



Versatile Mathematics

COMMON MATHEMATICAL APPLICATIONS



Josiah Hartley

Frederick Community College

Val Lochman

Frederick Community College

Erum Marfani

Frederick Community College

2nd Edition

2020

This text is licensed under a Creative Commons Attribution-Share Alike 3.0 United States License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA

You are **free**:

- to Share** – to copy, distribute, display, and perform the work
- to Remix** – to make derivative works

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same, similar, or a compatible license.

With the understanding of the following:

Waiver. Any of the above conditions can be waived if you get permission from the copyright holder.

Other Rights. In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights
- Apart from the remix rights granted under this license, the authors' moral rights
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights
- Notice — For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to the following web page: <http://creativecommons.org/licenses/by-sa/3.0/us/>

Attributions This book benefited tremendously from others who went before and freely shared their creative work. The following is a short list of those whom we have to thank for their work and their generosity in contributing to the free and open sharing of knowledge.

- David Lippman, author of *Math in Society*. This book uses sections derived from his chapters on Finance, Growth Models, and Statistics. He also administers MyOpenMath, the free online homework portal to which the problems in this text were added.
- The developers of onlinestatbook.com.
- OpenStax College (their book *Introductory Statistics* was used as a reference)
OpenStax College, *Introductory Statistics*. OpenStax College. 19 September 2013. <<http://cnx.org/content/col11562/latest/>>
- The authors of OpenIntro Statistics, which was also used as a reference.
- The Saylor Foundation Statistics Textbook: <http://www.saylor.org/site/textbooks/Introductory%20Statistics.pdf>

Thanks The following is a short list of those whom we wish to thank for their help and support with this project.

- The President's office at Frederick Community College, for providing a grant to write the first chapters.
- Gary Hull, who in his tenure as department chair gave us his full support and gave us the impetus to start the project, and generously shared his notes for MA 103.
- The entire FCC math department, who provided untold support and encouragement, as well as aid in reviewing and editing the text.

Example Videos

Every example in the book has an accompanying video; to view the video, click on the title of the example or the example number in the margin.

EXAMPLE 1



CONVERTING TO PERCENTAGES

Convert each of the following to a percentage:

(a) $\frac{2}{5}$ (b) 0.15 (c) $\frac{9}{2}$

Solution

$$(a) \frac{2}{5} = 0.40 = 40\% \quad (b) 0.15 = 15\% \quad (c) \frac{9}{2} = 4.50 = 450\%$$

Try It

Many examples are followed by Try It examples, which can be used for extra practice. Clicking on the words Try It in the margin will open a web page where students can enter their answers and check them.

TRY IT



Convert each of the following to a percentage:

(a) $\frac{3}{5}$ (b) 0.7 (c) 2

Free Online Homework

Versatile Math includes free online homework, provided through MyOpenMath.

● Question 23

0/1 pt 100 99 Details

In the fall of 2009 FCC enrolled 6,233 students. Of those enrolled, 2,810 are in the 18-21 age group. What percent of FCC students does this represent? Give your answer to at least one decimal place.

 %

Question Help: [Video](#) [Message instructor](#)

Submit Question

What's New in the 2nd Edition

Chapter 1

- A new introduction section was added, describing many of the initial concepts used in the chapter.
- The income tax section was moved to the end of the chapter and updated to reflect changes to the U.S. tax code (including 2020 tax brackets); the discussion of deductions and credits was expanded with new graphics.
- The discussion of inflation was removed, since it made the section on simple and compound interest too long.
- New examples were added to the retirement section, showing how to plan fully for retirement.
- A description of the use of Excel and the TVM solver on TI graphing calculators was added to several sections.

Chapter 2

- A new section on quadratic models was added.
- A discussion on using a calculator to do regression with each type of model was added.
- In the section on exponential models, the discussion was condensed by eliminating models of the form $P_t = P_0 e^{kt}$ and Newton's law of cooling.

Chapter 3

- The entire chapter was reordered:
 - The old first section was split, expanding the discussion of gathering data and graphing it into two separate sections.
 - The discussion of sampling methods was expanded, and new graphs were introduced, including dot plots and scatterplots.
 - The next two sections (measures of center and measures of spread) were merged into a single section on describing data with statistics.
 - A new section was added on the use of linear regression.

Chapters 4–7 These remained mostly unchanged.

Chapter 8 A new chapter on Graph Theory was added, with all-new videos and homework.

Chapter R A short review of some algebra concepts used throughout the book was added.



Contents

3	Statistics	127
3.1	Gathering Data	128
3.2	Visualizing Data	138
3.3	Describing Data with Statistics	159
3.4	Linear Regression	175
3.5	The Normal Distribution	190

Statistics



More than ever before, we are surrounded by tremendous amounts of data. Machine learning, an advanced form of statistics, is one of the fastest-growing fields in the world, with applications in every industry imaginable. Even something as simple as unlocking a phone often involves a fingerprint or face scan, and self-driving cars have gone from science fiction to reality.

Statistics are everywhere, from crime rates in your city to weight percentiles for children on growth charts. When a research team is testing a new treatment for a disease, they can use statistics to make conclusions based on a relatively small trial and show that there is good evidence that their drug is effective. Statistics allowed prosecutors in the 1950's and 60's to demonstrate that racial bias existed in jury panels.

How do we make sense of all of this information that surrounds us? There are many tools that have been developed to do exactly that, and in this chapter, you will learn about a few basic ones.

Broadly speaking, statistics is the study of how to gather and make sense of data. We will begin by learning to gather data carefully, which generally involves picking a good sample. Then we will see how to describe data both visually and numerically, and finally, we will discuss a few ways to put data to work answering questions.

SECTION 3.1 Gathering Data



Every 10 years, the U.S. Census Bureau undertakes the enormous task of gathering all kinds of data on people residing in the country. This is incredibly difficult, but equally important, because everything from representation in Congress to federal funding depends on having accurate counts.

It's not hard to see why this process is so complicated. Not only are there hundreds of millions of people in the country, but many live in rural areas, others are transient or homeless, and many may be suspicious of someone showing up at their door with a clipboard, asking questions. Because of this, and because it is crucial to count everyone, not just those who are easy to count, the Census Bureau devotes a tremendous amount of time, energy, and money to designing their procedure as carefully as possible.

Sampling

Now think of a different example: a presidential election. The election occurs on a specific day, and you can think of the election as a massive survey with the question, "Who do you want to be President?"

However, before the election occurs, we generally like to have an idea of how the vote will turn out. How can we predict it? Can we go around and ask every voter how they plan to vote, and then repeat this every so often (because people may change their minds)? Clearly, that's impractical, so we turn to the statistician's secret weapon: **sampling**.

To understand why sampling works, think about cooking a pot of soup. Throughout the process, a good cook will frequently take a small spoonful from the simmering pot and taste it. The cook doesn't need to eat the entire potful to get a sense of how it's doing; even though there are many ingredients in the pot, if it is well-stirred, the small spoonful will taste the same as the entire pot, more or less.

This is exactly what happens when a statistician selects a sample to make a prediction. Rather than surveying every voter, if we carefully choose a relatively small group of voters and ask them the question, "Who do you plan to vote for?" we can assume that the mixture of answers will be relatively close to the mixture in the full population of voters.

Population and Sample

Population: the full group that we're interested in understanding.

Example: everyone who will vote on Election Day

Sample: a small group chosen from the population that we use to estimate the response of the full group.

Example: a few voters chosen to take a poll

Because it is generally not possible or feasible to study the entire population, nearly every statistical study involves taking a sample, and using that sample to predict what the population looks like.

Solution

TRY IT

Remember how they sent out 10 million surveys and received 2 million responses? This is a red flag, because their survey was a *voluntary response* survey, in which people had to take initiative to respond. That was their second error.

First of all, their flawed sampling method skewed the sample toward more wealthy people. Remember that this took place in the midst of the Great Depression, so people who had the disposable income for a magazine subscription, or owned a car or telephone, were likely to be wealthier than the average voter. These wealthier voters had less stake in the policies of the New Deal, and since they were less likely to take advantage of its benefits, they were more likely to agree with Governor Landon that the New Deal was wasteful and inefficient.

Second, since people had to take an active step (filling out the questionnaire and mailing it in), responses were more likely to come from people who felt strongly about the question, and this tends to attract people who are unhappy with the status quo (remember that Roosevelt was the incumbent). People who were mostly content with the direction of the country, who would vote to re-elect Roosevelt, were less inclined to respond. This is a major problem with voluntary response surveys in general, and the main reason that they are generally considered unreliable.

Largely due to the very public failure of their polling system, the *Literary Digest* folded in less than two years. The same election, though, saw the debut of the well-known Gallup poll, when a pollster named George Gallup used a much smaller, carefully chosen sample to correctly predict the results.

What's the lesson? When sampling, **a representative sample is better than a large sample**. Remember the metaphor of the pot of soup; larger samples might improve the results, but not by as much as carefully stirring the pot first.

EXAMPLE 2 REPRESENTATIVE SAMPLES

Decide whether each of the following sampling methods is likely to produce a representative sample.

- (a) To find the average annual income of all adults in the United States, sample representatives in the US Congress.
- (b) To find out the most popular cereal among children under the age of 10, stand outside a large supermarket one day and poll every twentieth child under the age of 10 who enters the supermarket.

Solution

- (a) This is probably **not representative**, if for no other reason than that the salary of congresspeople is fixed by law at a single value. It happens to be much higher than the average salary in the U.S., but this is partly by design, in order to reduce the incentive to accept bribes.
- (b) This seems likely to be **representative**; while there could be regional differences, for instance, there are no obvious biases.

Okay, so if we want a good, representative sample, how do we do this? How do we make sure that our sample looks like the population (that, for instance, the racial and ethnic proportions match)? This is a hard question; in fact, it is so hard that we don't directly tackle it.¹ Instead, we rely on **randomness**.

Representative Samples

Random selection leads to representative samples.

Say you have a question about the behavior of college students, like how much they spend on textbooks in a typical semester. There are all sorts of variations that you may or may not think of; for instance, it may depend on what year they are in, since senior-level textbooks may be more expensive. It may depend on what major they are in, what college or university they attend, and a hundred other factors that may not even occur to you. You could try to build an exhaustive list of all of these factors, and then decide to get equal numbers of students in each year. Maybe then you figure out the proportions of students in each major and try to match those proportions in your sample. Is it starting to sound like a really thorny problem?

Instead, if you can find a good way to *randomly* select college students, the problem will take care of itself. If students are divided evenly in freshmen, sophomores, juniors, and seniors, and you truly select people randomly, probability suggests that you will select *approximately* a quarter of your sample from each group.

¹At times, pollsters *do* adjust their samples to reflect some demographic trends.

