The dataset that the response variable depends on contains values

of what we call the **explanatory variable** (or **independent variable**). If no dependence relationship can be identified, then we can assign either dataset to either role.

Example 1 **Identifying Explanatory and Response Variables**

For each of the following pairs of related datasets, identify which (if any) should be assigned the role of response variable and which should be assigned to be the explanatory variable.

1. A person's height and weight
2. A professional basketball player's salary and their average points scored per game (which is a measure of how good they are at basketball)
3. The length and width of leaves on a tree

Solution

1. As people get taller, their weight tends to increase. But if a person goes on a diet and loses weight, we don't expect them to also get shorter. So, weight depends on height. That means we'll say that the response variable is *weight* and the explanatory variable is *height*.
2. The more points a basketball player scores, the more money they should make. But if a basketball player gets a raise, we wouldn't expect them to get better at basketball as a result. So, the response is *salary* and the explanatory is *points per game*.
3. The way that the length and width of leaves are connected isn't clear. It seems reasonable that as the width goes up, so would the length. But the other direction is also plausible: as the length goes up, so does the width. Without a clear dependence relationship, we're free to declare either to be the response and the other to be the explanatory.

Once we've assigned roles to our two datasets, we can take the first step in visualizing the relationship between them: creating a scatter plot.

Creating Scatter Plots

A **scatter plot** is a visualization of the relationship between two quantitative sets of data. The scatter plot is created by turning the datasets into ordered pairs: the first coordinate contains data values from the explanatory dataset, and the second coordinate contains the *corresponding* data values from the response dataset. These ordered pairs are then plotted in the xy-plane. Let's return to our exam example to put this into practice.

Example 2 **Creating Scatter Plots Without Technology**

Students are exploring the relationship between scores on the midterm exam and final exam in their math course. Here are some of the scores reported by their classmates:

Name	Midterm grade	Final grade
Student 1	88	84
Student 2	71	80
Student 3	75	77
Student 4	94	95
Student 5	68	73

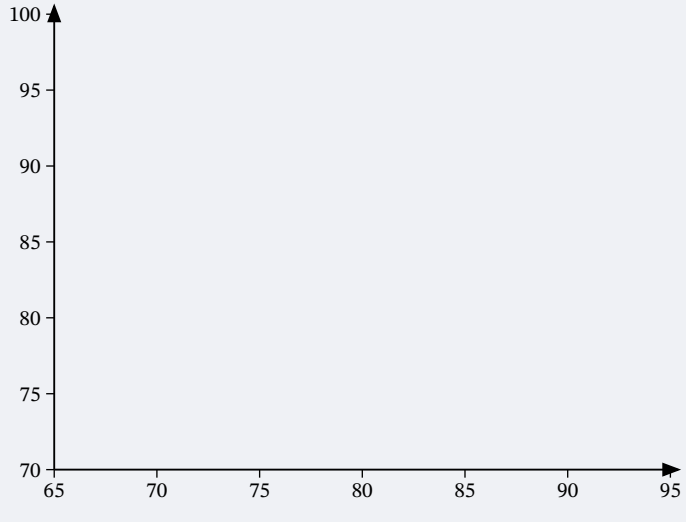
Create a scatter plot to visualize the data.

Solution

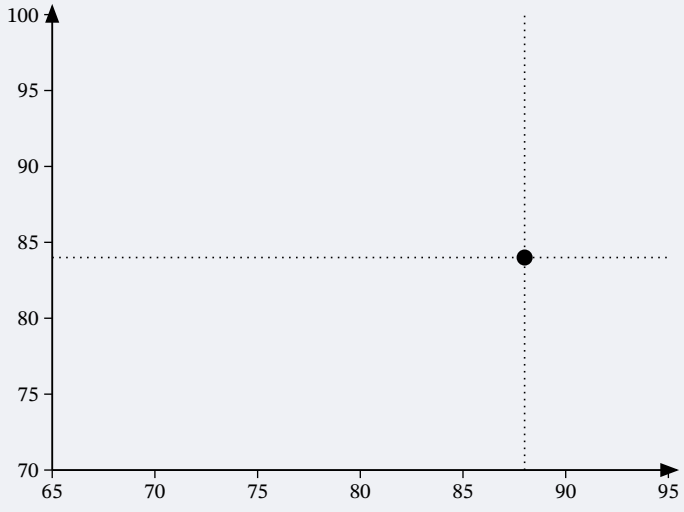
Step 1: Since it makes more sense to think of the final exam score as being dependent on the midterm exam score, we'll let the final grade be the response. So, let's think of these two datasets as a set of ordered pairs, midterm first, final second:

Name	(Midterm, Final)
Student 1	(88, 84)
Student 2	(71, 80)
Student 3	(75, 77)
Student 4	(94, 95)
Student 5	(68, 73)

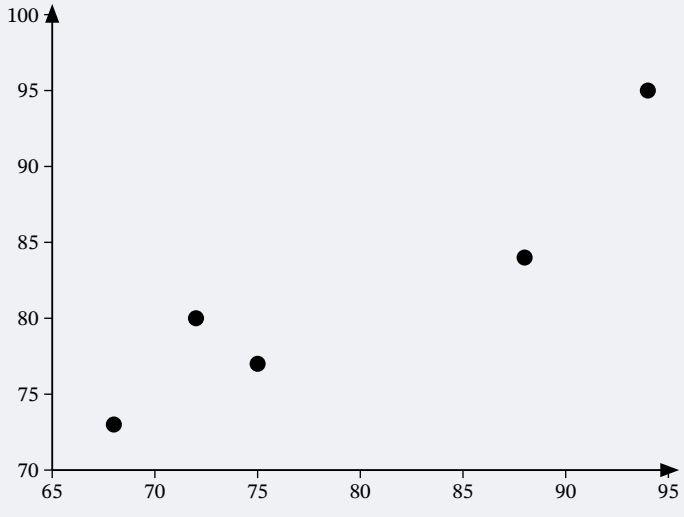
Step 2: Next, let's make the axes. On the horizontal axis, make sure the range of values is sufficient to cover all of the explanatory data. For our data, that's 68 to 94. Similarly, the vertical axis should cover all of the response data (73 to 95):



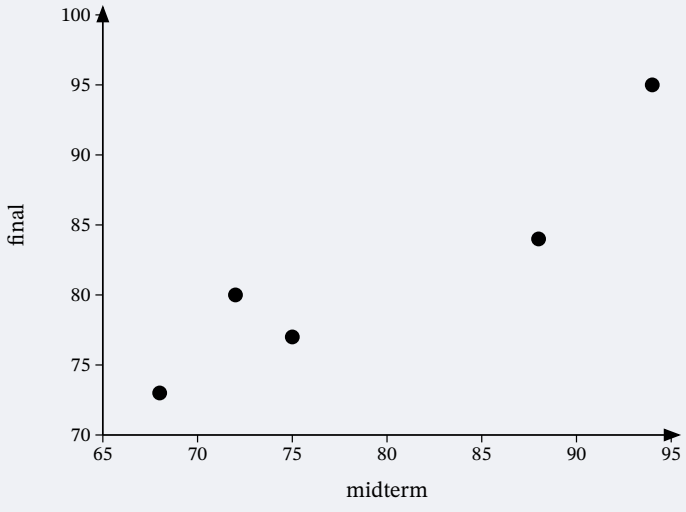
Step 3: Our first point is (88, 84). So, we'll locate 88 on the horizontal axis, 84 on the vertical axis, and identify the point that's directly above the first location and horizontally level with the second:



Step 4: Repeat this process to place the other four points on the graph:



Step 5: Finally, label the axes:



For large datasets, it's impractical to create scatter plots manually. Luckily, Google Sheets automates this process for us.