Random Samples

The key to a random sample is that **each member of the population is equally likely to be selected.**

Examples of Random Samples

- If the population is students in a particular classroom, number the students in the classroom from 1 to n and use a random number generator to select numbers between 1 and n .
- If the population is FCC students, list their FCC email accounts and number them, then pick random numbers between 1 and n , where n is the number of students.

Examples of Biased Samples

- If the population is residents of Frederick County, number the entries in the phone book and use a random number generator to select a sample.
- If the population is American citizens, go to the entrance of Yankee Stadium and poll everyone entering.
- If the population is FCC students, poll the students in one class.

If you're ever unclear on whether or not a sampling method is random, simply ask whether or not every member of the population has an equal chance of being selected.

Not every resident of the county has a phone, let alone a phone listed in the phone book

PICKING RANDOM NUMBERS WITH A GRAPHING CALCULATOR

Your graphing calculator can select random numbers. To access this menu, press the **MATH** button and use the left and right arrows to navigate to the PRB (probability) tab.

```
MATH NUM CPX PRB
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(
```

Selecting **rand** will select a “random” (technically pseudo-random, but close enough for us) number between 0 and 1. Selecting **randInt(** will allow you to choose a random integer between a given lower and upper bound, or several of those.

```
randInt(1,10)
10
randInt(1,10,3)
(10 2 6)
```

To select a single random integer, enter **randInt(lower bound, upper bound)**, with whatever numbers you want for the lower and upper bounds. To select n random integers, enter **randInt(lower bound, upper bound, n)**.

CARTWHEELS

A coach is interested in how many cartwheels the average college freshman can do at his university. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping. Is this sample random and representative?

The population is the class of all freshmen at the coach's university. The sample is composed of all freshmen so that is good. However, the sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman: people who cannot do cartwheels probably did not volunteer! Hence, this sample is not random, and thus unlikely to be representative.

EXAMPLE 3

Solution

A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their responses that the class did extremely well. What is the sample and population? Is the sample random and representative? Why or why not?

TRY IT

Sampling Methods

Naturally, this brings up another question: how do we pick a random sample? There are many ways to do this; let's look at a few.

Simple Random Sampling

Number every member of the population and use a random number generator (like a calculator) to pick as many members as you need for the sample.

Simple random sampling is the most common method; most studies default to this one unless there's a good reason to use a more exotic approach. The hardest part is often acquiring as thorough a list of the population as possible.

Convenience Sampling

Pick members of the population that are easy to pick.

Convenience sampling is generally not as reliable as other methods, but it is the easiest. There are cases, though, where the benefits outweigh the costs. For instance, say you were the quality assurance officer at a facility that manufactured something heavy like cinder blocks. If a full pallet is delivered, and you need to select a few blocks to test for strength, a simple random sample would require digging through the full stack and moving a lot of the blocks. Instead, you could assume that this pallet all comes from the same batch, and so testing one of the blocks on the top layer will give you a representative idea of the batch quality. It can be dangerous to make assumptions, but those with specialized knowledge can sometimes do so.

Systematic Sampling

Randomly pick a starting point and select every n th member.

With systematic sampling, the randomness all comes from the random starting point, so that part of the process is crucial; you can't start from the beginning of the list and select every third person and call that systematic sampling.

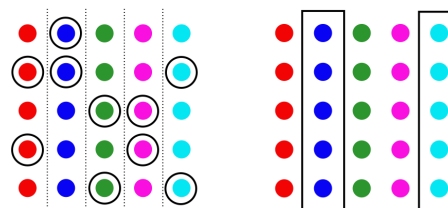
Stratified Sampling

Divide the population into categories (for instance, divide college students into freshmen, sophomores, juniors, and seniors), then randomly select an equal number from each category.

Cluster Sampling

Like stratified sampling, divide the population into groups, but this time, select one or more *entire* groups, rather than a few from each group.

It's easy to confuse stratified and cluster sampling, because they sound similar.



Stratified sampling

Cluster sampling

Stratified sampling means picking a few from *each* cluster; cluster sampling means selecting a few *entire* clusters.

For instance, suppose you're canvassing a neighborhood to poll residents about a new local ordinance, and say this neighborhood consists of 10 apartment building, with 12 apartments in each building. If you decided to use stratified sampling, you could enter each building and use a random number generator to pick 2 apartments; this would give you a random sample of 20 homes. On the other hand, if you used cluster sampling, you could randomly select two buildings and survey every apartment in those buildings, for a total of 24 responses. In this case, if there's no need to get a response from each building, cluster sampling would be more convenient, since you wouldn't need to travel as far.

SAMPLING METHODS

EXAMPLE 4

Determine the type of sampling used in each of the following scenarios.

- (a) A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

This is **stratified sampling**, since the group is divided into categories, and a few are taken from each category.

Solution

- (b) A pollster interviews all human resource personnel in five different high tech companies.

This sounds like **cluster sampling**; the companies are the clusters, and these five companies have been selected in entirety.

Solution

- (c) A high school educational counselor interviews 50 female teachers and 50 male teachers.

This is **stratified sampling**, with teachers categorized by gender.

Solution

- (d) A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

The key here is that *every third* patient is selected; that's the sign of **systematic sampling**.

Solution

- (e) A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

There are no categories, and no systematic progression through a list; this is purely random, and indeed is **simple random sampling**.

Solution

- (f) A student interviews classmates in his algebra class to determine how many pairs of jeans a student at his school owns, on the average.

Since this student simply interviews the nearest available students, this represents **convenience sampling**; there was no attempt at randomness.

Solution

EXAMPLE 5 QUIZ SCORE SAMPLES

Use the random number generator on your calculator to generate different types of samples from the data below. Find the average score for each sample and compare the results for each method.

This table displays six sets of quiz scores (out of 10 points) for an elementary statistics class.

A	B	C	D	E	F
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

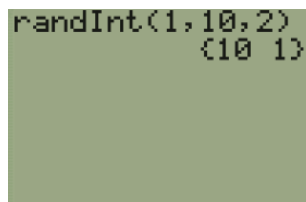
Create a sample of 12 scores using each of the following methods.

- Stratified sampling
- Cluster sampling
- Simple random sampling
- Systematic sampling
- Convenience sampling

Solution

- To create a stratified sample, we need to decide what to use as the categories (also called *strata*). Since the data is separated into 6 categories already (A-F), we can take advantage of this natural division. To get a sample of 12 scores, we need to select 2 from each category. We'll show the process for group A, but omit the rest, since it follows the same pattern.

Group A: to select two scores from this category, we need to generate two random values between 1 and 10 (since there are 10 scores in the group). We'll use the **randInt** function on the calculator. Remember, to access this, press the **MATH** button and navigate to the **PRB** menu, then look for the option labeled **5: randInt(**. Since we want two values, we'll type in **randInt(1,10,2)** (the comma button is above the **7** key).



```
randInt(1,10,2)
{10 1}
```

This means that we have selected the 10th and 1st values in the list; if we scan down the list in the first column, the 10th value is 8 and the 1st value is 5, so the first two values in our sample are 8 and 5.

For the other groups, we repeat this process. If you follow along on your calculator, you may get different results, but the final sample is

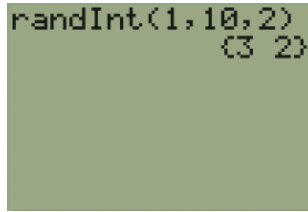
8, 5, 7, 7, 9, 10, 9, 5, 7, 7, 4, 7

We'll cover this in more detail in another section, but you are probably familiar with the average; to calculate it, add up all the values and divide by how many values there are.

$$\begin{aligned}\text{Average score:} &= \frac{8 + 5 + 7 + 7 + 9 + 10 + 9 + 5 + 7 + 7 + 4 + 7}{12} \\ &= \boxed{7.1}\end{aligned}$$

- (b) Since we want a total of 12 scores in our sample, we need to define clusters in such a way that a few of them will total 12 in size. We could, of course, define the columns to be the clusters, but then if we selected two columns, for instance, we would need to toss out some values.

Instead, it will be simpler to define the rows of the table as the clusters; this way, we can randomly select two of them and wind up with 12 values. Use the calculator as before; this time, we only need to select random values once:



```
randInt(1,10,2)
{3 2}
```

We selected the 2nd and 3rd rows, so the sample consists of these values:

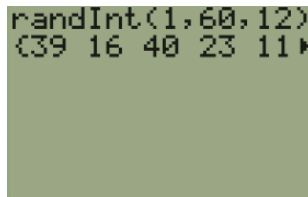
$\boxed{10, 5, 9, 8, 7, 6, 9, 10, 8, 6, 7, 9}$

Calculate the average:

$$\begin{aligned}\text{Average score:} &= \frac{10 + 5 + 9 + 8 + 7 + 6 + 9 + 10 + 8 + 6 + 7 + 9}{12} \\ &= \boxed{7.8}\end{aligned}$$

- (c) Simple random sampling entails listing all of the values (in some order) and randomly selected 12, in our case, from the full pool. We need to decide on some order; for no particular reason, let's list the values in order by category. So the scores in group A will come first, followed by B, and so on. This means when we count, we'll start in the upper left-hand corner of the table, and count down one column at a time.

There are a total of 60 values; using the calculator, we can select 12 of them by typing `randInt(1,60,12)`. When selecting this many values, there's a good chance that some values will be repeated. To avoid this, there is another option in the menu called `randIntNoRep`, which will avoid repeated values. However, in this case, we won't bother with this; we'll allow repetition.



```
randInt(1,60,12)
{39 16 40 23 11}
```

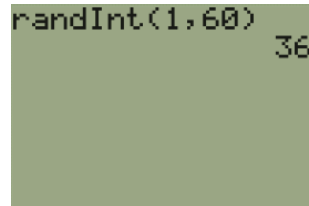
We start by looking for the 39th value in the list, then the 16th, and so on (this quickly gets tedious). The result is

$\boxed{7, 9, 9, 8, 7, 9, 10, 9, 8, 10, 8, 9}$

and the average is

$$\begin{aligned}\text{Average score:} &= \frac{7 + 9 + 9 + 8 + 7 + 9 + 10 + 9 + 8 + 10 + 8 + 9}{12} \\ &= \boxed{8.6}\end{aligned}$$

- (d) To create a systematic sample, we simply need a starting point and a step size. If we decide to pick every 3rd value, we can use the calculator to pick the starting point:



```
randInt(1,60)
36
```

Starting at the 36th position and counting that value and every 3rd one after that yields the following sample:

10, 7, 7, 7, 8, 3, 9, 8, 7, 9, 9, 9

Notice that we wrapped around at the end of the list and came back to the beginning. The average is

$$\begin{aligned}\text{Average score: } &= \frac{10 + 7 + 7 + 7 + 8 + 3 + 9 + 8 + 7 + 9 + 9 + 9}{12} \\ &= \boxed{7.8}\end{aligned}$$

- (e) Finally, a convenience sample simply means picking the easiest values; let's use the top two rows, since we can easily read and copy those without jumping around the table:

5, 7, 10, 9, 8, 3, 10, 5, 9, 8, 7, 6

$$\begin{aligned}\text{Average score: } &= \frac{5 + 7 + 10 + 9 + 8 + 3 + 10 + 5 + 9 + 8 + 7 + 6}{12} \\ &= \boxed{7.3}\end{aligned}$$

Notice how the average for each sample was slightly different, but all relatively close to each other, ranging from 7.1 to 8.6. This illustrates an important point: **different samples from the same population will give different results**, but as long as the sample size is not too small and the sample is not skewed or biased in some way, the results should be relatively similar.