VIDEO
Making Scatter Plots in Google Sheets

Example 3 **Creating Scatter Plots in Google Sheets**

The dataset "NHL19" gives the results of the 2018–2019 National Hockey League season. The columns are team, wins (W), losses (L), overtime losses (OTL), total points (PTS), goals scored by the team (GF), goals scored against the team (GA), and goal differential (the difference in GF and GA). Use Google Sheets to create a scatter plot for GF vs. GA.

CHECKPOINT
When we talk about plotting one set versus another, the first is the response and the second is explanatory.

Solution

Step 1: Open the dataset in Google Sheets, and click and drag to select the data we want to visualize (in this case, we want the columns for GF and GA; make sure you include those labels in the selection).

Step 2: Next, click on the "Insert" menu, then click "Chart." Sheets will automatically choose a chart format; if the result isn't a scatter plot, click on the drop-down menu under "Chart type" in the Chart Editor on the right side of the window and select "Scatter chart."

Step 3: Next, check that the correct choices were made for the horizontal and vertical axes. In this case, we want to see GF on the vertical axis and GA on the horizontal axis. If Sheets

made the wrong choice, we can fix it in the Chart Editor by clicking on the name of the dataset under "X-axis" to open up a dropdown menu, then selecting the variable that should go on the horizontal axis (GA in this case).

Step 4: Now, click on the variable under "Series" and select the one that should go on the vertical axis (GF). If you had to make that change, the axis labels in the graph may also need changing; those labels can be fixed using the "Customize" tab in the Chart Editor under "Chart & axis titles." Your result should look like this:

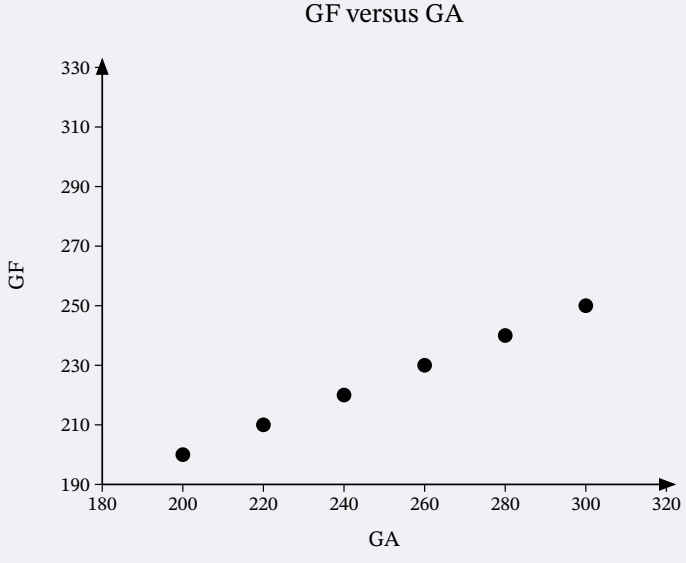


Figure 71: (data source: www.nhl.com)

Reading and Interpreting Scatter Plots

Scatter plots give us information about the existence and strength of a relationship between two datasets. To break that information down, there are a series of questions we might ask to help us. First: Is there a curved pattern in the data? If the answer is "yes," then we can stop; none of the linear regression techniques from here to the end of this section are appropriate. and show several examples of scatter plots that can help us identify these curved patterns.