Larger samples tend to be more consistent, but as we saw with the example of the *Literary Digest*, larger sample sizes alone don't guarantee better outcomes.

This point about the natural variability of samples is a good one to keep in mind during election season, for instance, when multiple polls are released, and they all give different levels of support for one candidate or another. A good approach is to compare a variety of reputable polls in order to gain a fuller picture; the website FiveThirtyEight, among others, does this well.

Exercises 3.1

For problems 1–2, identify the population and sample.

1. A political scientist surveys 28 of the current 106 representatives in a state's congress. Of them, 14 said they were supporting a new education bill, 12 said they were not supporting the bill, and 2 were undecided.

2. The city of Frederick has 9500 registered voters. There are two candidates for the city council in an upcoming election: Marfani and Rahman. The day before the election, a telephone poll of 350 randomly selected registered voters was conducted. Of them, 112 said they would vote for Marfani, 207 said they would vote for Rahman, and 31 were undecided.

For problems 3–4, decide whether the sampling method described is likely to produce a representative sample, and why or why not.

3. A marriage counselor is interested in the studying divorce rates, so she gives her clients a survey.

4. A fitness center wants to see how much use their treadmills get, so they pick random times during the day and record how many treadmills are in use each time.

For problems 5–10, determine the type of sampling used in the given scenario.

5. A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

6. To check their accuracy, the Census Bureau draws a sample of several city blocks and recounts everyone in those blocks.

7. A pollster walks around a busy shopping mall and asks people passing by how often they shop at the mall.

8. Police at a DUI checkpoint stop every tenth car to check whether the driver is sober.

9. A restaurant samples 100 sales from the past week by numbering all their receipts, generating 100 random numbers, and picking the receipts that correspond to those numbers.

10. The provost at a university wants to know how a particular policy is affecting faculty, so she randomly selects 3 members of each department to survey.

SECTION 3.2 Visualizing Data



Suppose you're doing some sort of study on players in the NBA. You decide that rather than gathering data on *all* NBA players, you can select a sample of, say, around 30 players. In fact, you realize that since there are 30 teams, you can use stratified sampling and randomly select one player from each team.

Note: the data below is from the 2019-20 NBA season.

Name	Team	No.	Age	Position	Height	Pts	Reb	Salary
Jayson Tatum	Celtics	0	22	PF	2.03 m	23.4	7.0	\$7,830,000
Joe Harris	Nets	12	28	SF	1.98 m	14.5	4.3	\$7,666,667
RJ Barrett	Knicks	9	20	SG	1.98 m	14.3	5.0	\$7,839,960
Ben Simmons	76ers	25	24	PG	2.08 m	16.4	7.8	\$8,113,930
Matt Thomas	Raptors	21	26	SG	1.93 m	4.9	1.5	\$898,310
Daniel Gafford	Bulls	12	21	PF	2.08 m	5.1	2.5	\$898,310
Andre Drummond	Cavaliers	3	27	C	2.08 m	17.7	15.2	\$27,093,019
Langston Galloway	Pistons	9	28	SG	1.85 m	10.3	2.3	\$7,333,333
Justin Holiday	Pacers	8	31	SF	1.98 m	8.3	3.3	\$4,767,000
Khris Middleton	Bucks	22	29	SF	2.01 m	20.9	6.2	\$30,603,448
Skal Labissiere	Hawks	7	24	PF	2.08 m	5.8	5.1	\$2,338,847
PJ Washington	Hornets	25	22	PF	2.01 m	12.2	5.4	\$3,831,840
KZ Okpala	Heat	4	21	SF	2.03 m	1.4	1.0	\$898,310
Wes Iwundu	Magic	25	25	SF	1.98 m	5.8	2.5	\$1,618,420
Rui Hachimura	Wizards	8	22	PF	2.03 m	13.5	6.1	\$4,469,160
Andrew Wiggins	Warriors	22	25	SF	2.01 m	21.8	5.1	\$27,504,630
Paul George	Clippers	13	30	SG	2.03 m	21.5	5.7	\$30,560,700
Avery Bradley	Lakers	11	29	PG	1.91 m	8.6	2.3	\$4,767,000
Jalen Lecque	Suns	0	20	PG	1.93 m	2.0	0.4	\$898,310
Harry Giles III	Kings	20	22	C	2.11 m	6.9	4.1	\$2,578,800
J.J. Barea	Mavericks	5	36	PG	1.78 m	7.7	1.8	\$1,620,564
Bruno Caboclo	Rockets	5	24	SF	2.06 m	3.0	2.0	\$1,845,301
Josh Jackson	Grizzlies	20	23	SG	2.03 m	9.0	3.1	\$7,059,480
JJ Redick	Pelicans	4	36	SG	1.91 m	15.3	2.5	\$13,486,300
Trey Lyles	Spurs	41	24	C	2.06 m	6.4	5.7	\$5,500,000
Troy Daniels	Nuggets	30	29	SG	1.93 m	4.3	1.1	\$384,541
Naz Reid	Timberwolves	11	21	C	2.06 m	9.0	4.1	\$898,310
Luguentz Dort	Thunder	5	21	SG	1.91 m	6.8	2.3	\$155,647
Jusuf Nurkic	Trail Blazers	27	26	C	2.13 m	17.6	10.3	\$13,125,000
Donovan Mitchell	Jazz	45	23	SG	1.85 m	24.0	4.4	\$3,625,760

Now, how helpful is the table above? How quickly can you glean information from it, or get an answer to a question you may have; for instance, if you wanted to have a sense of how diverse the salaries are in the NBA, would the list in the table above give you that?

It's unlikely that we can draw meaningful conclusions from a list of data like this; when we scan a table like this, there's no good way to aggregate the results that our eyes pass over.

Because of this, much of statistics is concerned with describing and interpreting data in ways that we can quickly grasp. In this section, we'll see how to use visual displays to do this; graphs are good because we can gather a lot of information at a glance. In the next section, we'll discuss the use of numerical summaries, which are not quite as user-friendly as pictures, but more precise, and equally instructive once you know how to read them. By using a combination of both approaches, we can get a full picture of a dataset like the one above.

Types of Data

Before we start summarizing data, we need to recognize that there are different types of data that need to be treated differently. As an example, notice that one variable recorded above is the team that each player plays for, which cannot be treated the same way as their height.

If we wanted to compare players' heights, we could find the average, among other things. This doesn't make any sense with their teams; we can't calculate an average, nor would it mean anything if we could. This brings us to our first distinction: numerical versus categorical data.

Numerical and Categorical Data

Numerical (or quantitative) data: a quantitative variable is one that we count or measure.

Categorical (or qualitative) data: a qualitative variable is one that divides items into categories.

Notice that of the nine columns in the table, two of them are identifiers (name and jersey numbers) and thus not really variables that we're interested in summarizing. You could certainly take the average of the jersey numbers, but what would that answer tell you?

We can break down the remaining seven columns into numerical and categorical variables:

Variable	Type
Team	Categorical
Age	Numerical
Position	Categorical
Height	Numerical
Points	Numerical
Rebounds	Numerical
Salary	Numerical

Note: in case you're not familiar with positions in basketball, there are five players on the court for each team at any time: the point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C), and the order listed roughly corresponds to the typical size of the players at each position (guards tend to be small and quick, and forwards and centers tend to be larger and stronger).

This distinction between numerical and categorical data is important because we will draw different types of graphs for the two different types.

There is another distinction within numerical data; notice that the definition above says that a numerical variable is one that we "count or measure." It turns out that these two options distinguish two types of numerical data: discrete and continuous variables.

Discrete and Continuous Data

Discrete variable: a discrete variable is one that is counted, and it is limited to specific values. For example, the number of children that a person has will always be a whole number; it cannot be anything in between.

Continuous variable: a continuous variable is one that is measured; the values can be any number in a given range. For example, a person's height can be measured as precisely as desired, so it's not limited to specific values.

The line between discrete and continuous variables can sometimes be fuzzy. For instance, is age a discrete or continuous variable in the NBA dataset? You could make an argument either way. Someone's age can be measured as precisely as desired, down to the minute or second or further, but as it is listed in the table (and usually written) it's given in terms of years, so it is limited to specific values (whole numbers).

The good news is that this distinction is not as crucial for us in this chapter; the work that we'll do with quantitative data will not change from discrete to continuous variables. It is good to be aware of the distinction, but especially in fuzzy cases like age, don't spend too much time worrying about which category that variable falls into.

At this point, we're ready to start summarizing data with charts and graphs; for the examples that follow, we'll keep returning to the NBA dataset above.

Dot Plots

A dot plot is one of the simplest kinds of graphs, and it can be drawn for both numerical and categorical data. We'll use the age of the NBA players as an example.

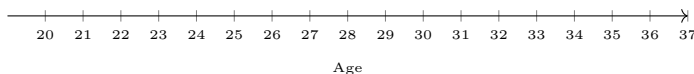
EXAMPLE 1 DOT PLOT

Draw a dot plot to summarize the following data, the ages of 30 randomly chosen NBA players:

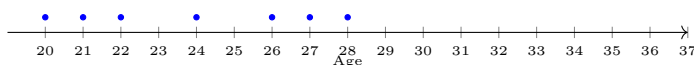
22, 28, 20, 24, 26, 21, 27, 28, 31, 29, 24, 22, 21, 25, 22,
25, 30, 29, 20, 22, 36, 24, 23, 36, 24, 29, 21, 21, 26, 23

Solution

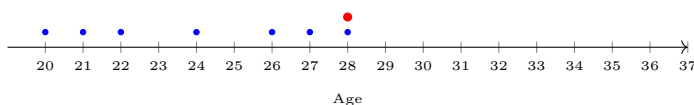
To begin, draw an axis scaled to cover the full range of the data. The age values range from 20 to 36, so our scale needs to extend at least that far.



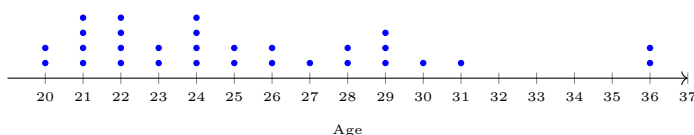
Now, for each value in the dataset, place a dot at that position. After placing the first seven values (22, 28, 20, 24, 26, 21, and 27), the graph looks like this:



At the eighth value, we encounter our first duplicate (28). Simply place this dot above the one that is already placed at 28.



The final picture is shown below, after all the data is accounted for.



It's important as we start adding duplicate values that we space these dots evenly so that we can see at a glance where the grouping of the data occurs.

TRY IT

Draw a dot plot for the NBA players' positions.

What you'll see as we encounter other graph types in this section is that the primary goal of most of them is to visualize how data is arranged, where it is tightly grouped together and where it is spread out. This is true for dot plots as well.

For instance, looking at the dot plot in the previous example, what jumps out is that most of the players are between the ages of 20 and 29; only four of the players are 30 or older, and only two of them are older than 31. Those two oldest players, at 36, are examples of what we often call *outliers*, which are data points that are isolated from the main cluster of points.

Limitations: a dot plot is designed to handle relatively small amounts of data; they are tedious to build and hard to read for large datasets. For the dataset in the example, with 30 observations, a dot plot is a good way to get a quick glimpse of the data arrangement.

Frequency Distributions

The next tool, a frequency distribution or frequency table, is actually just another form of a dot plot. Rather than drawing a dot for each individual value, though, a frequency table simply counts the **frequency** of each value, or how many times it occurs.

For instance, in the example of the NBA players' ages, there are two players aged 20, four at 21, and so on. Continuing the pattern, we can build the table below.

Age	Frequency
20	2
21	4
22	4
23	2
24	4
25	2
26	2
27	1
28	2
29	3
30	1
31	1
32	0
33	0
34	0
35	0
36	2

In a frequency table, the first column lists the variable that we're summarizing and all the values of the variable that appear in our dataset; the second column lists the corresponding frequency for each value (note that the frequencies should all add up to 30, the size of the full dataset).

Notice that there were several missing values (from 32 to 35), but we included them in the table anyway, and recorded zeros for the frequency. This is mostly a matter of preference; some would remove those empty rows. However, since our goal is to visualize how the data clusters and spreads out, those rows give the same sense that we got from the dot plot, that the two players at 36 are unusual within our dataset.

Frequency tables often include a third column, which contains the **relative frequency**.

Relative Frequency

The **relative frequency** of a value is the proportion (or percentage) of its frequency to the total size of the dataset.

Note: the relative frequencies of all values will always add up to 1 (although because of rounding, this may not always appear to be true).

For instance, the first age, 20, appears twice in the full dataset, so its relative frequency is 2 out of 30, which could be simplified to 1 out of 15 (but generally we don't simplify these fractions) or written as a decimal (0.067) or percentage (6.7%). Any of these representations is perfectly acceptable.

If we add this column to the frequency table, we get the following:

Age	Frequency	Relative Frequency
20	2	$2/30$ or 0.067
21	4	$4/30$ or 0.133
22	4	$4/30$ or 0.133
23	2	$2/30$ or 0.067
24	4	$4/30$ or 0.133
25	2	$2/30$ or 0.067
26	2	$2/30$ or 0.067
27	1	$1/30$ or 0.033
28	2	$2/30$ or 0.067
29	3	$3/30$ or 0.100
30	1	$1/30$ or 0.033
31	1	$1/30$ or 0.033
32	0	$0/30$ or 0.000
33	0	$0/30$ or 0.000
34	0	$0/30$ or 0.000
35	0	$0/30$ or 0.000
36	2	$2/30$ or 0.067

EXAMPLE 2 DAILY COMMUTE

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. They responded as follows:

2, 5, 7, 3, 2, 10, 18, 15, 20, 7, 10, 18, 5, 12, 13, 12, 4, 5, 10.

This is summarized in the frequency table below:

Data Value	Frequency	Relative Frequency
3	3	$3/19$
4	1	$1/19$
5	3	$3/19$
7	2	$2/19$
10	3	$4/19$
12	2	$2/19$
13	1	$1/19$
15	1	$1/19$
18	1	$1/19$
20	1	$1/19$

- (a) Is the table correct? If it is not correct, what is wrong?

It is incorrect, because the frequency column sums to 18, not 19 as it should. One of the data values was left out. Besides, two people responded that they commute 2 miles, and that doesn't appear on the table at all.

- (b) True or false: Three of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.

False. The frequency for 3 miles should be 1. When building the table, the two that responded 2 miles got lumped into the 3 category.

- (c) What fraction of the people surveyed commute five or seven miles?

$5/19$

- (d) What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between 5 and 13 miles (not including those who commute exactly 5 or exactly 13 miles)?

$7/19$, $12/19$, $7/19$

We can also create frequency distributions for categorical data.

CATEGORICAL FREQUENCY DISTRIBUTION

EXAMPLE 3

Create a frequency table for the NBA players' positions from the dataset at the beginning of this section, including a relative frequency column.

Since there are five possible results for position, we can start the table by creating the first column with these five values. Then simply count the frequency of each position and record the result; the results are below.

Solution

Position	Frequency	Relative Frequency
PG	4	13.3%
SG	9	30.0%
SF	7	23.3%
PF	5	16.7%
C	5	16.7%

Clearly, the most popular positions are in the middle of the list, shooting guards and small forwards. This makes sense because these tend to be versatile players who can fill multiple roles, so teams are more likely to sign them.

Grouped Frequency Distributions

What if we wanted to build a frequency distribution for points per game in the NBA dataset? Since there is very little, if any, repetition in the values, the resulting table would have a lot of columns with 0's or 1's in them, and it wouldn't give us any sense of the grouping of the data.

Instead, we can create a **grouped frequency distribution**, which counts the frequency of specific *ranges* instead of specific *values*.

For instance, suppose we split the values from 0 to 25 into **classes** that cover 5 points each. The first class would cover everything from 0 to 5, the second from 5 to 10, and so on.

Hold On! You may have already spotted a potential problem. What if someone averages exactly 5 points; in which category do we count them? If the classes go from 0 to 5 and 5 to 10, the overlap causes a problem. To fix this, we'll make the first class from 0 to **just less than 5**, so that if a player averages 4.9 points per game, he'll belong to the first class, and if he averages 5.0 points per game, he'll belong to the second class.

This is a very important point, because this is the most common error that students make in creating grouped frequency tables; be careful with this.

Here are our classes:

Points per Game

0.0 – 4.9
5.0 – 9.9
10.0 – 14.9
15.0 – 19.9
20.0 – 24.9

There are, of course, values between 4.9 and 5.0, so instead of writing 4.9, we could also write "less than 5." However, as long as we're only considering the first decimal point, this is clear enough that anyone reading it can understand that what we really mean is "less than 5." If we counted two decimal places, we'd need to replace 4.9 with 4.99 to make ourselves clear.

Choosing Classes

When building a grouped frequency distribution, you'll usually have the freedom to choose how you want to separate the classes. Here are some guidelines you should follow:

- Each class should be the same width. Notice in the example above that each class was five units wide. If not, the grouped frequency distribution would not give a clear picture of how the data is arranged.
- Classes cannot overlap.
- Avoid empty classes if possible. This can occur if you choose to have too many classes.
- Don't make open-ended classes. For instance, in the example above we didn't make the last class "20 and above," which would have been an open-ended class. The reason not to do this is that it violates the first guideline about having all classes have the same width.

To find the appropriate class width to use, you can start by deciding on the number of classes you want to use, and the class width is found by dividing the distance from the minimum to the maximum by the number of classes.

Class Width

$$\text{Class Width} = \frac{\text{Maximum} - \text{Minimum}}{\text{Number of Classes}}$$

Round UP to the next whole number.

EXAMPLE 4

GROUPED FREQUENCY DISTRIBUTION

Build a grouped frequency table for points per game for the NBA players dataset, using a class width of 5.

Solution

We already defined the classes; all that's left is to count the number of players that fall into each range:

Points per Game	Frequency	Relative Frequency
0.0 – 4.9	5	16.7%
5.0 – 9.9	11	36.7%
10.0 – 14.9	5	16.7%
15.0 – 19.9	4	13.3%
20.0 – 24.9	5	16.7%

The most common category by far is the range from 5 to 10 points (just less than 10, technically). The other players are equally split among the other four categories.

TRY IT

Create a grouped frequency distribution for the salaries of the NBA players in the same dataset, using a class width of \$5 million.

Remember that frequency distributions are really the same concept as dot plots; the results are displayed differently, but both basically show a list of categories and the count in each category.

Keep this in mind, because this trend will continue; all of the next three graph types present variations on this same theme. As you'll see, this goal of visualizing how data is arranged—where it is clustered and where it is spread out—is a very popular one, so we have many tools to achieve it.

Histograms and Bar Charts

We'll lump two of the types of graphs together, because they are very similar. If you go back and look at the example of a dot plot at the beginning of this section, and you blur the picture, what you'll see is basically a series of bars rising up from a horizontal axis, and the height of each bar represents how many times that value appears (its frequency, in other words).

This is exactly what a histogram or bar chart does; it simply replaces the series of dots with a bar of corresponding height. Generally, we'll start by building a frequency distribution, and then draw a bar for each row of the table. Thus, a histogram/bar chart contains *exactly* the same information as the corresponding frequency table; the only difference is how this information is expressed. Histograms and bar charts are used because we'd often rather look at a picture of the grouping than read a list of numbers in a table and interpret them.

What's the difference? Okay then, what differentiates a histogram from a bar chart? The difference is what kind of data we're summarizing: numerical or categorical. Remember that we can draw a dot plot or frequency table for both types. However, when we draw a histogram/bar chart, we make a (relatively small) distinction between them.

Visually, the difference is subtle, but it's based on the difference in the way that we can think of these two types of data. With numerical data, there's a natural transition from one category to the next (from the 0.0 – 4.9 class to the 5.0 – 9.9 class, for instance), so we draw the bars in an unbroken sequence, without gaps between them. Categorical data, on the other hand, has no such smooth transition; there's a sharp break between categories (the Knicks and the Nets are completely distinct, with no blending from one to the other). To visualize this, we draw the bars with gaps between them.

Other than this, there is no fundamental distinction between histograms and bar charts.

HISTOGRAM

EXAMPLE 5

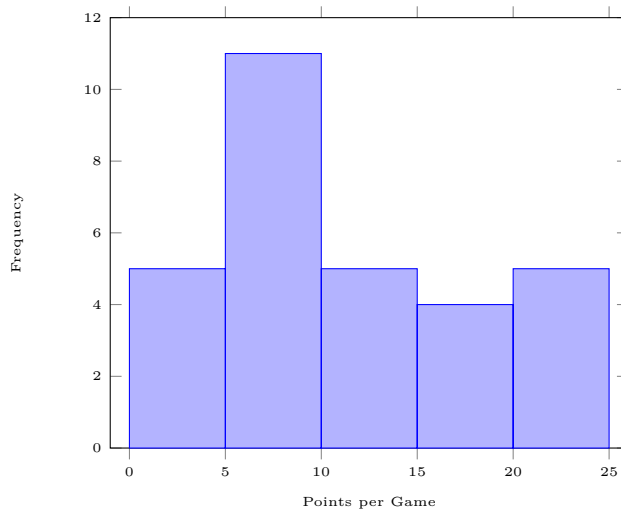
Build a histogram for points per game in the NBA dataset, using grouped classes with a class width of 5.