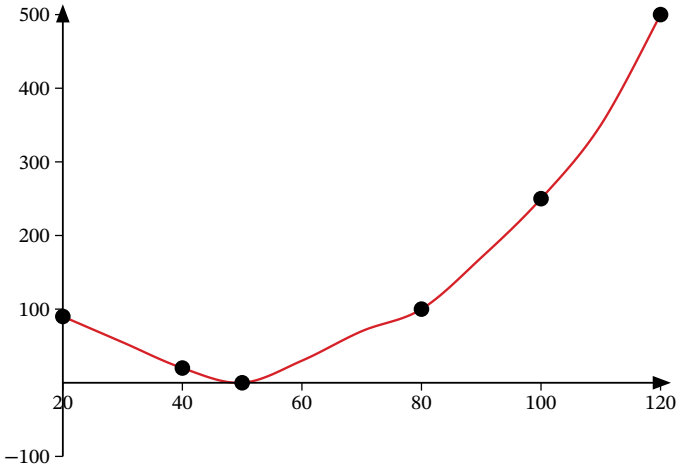


Figure 72: Curved pattern

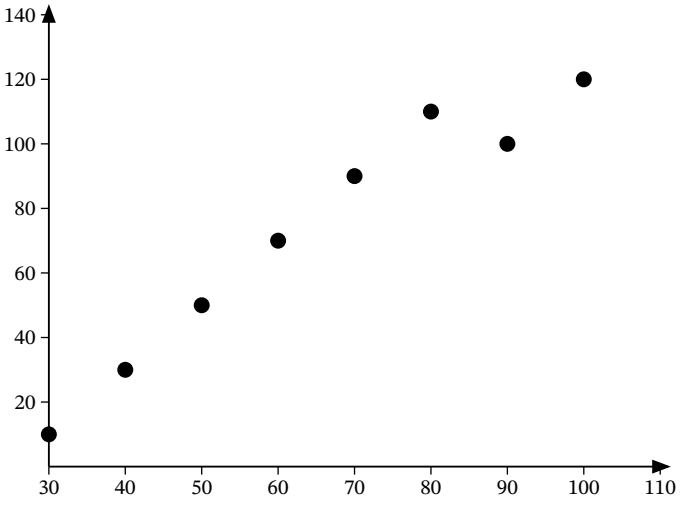


Figure 73: No curved pattern

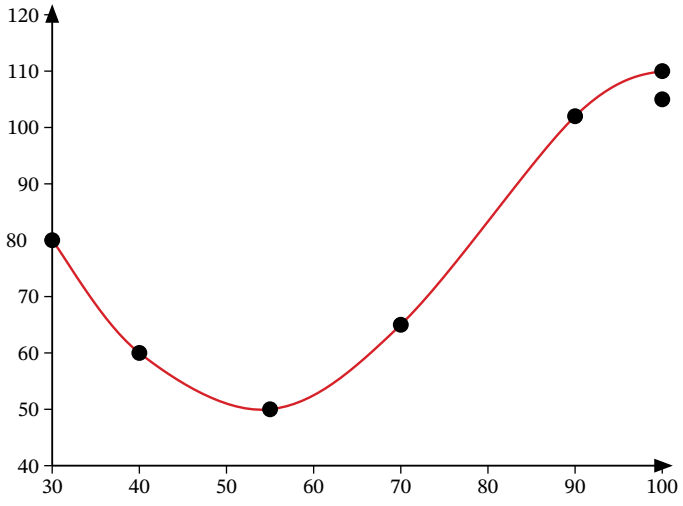


Figure 74: Curved pattern

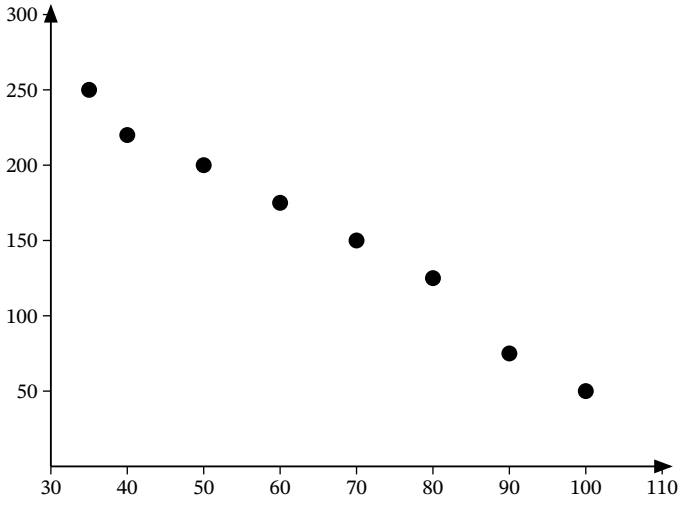
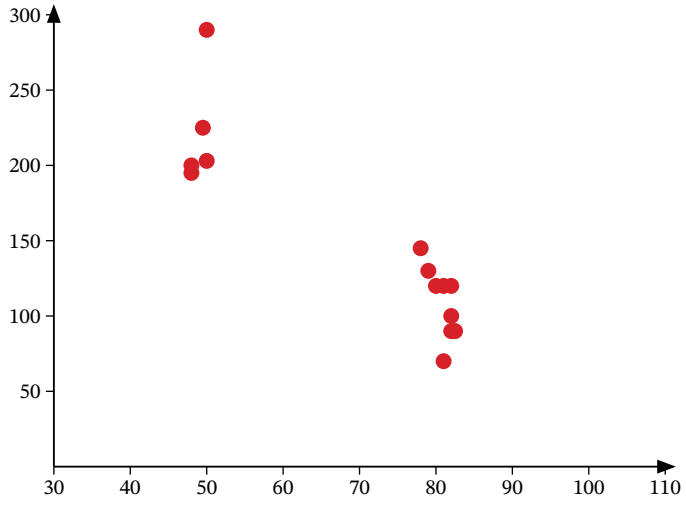
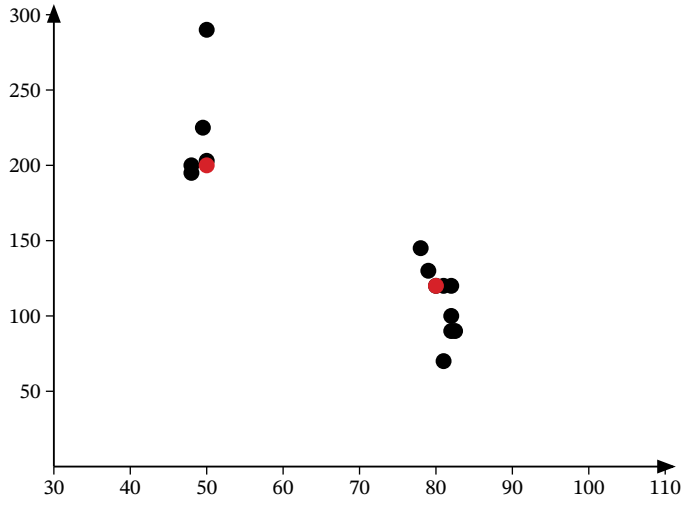


Figure 75: No curved pattern

Once we have confirmed that there is no curved pattern in our data, we can move to the next question: Is there a linear relationship? To answer this, we must look at different values of the explanatory variable and determine whether the corresponding response values are different, *on average*. It's important to look at the values "on average" because, in general, our scatter plots won't include just one corresponding response point for each value of the explanatory variable (i.e., there may be multiple response values for each explanatory value). So, we try to look for the center of those points. Let's look again at , but consider some different values for the explanatory variable. Let's highlight the points whose x-values are around 50 and those that are around 80:

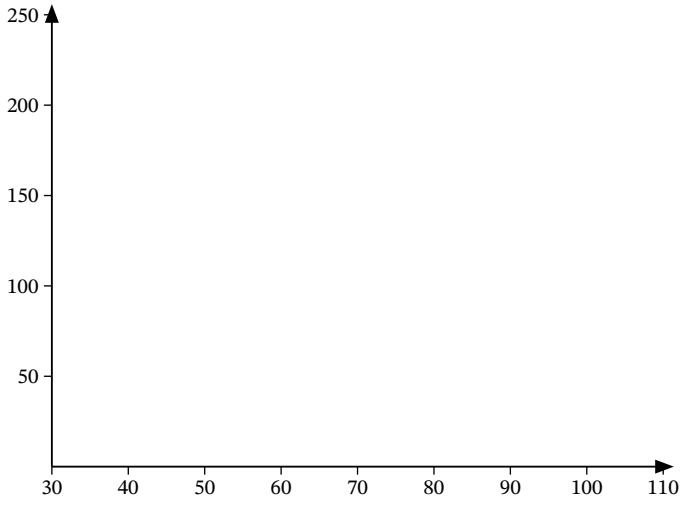


Now, we can estimate the middle of each group of points. Let's add our estimated averages to the plot as starred points:

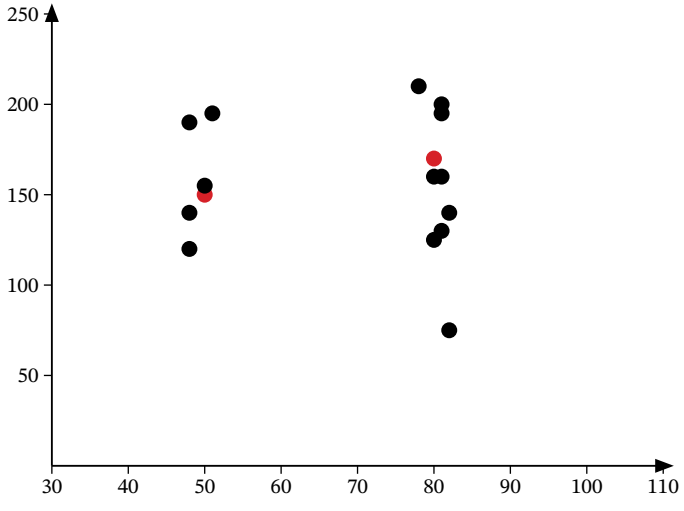


Since those two starred points occur at different heights, we can conclude that there's likely a relationship worth exploring.