We already built the frequency table for this one:

Points per Game	Frequency
0.0 – 4.9	5
5.0 – 9.9	11
10.0 – 14.9	5
15.0 – 19.9	4
20.0 – 24.9	5

Solution

Now all we have to do is represent this with a graph. Note that we will always draw vertical bars in this book, but this is simply a matter of preference; you can find histograms and bar charts that are drawn horizontally as well.



TRY IT

Build a histogram for the players' ages in the NBA dataset. There is no need to use grouping.

Note that although we don't show it in this section, we could also draw a histogram for the *relative frequency*. It turns out that the picture is identical, except that the vertical axis is relabeled; that's why we don't bother to cover that here.

Let's do an example with categorical data.

EXAMPLE 6 BAR CHART

Build a bar chart for the players' positions in the NBA dataset.

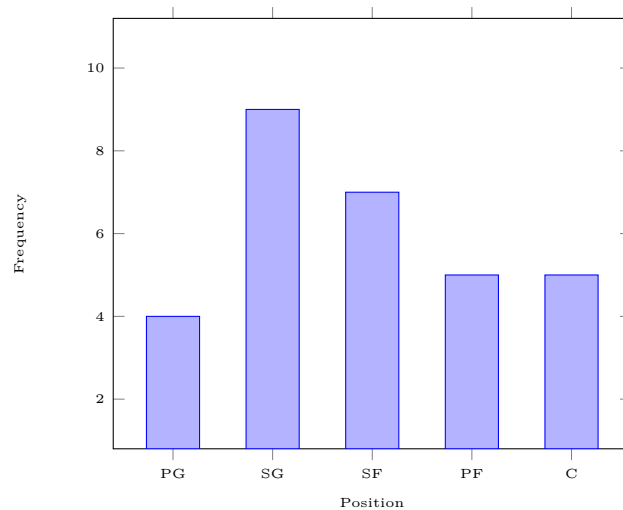
Solution

Remember that the reason we're calling this a bar chart is simply that position is a categorical variable rather than a numerical one.

Once again, we've already created the frequency table:

Position	Frequency
PG	4
SG	9
SF	7
PF	5
C	5

Now, draw the bar chart (note the gaps between the bars):

**Stem-and-Leaf Plots**

There is one major downside to grouped frequency distributions: some of the data gets lost in the summary. In other words, maybe in an example all we know is that there are 10 observations between 20 and 25, but we don't know exactly what all those observations are. This is an example of the trade-off between clarity and precision: often, the more precise we are, the less clear our summary will be.

To split the difference and display the data in a way that exhibits where it is clustered without losing any information about the data, we can use a **stem-and-leaf** plot. Here, the data is grouped by tens; each tens value is a stem, and all the data points that have that tens value are listed as the leaves. We'll illustrate with an example.

STEM AND LEAF PLOT

EXAMPLE 7

Suppose you gathered data on how long it took you to get ready in the morning. For 40 days, you measured the amount of time between when your alarm went off and when you left the house. The results are below, rounded to the nearest minute:

35	28	25	23	23	32	29	19	21	13
24	26	25	31	30	20	25	29	37	26
32	36	18	17	15	24	21	16	19	30
38	27	22	24	28	17	31	32	21	28

Build a stem-and-leaf plot for this data.

The tens places are 1, 2, and 3, and each of them gets a category:

Stems	Leaves
1	
2	
3	

Finally we go through (carefully) and find each value that begins with a 1 and list the ones place of each of them under the first category, and similarly with the other two categories.

Stems	Leaves
1	3 5 6 7 7 8 9 9
2	0 1 1 1 2 3 3 4 4 4 5 5 5 6 6 7 8 8 8 9 9
3	0 0 1 1 2 2 2 5 6 7 8

Notice that we arranged the leaves in order; this isn't strictly necessary, but it makes the data a bit more orderly.

Once again, this stem-and-leaf plot illustrates where the data is clustered, as the length of each row of leaves is equivalent to the height of a bar on a histogram, but it does this without losing any information. In other words, if we were simply given the stem-and-leaf plot, we could completely recreate the data set.

For three-digit data values (or longer), the leaves are usually still the last digits (the unit digits), and the stems are everything before that. For instance, observe the data set below and the corresponding stem-and-leaf plot.

135	128	125	123	123	132	129	119	121	113
124	126	125	131	130	120	125	129	137	126

Stems	Leaves
11	3 9
12	0 1 3 3 4 5 5 5 6 6 8 9 9
13	0 1 2 5 7

Build a stem-and-leaf plot for the NBA players' rebounds per game, using the ones place as the stems and the tenths place as the leaves.

Solution

TRY IT

Scatterplots

The final type of graph that we'll consider, scatterplots are different from the rest; the rest are all concerned basically with frequency, although they simply display the results in different ways, and they can be used with either numerical or categorical data. On the other hand, scatterplots are designed to compare one numerical variable to another, and we can use them to spot a connection or *correlation* between the two.

For instance, in the NBA dataset, we may suspect that there is some relationship between the number of points per game that a player averages and his salary. We'd expect to see that players who score more points are likely paid more. Looking at the raw data in the table, there's no way to observe that, but if we draw a scatterplot, we can see whether our guess is correct.

To draw a scatterplot, we will call one of our variables x and the other y . It is worth understanding the distinction between the two, but the good news is that if we reverse the choice, the fundamental relationship isn't changed, and we can still do everything we'd like to.

Explanatory Variable

When drawing a scatterplot, ask yourself this question:

“Which variable determines the other?”

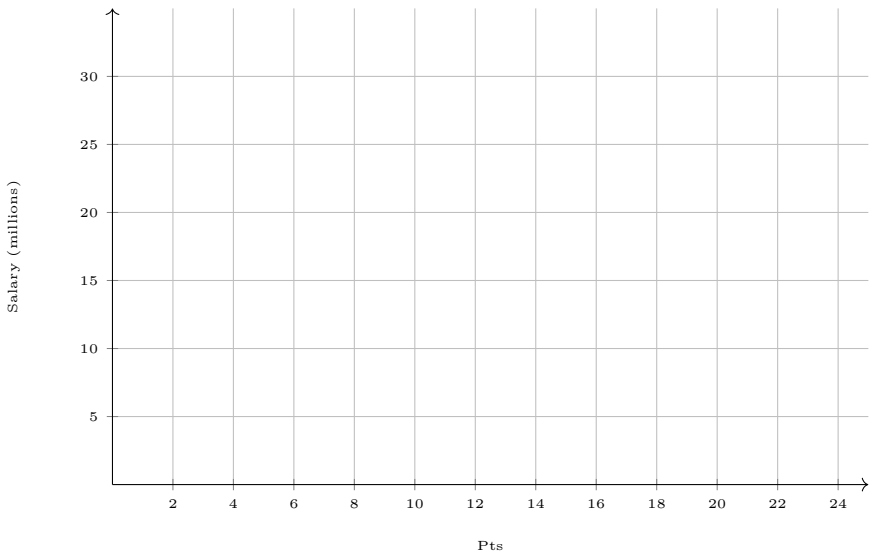
Your answer will be called the *explanatory variable* and labeled as x .

In our example, does it make more sense to say that the number of points a player scores determines his salary, or that his salary determines how many points he scores? It should be clear that x should represent points, and y should thus represent salary. Remember, though, if you reverse this, the basic relationship will still be evident.

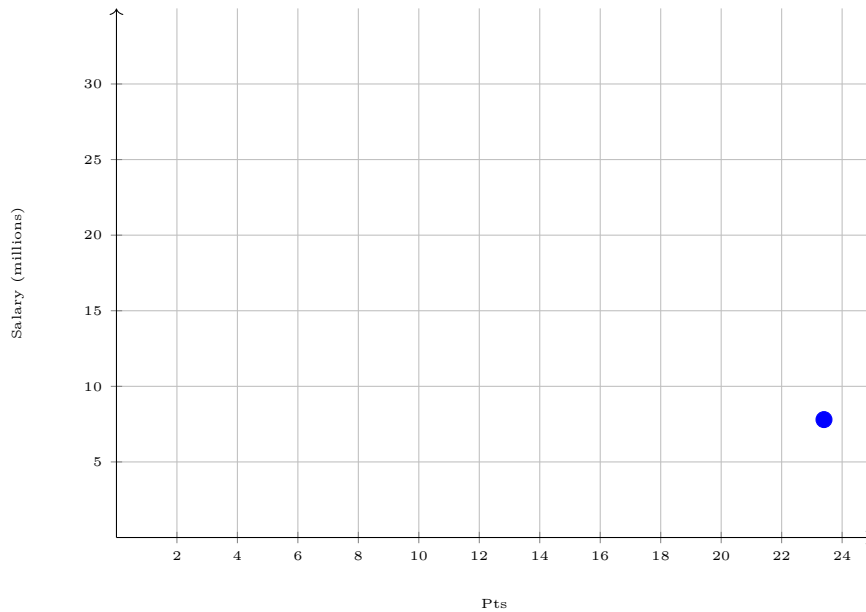
To illustrate the process, let's use only the first 10 rows of the full dataset, for the sake of simplicity. After labeling our variables, here's what we have:

Pts (x)	Salary (y) in millions
23.4	7.8
14.5	7.7
14.3	7.8
16.4	8.1
4.9	0.9
5.1	0.9
17.7	27.1
10.3	7.3
8.3	4.8
20.9	30.6

Now draw a grid with x on the horizontal axis and y on the vertical axis:

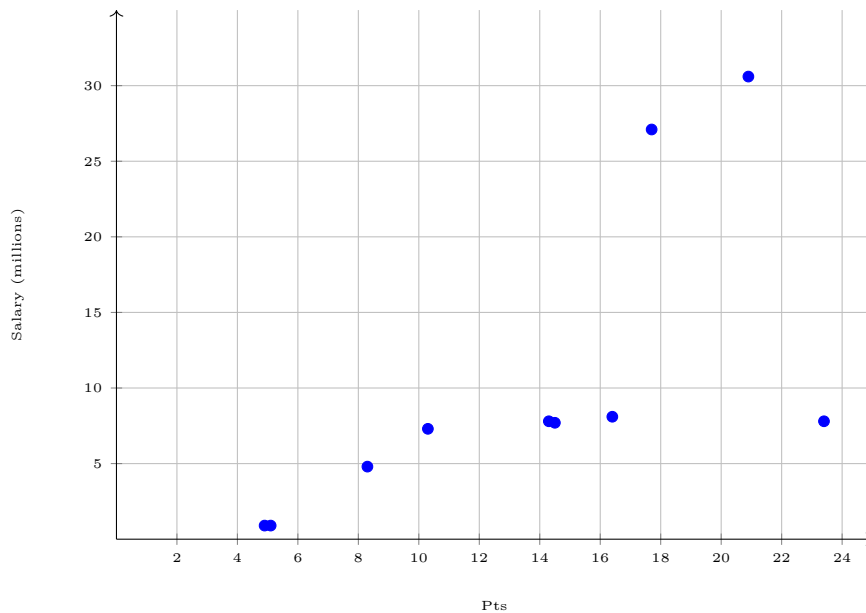


Each row of the table will correspond to a dot on this grid; the horizontal position is defined by x , and the vertical position is defined by y . For instance, the first dot will have a horizontal position of 23.4 and a vertical position of 7.8, which is shown below:



Notice that when we draw scatterplots by hand, the positions are all approximate, but since the goal is to get a big-picture view of the relationship between the two variables, that's okay.