Here's another example using a different set of data:



Let's look again at the points near 50 and near 80, and estimate the middles of those clusters:



Notice that there's not much vertical distance between our two starred points. This tells us that there's not a strong relationship between these two datasets.

Positive and Negative Linear Relationships

Another way to assess whether there is a relationship between two datasets in a scatter plot is to see if the points seem to be clustered around a line (specifically, a line that's not horizontal). The stronger the clustering around that line is, the stronger the relationship.

Once we've established that there's a relationship worth exploring, it's time to start quantifying that relationship. Two datasets have a **positive linear relationship** if the values of the response tend to increase, on average, as the values of the explanatory variable increase. If the values of the response *decrease* with increasing values of the explanatory variable, then there is a **negative linear relationship** between the two datasets. The strength of the relationship is determined by how closely the scatter plot follows a single straight line: the closer the points are to that line,

the stronger the relationship. The scatter plots in to depict varying strengths and directions of linear relationships.

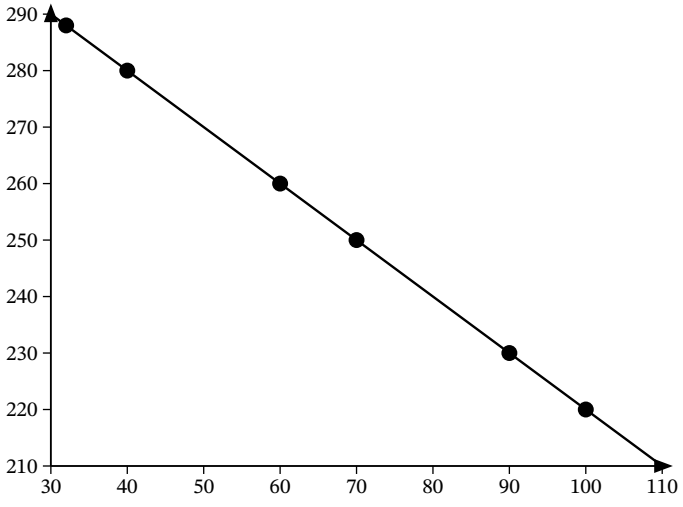


Figure 80: Perfect negative relationship

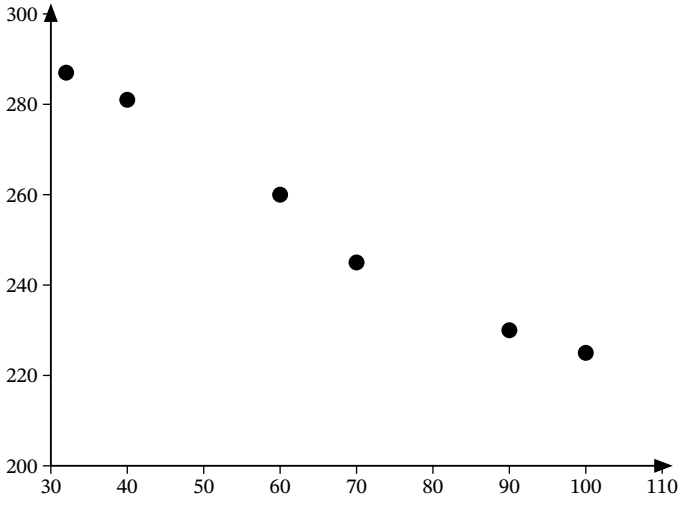


Figure 81: Strong negative relationship

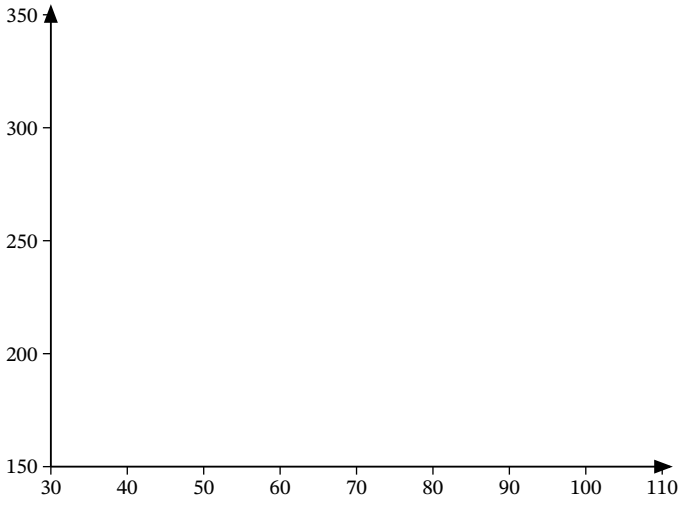


Figure 82: Weak negative relationship

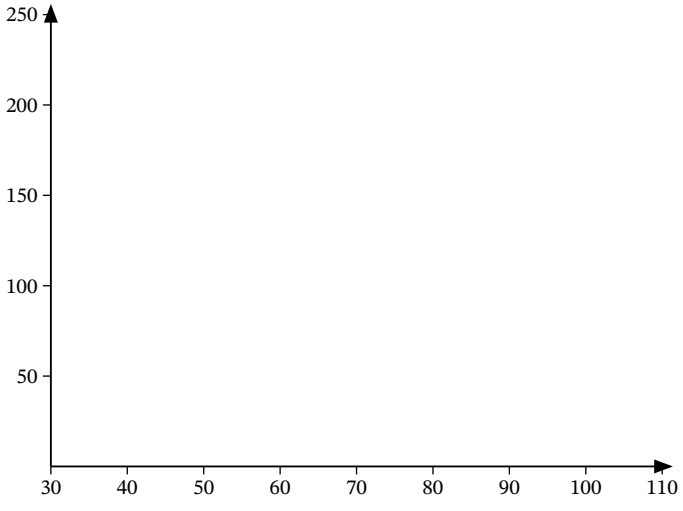


Figure 83: No relationship

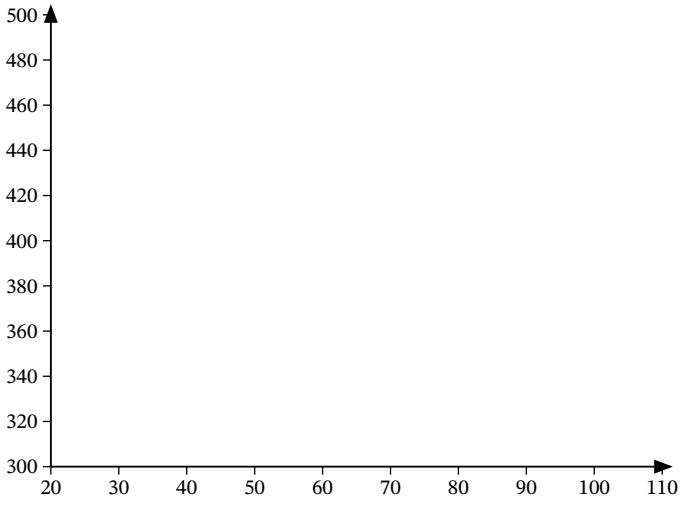


Figure 84: Weak positive relationship

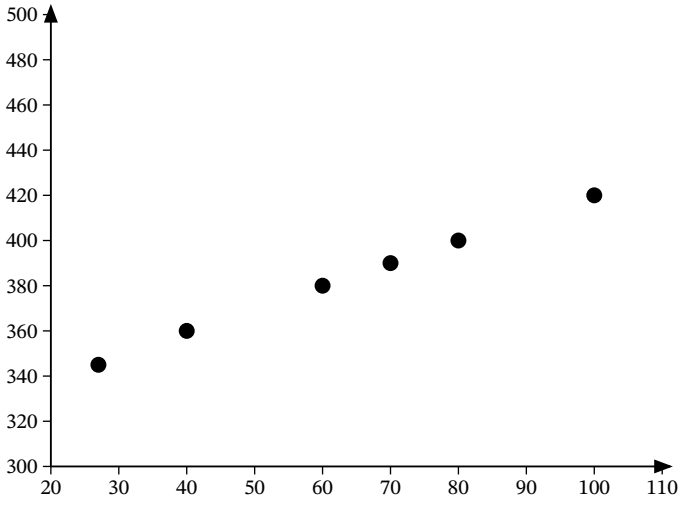


Figure 85: Strong positive relationship

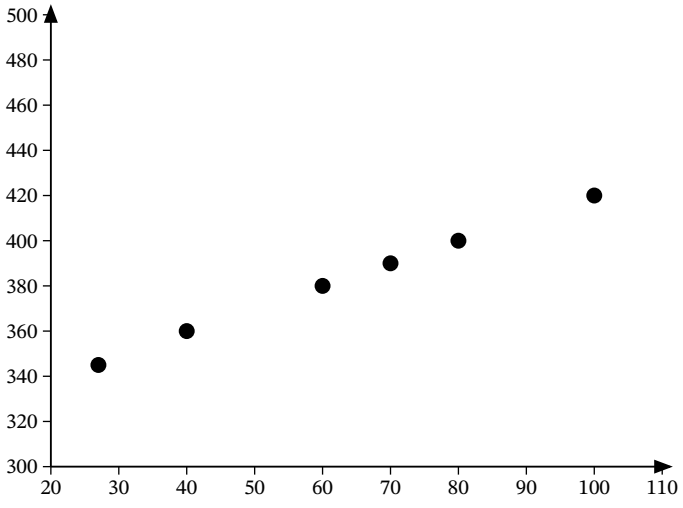


Figure 86: Perfect positive relationship

The strength and direction (positive or negative) of a linear relationship can also be measured with a statistic called the **correlation coefficient** (denoted r). Positive values of r indicate a positive relationship, while negative values of r indicate a negative relationship. Values of r close to 0 indicate a weak relationship, while values close to ± 1 correspond to a very strong relationship. Looking again at to , the correlation coefficients for each, in sequential order, are: –1, –0.97, –0.55, –0.03, 0.61, 0.97, and 1.

There's no firm rule that establishes a cutoff value of r to divide strong relationships from weak ones, but ± 0.7 is often given as the dividing line (i.e., if $r > 0.7$ or $r < -0.7$ the relationship is strong, and if $-0.7 < r < 0.7$ the relationship is weak).

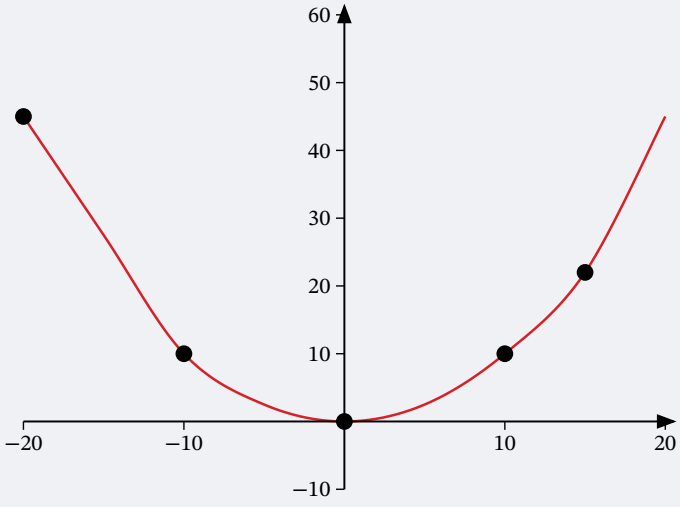
The formula for computing r is very complicated; it's almost never done without technology. Google Sheets will do the computation for you using the CORREL function. The syntax works like this: if your explanatory values are in cells A2 to A50 and the corresponding response values are in B2 to B50, then you can find the correlation coefficient by entering "=CORREL(A2:A50, B2:B50)". (Note that the order doesn't matter for correlation coefficients; "=CORREL(B2:B50, A2:A50)" will give the same result.)

Let's put all of this together in an example.

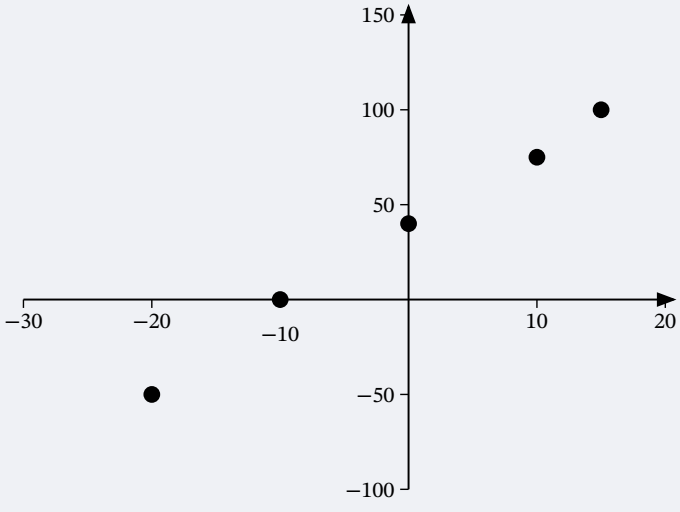
Example 4 **Interpreting Scatter Plots**

Consider the four scatter plots below:

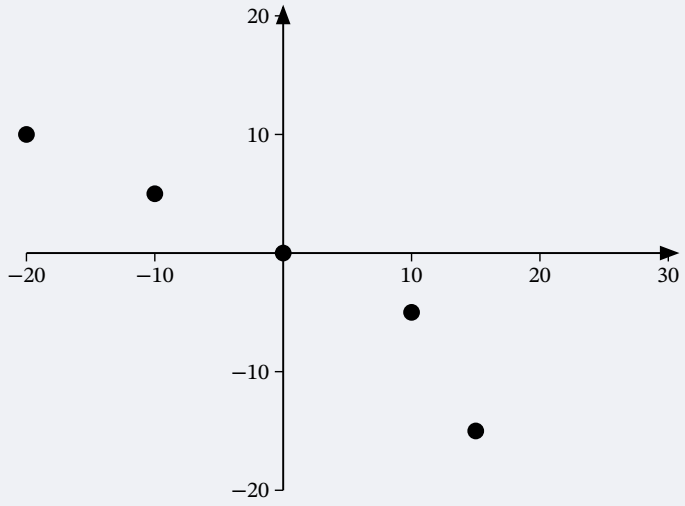
1.