Once we draw all the points, here is the final picture:



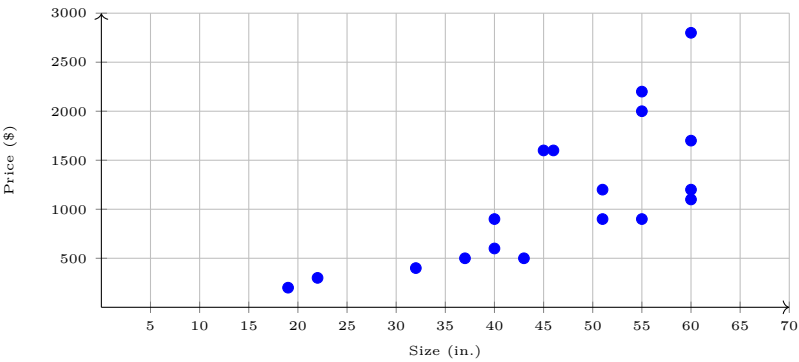
Notice that yes, there is an overall trend toward higher scorers being paid more, but that is not an unyielding rule; for instance, the first player (Jayson Tatum) is the highest scorer in this smaller list, but he's not nearly the highest-paid player. Clearly there are other factors at play (age, position, team, whether a player is still on his rookie contract, etc.). This is often the case; scatterplots can hint at relationships, but there are often other unseen relationships at play.

EXAMPLE 8 SCATTERPLOT: TV PRICE

The following table shows, for a sample of Samsung LCD TVs, their size and their price. Construct a scatterplot for this data.

Size (in.)	Price (\$)	Size (in.)	Price (\$)
43	500	60	1200
55	900	45	1600
51	900	19	200
32	400	55	2200
51	1200	60	1700
37	500	55	2000
60	2800	22	300
60	1100	40	600
46	1600	40	900

Solution We could pick either variable to be x , but it makes more sense to say that the size of a TV determines its price than to say that the price of a TV determines its size. The scatterplot below shows the relationship between the two.



Notice again that while there is a clear upward trend, meaning that larger TVs tend to be more expensive, there is plenty of variation, so there are other factors that play into the price, such as resolution or model.

TRY IT

The following table shows a sample of homes on the market, recording their size in square feet and their price in thousands of dollars (so for instance, the first home is selling for \$400,000).

Size (sq. ft.)	Selling Price (\$1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

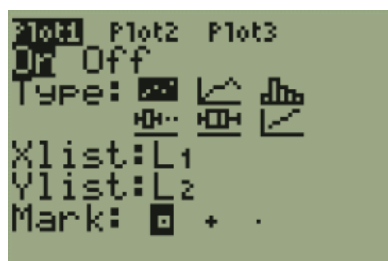
Construct a scatter plot for this data.

Using Your Calculator

On a TI graphing calculator, pressing **2ND** then **Y=** will open the statistics plots menu.



You can draw multiple plots at once, but there's no need to here. Pressing **ENTER** will open the first plot. By default, statistics plots are turned off, so if we want to use one, we have to turn it on first.

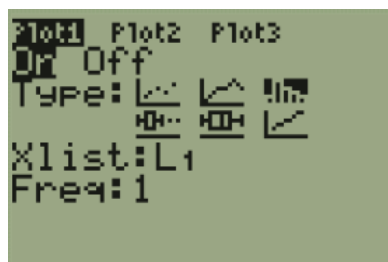


Notice that there are six options for the type of plot, but only two of them match ones we've discussed in this section: the scatterplot and the histogram. However, recall that dot plots, frequency distributions, and stem-and-leaf plots are all essentially equivalent to histograms, so we can use the histogram option to let the calculator handle the frequency counting for us, then we can draw the result however we like, as a table or graph.

HISTOGRAMS

We'll use the NBA players' ages as an example (before continuing, enter the data by pressing **STAT** **ENTER** to access the **Edit** menu).

Back in the statistics plot menu, select the histogram option (with the plot turned on).



Notice the options you're given: **Xlist** and **Freq**. The first option is asking for the list where you entered the data you want to graph. The second is a bit trickier; you have two options.

Option 1: Enter the raw data, writing every value in your data list into L1, and leave the frequency option as 1, meaning that every value you've listed appears once. In this case, the calculator will count the frequencies for you.

Option 2: If you have a frequency table already built, you can enter that in the calculator; write the values in L1 and the frequencies in L2. Then write L2 in the option for **Freq** in this menu (to write L2, press **2ND** then **2**).

Once you've entered the data and selected the appropriate options, pressing **GRAPH** will open the graphing window.

NOT SO FAST! If you press `GRAPH` now, you may see an empty graph. The reason is that the window is not necessarily set to the right parameters. Press the `WINDOW` button.

```

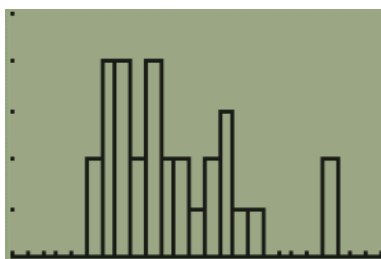
WINDOW
Xmin=15
Xmax=40
Xscl=1
Ymin=0
Ymax=5
Yscl=1
↓Xres=1

```

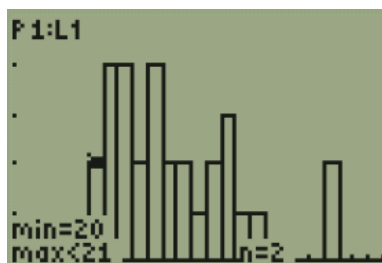
Most of the values are what you would expect: `Xmin` and `Xmax` define the lower and upper bounds of the horizontal axis, so we've set them to 15 and 40, respectively, to make sure that we can see all the results. The options for `Ymin` and `Ymax` are similar; it may take some trial and error to set the upper bound, since we don't know the max frequency before graphing.

The one that is not as obvious is that `Xscl` does more than set the distance between grid marks. For a histogram, this actually defines the **class width**, so if we wanted a grouped histogram, we would simply need to change this value.

In our example, though, we won't use any grouping, so we can now graph the histogram.



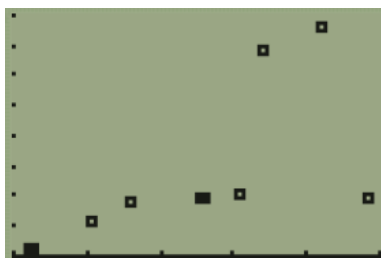
As it is, the histogram gives us the visual display of the data, but it's hard to read the actual frequencies. However, if you press the `TRACE` button, you can use the arrow keys to navigate from one bar to the next, seeing the frequency for each.



In this case, we've selected the bar "between" 20 and 21; really, this means anyone at 20 years old. Notice on the lower right the value of `n` is 2, which is the frequency for this category.

SCATTERPLOTS

Drawing a scatterplot on the calculator is simpler; we only have to enter the values of x in `L1` and y in `L2`, then adjust the window to ensure that all the points are visible. The scatterplot for the example comparing points per game and salary is shown below.

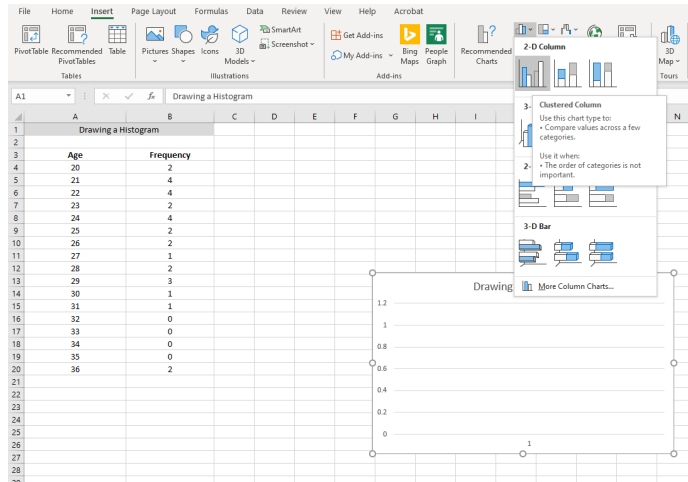


The low resolution on the calculator makes it hard to see, but there are a few overlapping points; the general trend, though, is visible.

Using Excel

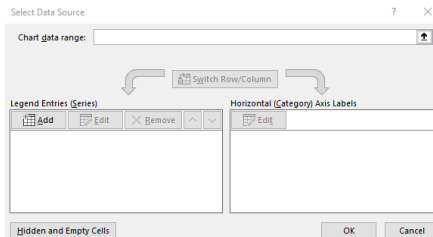
HISTOGRAMS

To draw a histogram in Excel, the data must be arranged in a frequency table to begin (there are ways to get Excel to build a frequency table from raw data, but these are too complicated to show here). For example, let's use the age data from the NBA dataset.

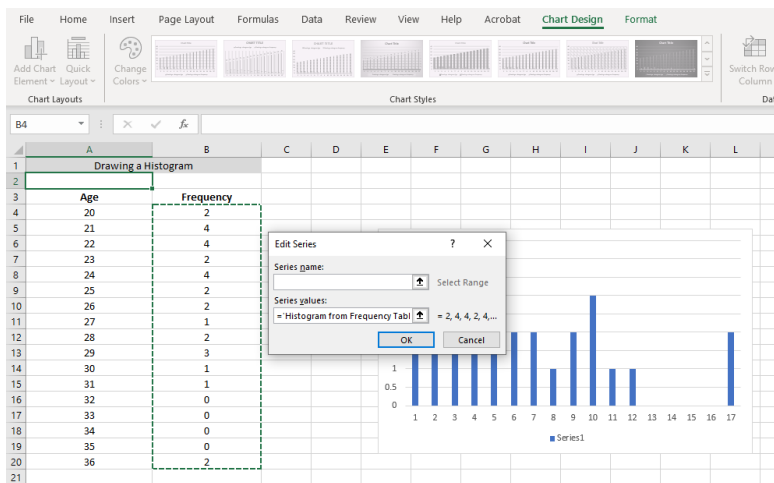


After entering the data, select the Insert tab along the top menu, and choose the first type of bar chart listed. When you do, Excel may try to guess what data you're trying to use, but it probably won't guess correctly.