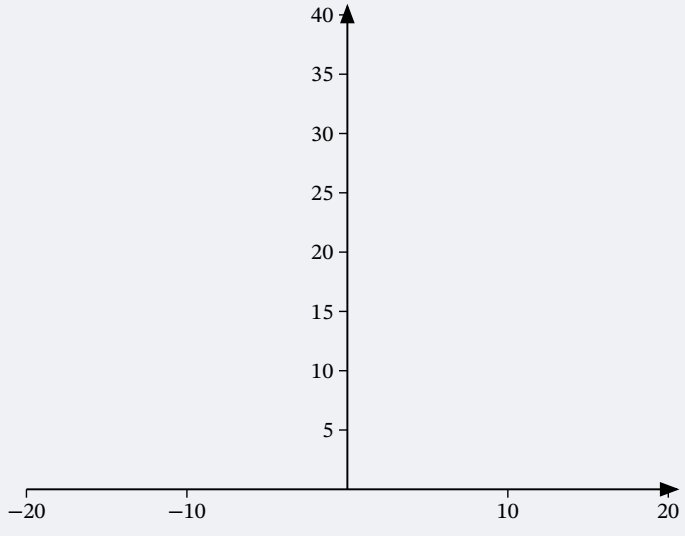
2.



3.



4.



For each of these, answer the following questions:

1. Is there a curved pattern in the data? If yes, stop here. If no, continue to part b.
2. Classify the strength and direction of the relationship. Make a guess at the value of r.

Solution

1. 1. Yes, there is a curved pattern.
2. 1. No, there's no curved pattern.
3. Since the points tend upward as we move from left to right, this is a positive relationship. The points seem pretty closely grouped around a line, so it's fairly strong. Comparing this scatter plot to those in to , we can see that the relationship is stronger

than the one in ($r = 0.61$) but not as strong as the one in ($r = 0.97$). So, the value of the correlation coefficient is somewhere between the two. We might guess that $r = 0.9$.

4. 1. No, there's no curved pattern.
5. Since the points tend downward as we move from left to right, this is a negative relationship. The points are not tightly grouped around a line, but the pattern is clear. It looks like it has approximately the same strength as the plot in , just with the opposite sign. So, we might guess that $r = -0.6$.
6. 1. No, there's no curved pattern.
7. Since the points don't really tend upward or downward as we move from left to right, there is no real relationship here. Thus, $r \approx 0$.

Example 5 **Finding the Correlation Coefficient**

The data that were plotted in the previous example can be found in the dataset "correlationcoefficient1". All of them share the same values for the explanatory variable x. The four responses are labeled $y1$ through $y4$. Compute the correlation coefficients for each, if appropriate, using Google Sheets. Round to the nearest hundredth.

Solution

Step 1: There is a curved pattern in the data, so the correlation coefficient isn't meaningful.

Step 2: Using "=CORREL(A2:A101, C2:C101)" we get $r = 0.89$.

Step 3: Using "=CORREL(A2:A101, D2:D101)" we get $r = -0.66$.

Step 4: Using "=CORREL(A2:A101, E2:E101)" we get $r = -0.04$.

WORK IT OUT

Winning with Statistics

Billy Beane, the former general manager of the Oakland A's baseball team, famously took his low budget team to unprecedented heights by using statistics to identify undervalued players; his story is recounted in the book *Moneyball* (which was later made into a movie, with Brad Pitt playing Beane). You can do the same thing: Take a look at team statistics in the sport of your choice and try to identify a statistic that's most closely related to winning (meaning that it has the highest correlation coefficient with team wins).

Linear Regression

The final step in our analysis of the relationship between two datasets is to find and use the equation of the regression line. For a given set of explanatory and response data, the **regression line** (also called the **least-squares line** or **line of best fit**) is the line that does the best job of approximating the data.

What does it mean to say that a particular line does the "best job" of approximating the data? The way that statisticians characterize this "best line" is rather technical, but we'll include it

for the sake of satisfying your curiosity (and backing up the claim of "best"). Imagine drawing a line that looks like it does a pretty good job of approximating the data. Most of the points in the scatter plot will probably not fall exactly on the line; the distance above or below the line a given point falls is called that point's *residual*. We could compute the residuals for every point in the scatter plot. If you take all those residuals and square them, then add the results together, you get a statistic called the *sum of squared errors* for the line (the name tells you what it is: "sum" because we're adding, "squared" because we're squaring, and "errors" is another word for "residuals"). The line that we choose to be the "best" is the one that has the smallest possible sum of squared errors. The implied minimization ("smallest") is where the "least" in "least squares" comes from; the "squares" comes from the fact that we're minimizing the sum of *squared* errors. This is very similar to the process we outlined in the "game" that we used to introduce the mean. Both the regression line and the mean are designed to minimize a sum of squared errors. Here ends the super technical part.

Finding the Equation of the Regression Line

So, how do we find the equation of the regression line? Recall the point-slope form of the equation of a line:

FORMULA

If a line has slope m and passes through a point (x_0, y_0), then the point-slope form of the equation of the line is:

$$y = m(x - x_0) + y_0$$

The regression line has two properties that we can use to find its equation. First, it always passes through the *point of means*. If $\overline{x}$ and $\overline{y}$ are the means of the explanatory and response datasets, respectively, then the point of means is $(\overline{x}, \overline{y})$. We'll use that as the point in the point-slope form of the equation. Second, if s_x and s_y are the standard deviations of the explanatory and response datasets, respectively, and if r is the correlation coefficient, then the slope is $m = r \times \frac{s_y}{s_x}$. Putting all that together with the point-slope formula gives us this:

FORMULA

Suppose x and y are explanatory and response datasets that have a linear relationship. If their means are $\overline{x}$ and $\overline{y}$ respectively, their standard deviations are s_x and s_y respectively, and their correlation coefficient is r, then the equation of the regression line is:

$$y = r\left(\frac{s_y}{s_x}\right)(x - \overline{x}) + \overline{y}$$

.

Let's walk through an example.

Example 6 **Finding the Equation of the Regression Line from Statistics**

Suppose you have datasets x and y with the following statistics: x has mean 21 and standard deviation 4, y has mean 8 and standard deviation 2, and their correlation coefficient is -0.4. What's the equation of the regression line?

Solution

Step 1: We're given $\overline{x} = 21$, $s_x = 4$, $\overline{y} = 8$, $s_y = 2$, and $r = -0.4$. Let's start with the formula for the equation of the regression line:

$$y = r\left(\frac{s_y}{s_x}\right)(x - \overline{x}) + \overline{y}$$

Step 2: Plugging in our values gives us:

$$y = -0.4\left(\frac{2}{4}\right)(x - 21) + 8$$

Step 3: Our final regression line equation is:

$$y = -0.2x + 12.2$$

As you can see, finding the equation of the regression line involves a lot of steps if you have to find all of the values of the needed quantities yourself. But, as usual, technology comes to our rescue. This video (which you actually watched earlier when learning how to create scatter plots) covers the regression line at around the 3:30 mark. Note that Google Sheets calls it the "trendline."

Let's put this into practice.

Example 7 **Finding the Equation of the Regression Line Using Google Sheets**

In Example 3, we considered the relationship between goals scored (GF) and goals against (GA) using the dataset "NHL19". Recreate the scatter plot in Google Sheets, and use it to find the equation of the regression line.

Solution

Once we have recreated the scatter plot, we find the equation of the regression line by clicking the three dots at the top right of the plot, selecting "Edit chart," then clicking on "Customize" and "Series." We add the regression line by checking the box next to "Trendline," and then we show the equation by selecting "Use Equation" in the drop-down menu under "Label." The equation of the tangent line is $y = -0.0554x + 261$.

Using the Equation of the Regression Line

Once we've found the equation of the regression line, what do we do with it? We'll look at two possible applications: making predictions and interpreting the slope.

We can use the equation of the regression line to predict the response value y for a given explanatory value x. All we have to do is plug that explanatory value into the formula and see what response value results. This is useful in two ways: first, it can be used to make a guess about an unknown data value (like one that hasn't been observed yet). Second, it can be used

to evaluate performance (meaning, we can predict an outcome given a particular event). In Example 2, we created a scatter plot of final exam scores vs. midterm exam scores using this data:

Name	Midterm Grade	Final Grade
Allison	88	84
Benjamin	71	80
Carly	75	77
Daniel	94	95
Elmo	68	73

The equation of the regression line is $y = 0.687x + 27.4$, where y is the final exam score and x is the midterm exam score. If Frank scored 85 on the midterm, then our prediction for his final exam score is $0.687 \times 85 + 27.4 = 85.795$. To use the regression line to evaluate performance, we use a data value we've already observed. For example, Allison scored 88 on the midterm. The regression line predicts that someone who scores an 88 on the midterm will get $0.687 \times 88 + 27.4 = 87.856$ on the final. Allison actually scored 84 on the final, meaning she underperformed expectations by almost 4 points ($87.856 - 84$).

The second application of the equation of the regression line is interpreting the slope of the line to describe the relationship between the explanatory and response datasets. For the exam data in the previous paragraph, the slope of the regression line is 0.687. Recall that the slope of a line can be computed by finding two points on the line and dividing the difference in the y-values of those points by the difference in the x-values. Keeping that in mind, we can interpret our slope as $0.687 = \frac{\text{difference in final scores}}{\text{difference in midterm scores}}$. Multiplying both sides of that equation by the denominator of the fraction, we get $0.687 \times$ difference in midterm scores = difference in final scores. Thus, a one-point increase in the midterm score would result in a predicted increase in the final score of 0.687 points. A ten-point drop in the midterm score would give us a decrease in the predicted final score of 6.87 points. In general, the slope gives us the predicted change in the response that corresponds to a one unit increase in the explanatory variable.

Example 8 **Applying the Equation of the Regression Line**

The data in "MLB2019Off" gives offensive team stats for the 2019 Major League Baseball season. Use that dataset to answer the following questions:

1. What is the equation of the regression line for runs (R) vs. hits (H)?
2. How many runs would we expect a team to score if the team got 1500 hits in a season?
3. Did the Kansas City Royals (KCR) overperform or underperform in terms of runs scored, based on their hit total? By how much?
4. Write a sentence to interpret the slope of the regression line.

Solution

1. Using Google Sheets, we find that the regression line equation is $y = 0.884x - 456$, where y is the number of runs scored and x is the number of hits.
2. Plugging 1500 into the equation of the regression line, we get $0.884 \times 1500 - 456 = 870$. We would predict that a team with 1500 hits would score 870 runs.
3. The Royals had 1356 hits, so we would predict their run total to be $0.884 \times 1356 - 456 \approx 743$. They actually scored 691 runs, so they underperformed expectations by 52 runs $(743 - 691)$.
4. The slope gives us the predicted change in the response that corresponds to a one unit increase in the explanatory variable. So, we expect one additional hit to result in 0.884 more runs. Since 0.884 runs doesn't really make sense, we can get a better interpretation by multiplying through by ten or one hundred: Ten additional hits will result in almost nine additional runs, or a hundred additional hits will yield on average just over 88 additional runs.

WHO KNEW?

Math and the Movies

Statistics and regression are used by Hollywood movie producers to decide what movies to make, and to predict how much money they'll earn at the box office. According to the American Statistical Association, not only do producers use statistics to identify the next potential blockbuster, but they've also pinned down how much money awards add to the bottom line. (An Academy Award is worth about $3 million!) In addition, studios use their streaming services to gather data about their customers and the types of movies they watch; this data helps them learn what kinds of entertainment their customers want more of.

WORK IT OUT

Collecting and Analyzing Your Own Data

This section has demonstrated many pairs of related quantitative datasets. Think about some quantitative variables that you can ask your classmates about, which might be related. Once you have some ideas, collect the data from your classmates. Then analyze the data by creating a scatter plot, finding the equation of the regression line (if appropriate), and interpreting it.

Extrapolation

A very common misuse of regression techniques involves **extrapolation**, which involves making a prediction about something that doesn't belong in the dataset.

Example 9 **More Applying the Equation of the Regression Line**

The data in "WNBA2019" gives team statistics from the 2019 WNBA season. Use that dataset to answer these questions about team wins (W) and the proportion of team field goals made (FG%, the number of shots made divided by the number of shots attempted. Even though this column is labeled using a percent sign, the values are not expressed as percentages):

1. What is the equation of the regression line for wins vs. proportion of made field goals?
2. How many wins would we expect for a team that makes 42% of its shots?
3. Did the New York Liberty overperform or underperform in terms of wins, based on the team's proportion of made field goals?
4. Write a sentence to interpret the slope of the regression line.