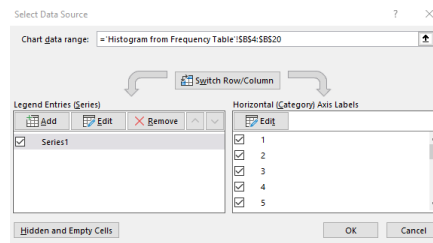
Right-click on the graph and choose the "Select Data..." option. If there are any data series automatically added, remove them.



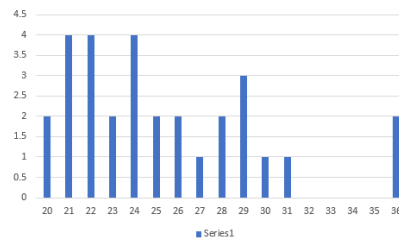
Now click on the Add button. The series name is optional, but delete the series values, and with that field selected, click and drag to select the *frequencies*.



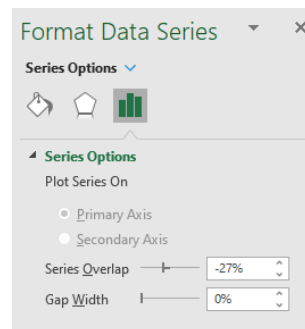
Notice that the graph so far shows the correct heights for the bars, but the horizontal axis labels are wrong. To fix this, we need to edit the "Horizontal (Category) Axis Labels" on the right of the Data Source menu.



When you select Edit, it will open a dialog that allows you to select the cells for the labels. Select the data in the age column.

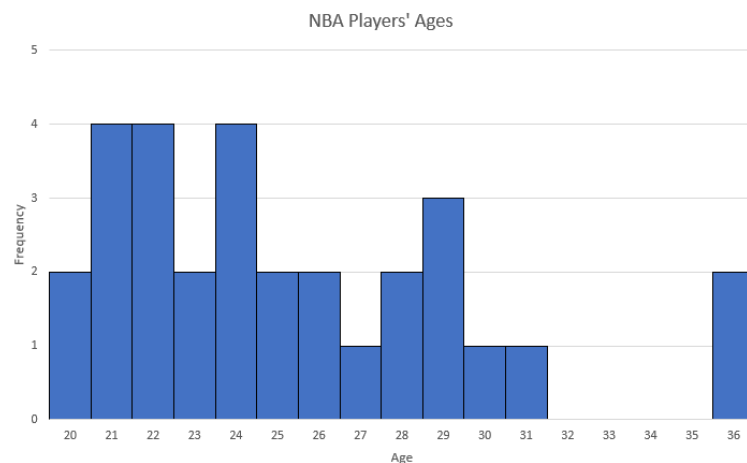


At this point, the graph is almost completed, but the bars are separated, and since this is a histogram (numerical data) instead of a bar chart (categorical data), we don't want a gap between them. To fix this, right-click on one of the bars and select the option labeled "Format Data Series." In the menu that opens to the right, find the option called "Gap Width," and change this to 0%.



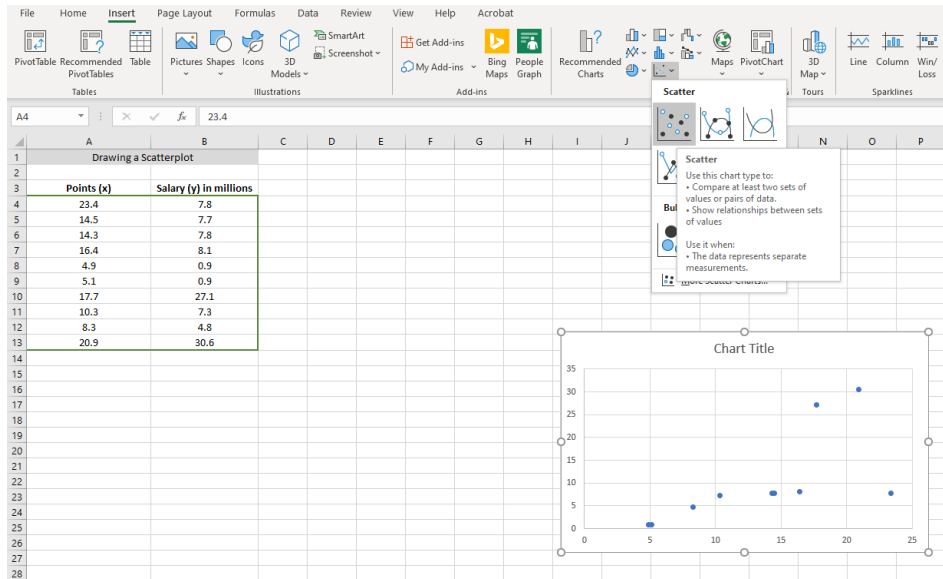
There are other small, optional options you can adjust, like adding axis and chart titles and removing the legend. Also, it can be helpful to add borders to the bars (in the same Format Data Series menu).

Here's the final histogram:

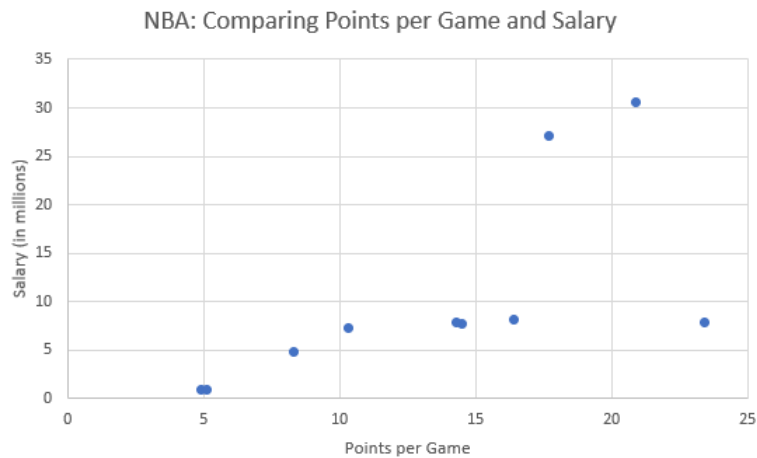


SCATTERPLOTS

Thankfully, it is much easier to create a scatterplot using Excel. With the data arranged in columns for x and y , simply select the data, and insert a chart; select the first option under “Scatter.”



After adding the chart and axis titles, here's the result:



Exercises 3.2

1. Nineteen immigrants to the U.S. were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

2, 5, 7, 2, 2, 10, 20, 15, 0,
7, 0, 20, 5, 12, 15, 12, 4, 5, 10

Draw a dot plot to summarize this data.

3. A store tracked how many iPads were sold each day for fifty days, and their data is below.

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

Construct a frequency table (including a relative frequency column) to describe this data.

2. A group of students earned the following final grades:

B, C, A, B, B, D, C, C, C, F, A, C, B, B, B, C, B, D

Draw a dot plot to summarize this data.

4. Twenty students were asked how many hours they worked per day. Their responses are as follows:

5 6 3 3
2 4 7 5
2 3 5 6
5 4 4 3
5 2 5 3

Construct a frequency table (including a relative frequency column) to describe this data.

5. Fifty part-time students were asked how many courses they were taking this semester. The (incomplete) results are shown below. Fill in the blank cells to complete the table.

Number of Courses	Frequency	Relative Frequency
1	30	0.6
2	15	
3		

6. A group of 20 students were polled and asked what year they belonged to, whether they were freshmen (FR), sophomores (SO), juniors (JR), or seniors (SR). The results are written below.

FR JR SO JR
SR FR SO SO
SO SR SO SR
SR FR SR SO
SR SO JR JR

Construct a frequency table (including a relative frequency column) to describe this data.

7. A group of 20 registered voters were polled and asked what party they belonged to, whether they were Republicans (R), Democrats (D), Green Party members (G), or independent (I). The results are written below.

R R D D
G D R D
I R R D
I D R I
R D R D

Construct a frequency table (including a relative frequency column) to describe this data.

8. The following is the average daily temperature for Frederick, Maryland for the month of June:

74, 60, 58, 58, 64, 67, 64, 74, 72, 70,
78, 80, 80, 79, 80, 80, 70, 83, 76, 78,
81, 78, 81, 70, 70, 71, 66, 66, 68, 74.

(a) Construct a grouped frequency and relative frequency distribution using a class width of 5, starting at 55.

(b) Construct a histogram from the frequency distribution.

9. A researcher gathered data on hours of video games played by school-aged children and young adults. She collected the following data:

0, 0, 1, 1, 1, 2, 2, 3, 3, 3,
4, 4, 4, 4, 5, 5, 5, 6, 6, 7,
7, 7, 8, 8, 8, 8, 8, 9, 9, 9,
10, 10, 11, 12, 12, 12, 12, 13.

(a) Construct a grouped frequency and relative frequency distribution using a class width of 2, starting at 0.

(b) Construct a histogram from the frequency distribution.

For exercises 10–13, use the frequency table below, which contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,356

10. What is the frequency of deaths measured from 2006 through 2009? 11. What percentage of deaths occurred after 2009 (from 2010 onwards)?
12. What is the relative frequency of deaths that occurred in 2003 or earlier? 13. What is the percentage of deaths that occurred in 2004?
14. What is wrong with the following grouped frequency distribution?

Grades	Frequency
50–55	2
55–60	4
60–70	9
70–80	15
80–90	7
90 and above	4

- (a) The classes do not all have the same width.
 (b) The classes overlap.
 (c) There are open-ended classes.
 (d) All of the above.
15. Draw a bar chart for the dataset in problem 6. 16. Draw a bar chart for the dataset in problem 7.
17. The scores for a math test are shown below, ordered from smallest to largest. 18. A basketball team's scores for the last 30 games are shown below, ordered from smallest to largest.

42 49 49 53 55 55
 61 63 67 68 68 69
 69 72 73 74 78 80
 83 88 88 88 90 92
 94 94 94 95 96 100

32 32 33 34 38 40
 42 42 43 44 46 47
 47 48 48 48 49 50
 50 51 52 52 52 53
 54 56 57 57 60 61

Build a stem-and-leaf plot for this data.

Build a stem-and-leaf plot for this data.

19. The following stem-and-leaf plots compare the ages of 30 actors and 30 actresses at the time they won the Oscar award for Best Actor or Actress.