Solution

1. Using Google Sheets, we get the equation $y = 178.097x - 58.543$, where x is the proportion of field goals made and y is the number of wins.
2. Since 42% corresponds to a proportion of 0.42, we'll plug 0.42 into the regression equation for x, which gives us $178.097 \times 0.42 - 58.543 \approx 16.26$. We would predict that a team that makes 42% of its shots would win about 16 games.
3. The New York Liberty made 41.4% of their shots, so we expect they would have $178.097 \times 0.414 - 58.543 \approx 15.19$wins. In fact, they had only 10 wins, so they underperformed expectations by over 5 wins.
4. **Step 1:** The slope gives us the expected increase in the response that corresponds to a one unit increase in the explanatory variable. If we simply go with that interpretation, we would get a sentence like "We expect an increase in proportion of field goals made of 1 would result in an additional 178 wins." However, that sentence doesn't make much sense. Let's consider why.
   Step 2: First, proportions must be between 0 and 1, the proportion of made field goals can't be increased by 1 and still make sense. Second, the total number of games played is only 34, so no team could get an additional 178 wins! So, we'll have to change the units.
   Step 3: Since the proportions of made field goals are often expressed as a percentage, we could try to use that. If we express the slope as a fraction with 1 in the denominator (remember, the denominator represents the proportion of field goals made), then convert the denominator to a percentage and simplify, we get $\frac{178.097}{1} = \frac{178.097}{100\%} \approx \frac{1.78}{1\%}$.
   Step 4: So, an increase in field goal percentage of 1% would result in an expected increase of 1.78 wins.

Correlation Does Not Imply Causation

One of the most common fallacies about statistics has to do with the relationship between two datasets. In the dataset "Public", we find that the correlation coefficient between the 75th percentile math SAT score and the 75th percentile verbal SAT score is 0.92, which is really strong. The slope of the regression line that predicts the verbal score from the math score is 0.729, which we might interpret as follows: "If the 75th percentile math SAT score goes up by 10 points, we'd expect the corresponding verbal SAT score to go up by just over 7 points."

Does the increasing math score *cause* the increase in the verbal score? Probably not. What's really going on is that there's a third variable that's affecting them both: To raise the SAT math score by 10 points, a school will recruit students who do better on the SAT in general; these students will also naturally have higher SAT verbal scores. This third variable is sometimes called a *lurking variable* or a *confounding variable*. Unless all possible lurking variables are ruled out, we cannot conclude that one thing causes another.

PEOPLE IN MATHEMATICS

Dr. Talithia Williams

Figure 91: A photo of Dr. Williams

Dr. Talithia Williams is a statistician on the faculty of Harvey Mudd College, and the first Black woman to achieve tenure at this university. She advocates for more women to become involved in the fields of engineering and science, and is on the board of directors for the EDGE Foundation, an organization that helps women obtain advanced degrees in

mathematics (EDGE standing for Enhancing Diversity in Graduate Education). In 2018, Dr. Williams published the book Power in Numbers: The Rebel Women of Mathematics, a retrospective look at historical female figures who have contributed to the development of the field of mathematics.

Dr. Williams earned a Master's degree in Mathematics from Howard University and a Master's in Statistics from Rice University, and also went on to earn her Ph.D. in Statistics from Rice. She has held research appointments at the Jet Propulsion Laboratory, the National Security Agency, and NASA. Her research focuses on the environmental and medical applications of statistics. In 2014, she gave a popular TED talk titled "Own Your Body's Data" that discussed the potential insights to be gained from collecting personal health data. She was even recently a host for the NOVA Wonders documentary series and a narrator for the NOVA Universe series on PBS.

To stay up-to-date on Dr. Williams's accomplishments, you can follow her on Twitter or her Facebook account.

WHO KNEW?

Statistics and Eugenics

Some of the brightest minds in the history of statistics unfortunately decided to use their considerable intellects to further a pseudoscience known as eugenics. Eugenicists took Charles Darwin's theories of evolution and ruthlessly applied them to the human race. Francis Galton (1822–1911), a cousin of Darwin and also the mathematician who invented the formula for standard deviation, claimed that people in the British upper classes possessed higher intelligence due to their superior breeding. Karl Pearson (1857–1936), who derived the formula for the correlation coefficient, argued in *National Life from the Standpoint of Science*, that, instead of providing social welfare programs, nations could better improve the fortunes of the poor by waging "war with inferior races." Ronald Fisher (1890–1962) was possibly the most important statistician of the 20th century, having invented several new techniques (including the ubiquitous analysis of variance), and yet he also founded the Cambridge University Undergraduates Eugenics Society, whose self-prescribed goal was to evangelize "not by precept only, but by example, the doctrine of a new natural ability of worth and blood."

When eugenics took hold in the United States, it was used to justify terrible acts by the government, including the forced sterilization of individuals with mental illness, epilepsy, a physical impairment (like blindness), or a criminal history. The Nazi regime took these ideas to their ultimate, terrible conclusion: killing people who had mental or physical disabilities, or who were born into an "inferior" race. Over six million people died in this Holocaust, one of the darkest events in human history. To learn more, watch this video about Francis Galton and the legacy of eugenics

Key Terms

- response variable (dependent variable)

- explanatory variable (independent variable)
- scatter plot
- positive linear relationship
- negative linear relationship
- correlation coefficient
- regression line (least-squares line, line of best fit)

Key Concepts

- If one variable affects the value of another variable, we say the first is an explanatory variable and the second is a response variable.
- Scatter plots place a point in the xy-plane for each unit in the dataset. The x-value is the value of the explanatory variable, and the y-value is the value of the response variable.
- The correlation coefficient r gives us information about the strength and direction of the relationship between two variables. If r is positive, the relationship is positive: an increase in the value of the explanatory variable tends to correspond to an **increase** in the value of the response variable. If r is negative, the relationship is negative: an increase in the value of the explanatory variable tends to correspond to a **decrease** in the value of the response variable. Values of r that are close to 0 indicate weak relationships, while values close to –1 or indicate strong relationships.
- The regression line for a relationship between two variables is the line that best represents the data. It can be used to predict values of the response variable for a given value of the explanatory variable.

Formulas

If a line has slope m and passes through a point (x_0, y_0), then the point-slope form of the equation of the line is:

$$y = m(x - x_0) + y_0$$

Suppose x and y are explanatory and response datasets that have a linear relationship. If their means are $\overline{x}$ and $\overline{y}$ respectively, their standard deviations are s_x and s_y respectively, and their correlation coefficient is r, then the equation of the regression line is:

$$y = r\left(\frac{s_y}{s_x}\right)(x - \overline{x}) + \overline{y}.$$

Videos

- Making Scatter Plots in Google Sheets

Projects

1. Browse through some news websites to find five stories that report on data and include data visualizations. Can you tell from the report how the data were collected? Was randomization

used? Are the visualizations appropriate for the data? Are the visualizations presented in a way that might bias the reader?

2. We discussed three measures of centrality in this chapter: the mode, the median, and the mean. In a broader context, the mean as we discussed it is more properly called the *arithmetic mean*, to distinguish it from other types of means. Examples of these include the geometric mean, harmonic mean, truncated mean, and weighted mean. How are these computed? How do they compare to the arithmetic mean? In what situations would each of these be preferred to the arithmetic mean?
3. *Simpson's Paradox* is a statistical phenomenon that can sometimes appear when we observe a relationship within several subgroups of a population, but when the data for all thegroups are analyzed all together, the opposite relationship appears. Find some examples of Simpson's Paradox in real-world situations, and write a paragraph or two that would explain the concept to someone who had never studied statistics before.