Actors	Stems	Actresses
	2	146667
98753221	3	00113344455778
88776543322100	4	11129
6651	5	
210	6	011
6	7	4
	8	0

- (a) What is the age of the youngest actor to win an Oscar?
- (b) What is the age difference between the oldest and the youngest actress to win an Oscar?
- (c) What is the oldest age shared by two actors to win an Oscar?

20. The table below shows the yearly tuition of 8 universities, as well as the average mid-career salaries for graduates of each university.

University	Tuition (\$)	Salary (\$)
Princeton	28,540	137,000
Harvey Mudd	40,133	135,000
CalTech	39,900	127,000
MIT	42,050	118,000
Lehigh University	43,220	118,000
NYU-Poly	39,565	117,000
Babson College	40,400	117,000
Stanford	54,506	114,000

Draw a scatterplot for this data, using x to represent tuition and y to represent salary.

21. The table below shows the frequency of chirps for the striped ground cricket compared to the ambient temperature.

Chirps per Second	Temperature (°F)
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2
14.7	69.7
17.1	82.0
15.4	69.4

Draw a scatterplot for this data, using x to represent the chirping frequency and y to represent temperature.

SECTION 3.3 Describing Data with Statistics



We'll keep using the NBA dataset from the last section; here is the data again for reference:

Name	Team	No.	Age	Position	Height	Pts	Reb	Salary
Jayson Tatum	Celtics	0	22	PF	2.03 m	23.4	7.0	\$7,830,000
Joe Harris	Nets	12	28	SF	1.98 m	14.5	4.3	\$7,666,667
RJ Barrett	Knicks	9	20	SG	1.98 m	14.3	5.0	\$7,839,960
Ben Simmons	76ers	25	24	PG	2.08 m	16.4	7.8	\$8,113,930
Matt Thomas	Raptors	21	26	SG	1.93 m	4.9	1.5	\$898,310
Daniel Gafford	Bulls	12	21	PF	2.08 m	5.1	2.5	\$898,310
Andre Drummond	Cavaliers	3	27	C	2.08 m	17.7	15.2	\$27,093,019
Langston Galloway	Pistons	9	28	SG	1.85 m	10.3	2.3	\$7,333,333
Justin Holiday	Pacers	8	31	SF	1.98 m	8.3	3.3	\$4,767,000
Khris Middleton	Bucks	22	29	SF	2.01 m	20.9	6.2	\$30,603,448
Skal Labissiere	Hawks	7	24	PF	2.08 m	5.8	5.1	\$2,338,847
PJ Washington	Hornets	25	22	PF	2.01 m	12.2	5.4	\$3,831,840
KZ Okpala	Heat	4	21	SF	2.03 m	1.4	1.0	\$898,310
Wes Iwundu	Magic	25	25	SF	1.98 m	5.8	2.5	\$1,618,420
Rui Hachimura	Wizards	8	22	PF	2.03 m	13.5	6.1	\$4,469,160
Andrew Wiggins	Warriors	22	25	SF	2.01 m	21.8	5.1	\$27,504,630
Paul George	Clippers	13	30	SG	2.03 m	21.5	5.7	\$30,560,700
Avery Bradley	Lakers	11	29	PG	1.91 m	8.6	2.3	\$4,767,000
Jalen Lecque	Suns	0	20	PG	1.93 m	2.0	0.4	\$898,310
Harry Giles III	Kings	20	22	C	2.11 m	6.9	4.1	\$2,578,800
J.J. Barea	Mavericks	5	36	PG	1.78 m	7.7	1.8	\$1,620,564
Bruno Caboclo	Rockets	5	24	SF	2.06 m	3.0	2.0	\$1,845,301
Josh Jackson	Grizzlies	20	23	SG	2.03 m	9.0	3.1	\$7,059,480
JJ Redick	Pelicans	4	36	SG	1.91 m	15.3	2.5	\$13,486,300
Trey Lyles	Spurs	41	24	C	2.06 m	6.4	5.7	\$5,500,000
Troy Daniels	Nuggets	30	29	SG	1.93 m	4.3	1.1	\$384,541
Naz Reid	Timberwolves	11	21	C	2.06 m	9.0	4.1	\$898,310
Luguentz Dort	Thunder	5	21	SG	1.91 m	6.8	2.3	\$155,647
Jusuf Nurkic	Trail Blazers	27	26	C	2.13 m	17.6	10.3	\$13,125,000
Donovan Mitchell	Jazz	45	23	SG	1.85 m	24.0	4.4	\$3,625,760

Now, let's say a young player is getting ready to sign his first contract. Before he does, though, he wants to get a sense of what he can expect to earn, and he decides to use this sample of current players to do so. How would he go about doing this?

Measures of Center and Spread

In the last section, we saw how to visualize a full dataset and gain a snapshot understanding of it. Visuals are good, and they are generally used at the beginning of a study to get a thousand-foot view. However, if we want to dig deeper, we need to turn from visual summaries to numerical summaries, and that's what we'll discuss in this section.

In fact, when we use the word “statistic,” we're generally referring to one of these numbers that describe a dataset, like the average, for instance (the first one we'll cover). The statistics we'll use in this section can be broadly divided into two categories: **measures of center** and **measures of spread**.

Center and Spread

Measures of Center: these give us a sense of what a typical value in the dataset looks like.

Measures of Spread: these describe how much variety the dataset contains.

These terms on their own are not important to memorize; they simply help us categorize statistics so that we can compare the statistics in one category to each other.

For instance, in the example of the young player entering the league and evaluating player salaries, he may want to start by finding out what a typical salary is (and maybe digging deeper to find a typical *rookie* salary); for this, he would use measures of center. But then he may wonder how close most players are to this typical value—are salaries very spread out, so that he could be far above or below the center, or are they tightly clustered together, so that he can expect to make about the same as the typical player? For this, he would use measures of spread.

Here are the statistics we'll use:

- Measures of Center:
 1. Average or Mean
 2. Median
 3. Weighted Mean
 4. Mode
- Measures of Spread:
 1. Range
 2. Standard Deviation

We will also combine several of these into what is called the *Five Number Summary* and draw a plot called a *boxplot* to illustrate it.

Note on Population and Sample Recall from the beginning of the chapter that one of the most fundamental concepts in statistics is the use of a sample to estimate the truth about a population (like we're doing by using this sample of NBA players to describe all players in the league). Thus, we can talk about, for instance, the *sample average* and the *population average*. For most of the statistics in this section, there will be no difference in how these are calculated, but there is a slight difference in the *standard deviation*. For the sake of simplicity, we will simply use the sample version (the difference is very small).

Note on Calculations For each of these statistics, we'll show the formula and calculate a few examples manually. However, a graphing calculator has the formulas built in, so after calculating a few the long way, we'll show how to get the answers more quickly using the calculator.

Average or Mean

The mean or average is probably the most familiar of these statistics; it is common measure used to get a sense of the center of a dataset.

To calculate the mean, add up all the values in the set and divide by the number of values. To write this as a formula, we'll use a new notation:

$$\sum x$$

The symbol \sum is the Greek letter *sigma*, and it means to add up whatever follows, so $\sum x$ means “the sum of x ,” so we'll use that to mean “sum all the values in the dataset.”

Mean

The mean of a sample of size n is

$$\bar{x} = \frac{\sum x}{n}$$

The notation for the mean, \bar{x} , is read as “ x bar.” This stands for the mean of a *sample*; the mean of a *population* is written with the Greek letter *mu*, μ , but we won’t worry about that. We’ll assume all the datasets in this section are samples.

AVERAGE NBA SALARY

EXAMPLE 1

Find the average salary of the players listed in the NBA dataset.

To calculate the average, simply add the 30 salaries together, then divide the result by 30:

$$\begin{aligned}\bar{x} &= \frac{7,830,000 + 7,666,667 + 7,839,960 + \cdots + 13,125,000 + 3,625,760}{30} \\ &= \frac{230,210,897}{30} \\ &= 7,673,697\end{aligned}$$

Solution

The average salary of this sample of players is \$7,673,697.

According to Basketball Reference, the average salary for all NBA players for the same season is \$7.7 million. Notice how the sample average is almost the same; this is the advantage of taking a sample. We were able to calculate the average of a small group of 30 randomly chosen players, and although the sample average is not *exactly* the same as the population average, it is an excellent estimate for the population value.

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Calculate the mean.

TRY IT

Median

We have one measure of center; isn’t that enough? Why do we need another? Why not just always use the mean?

Look back at the NBA dataset carefully. We found that the average salary is around \$7.7 million; is this really a typical salary? In other words, could our young player truly expect to make that much?

If you notice, only 11 players (about a third of the group) make more than \$7 million a year. The average is often not the best measure of a *typical* value; notice that a few players make over \$20 million, which artificially inflates the average.

Think of it this way; there is a minimum salary possible (\$0, or really whatever the league minimum is defined to be), but there is practically no maximum that someone could be paid. It turns out that this means that most players are clustered on the lower end (relatively speaking), and a few very highly-paid players increase the average value.

Let’s take a simpler example; suppose there are 5 people with the following salaries:

\$30,000 \$45,000 \$50,000 \$52,000 \$1,000,000

Calculating the mean for this data set yields \$235,400. Does the mean give a true picture for the center of the data? Can we rightfully say the average person in this data set earns roughly \$235,000? Obviously, the answer is no, since 80%—or 4/5—of the people in this group earn less than \$53,000. When we have an **outlier**—a number far removed from the majority of data values—in our data, the mean is *skewed*.

How else can we measure the center? Well, we can simply look for the *middle* value, when the data is arranged in order.

Median

The median of a sample, denoted with a capital M , is the value in the middle of the list when the data is ordered from smallest to largest (or vice versa).

If there are an odd number of values, the median is the middle one. If there are an even number of values, the median is the average of the two middle ones.

EXAMPLE 2 MEDIAN NBA SALARY

Find the median of the salaries listed in the NBA dataset.

Solution

First, we need to arrange the values in order:

\$155,647	\$384,541	\$898,310	\$898,310	\$898,310
\$898,310	\$898,310	\$1,618,420	\$1,620,564	\$1,845,301
\$2,338,847	\$2,578,800	\$3,625,760	\$3,831,840	\$4,469,160
\$4,767,000	\$4,767,000	\$5,500,000	\$7,059,480	\$7,333,333
\$7,666,667	\$7,830,000	\$7,839,960	\$8,113,930	\$13,125,000
\$13,486,300	\$27,093,019	\$27,504,630	\$30,560,700	\$30,603,448

Since there are an even number of salaries, we need to select the middle two (at the end of the third row and the beginning of the fourth row), and take the average of these two.

\$155,647	\$384,541	\$898,310	\$898,310	\$898,310
\$898,310	\$898,310	\$1,618,420	\$1,620,564	\$1,845,301
\$2,338,847	\$2,578,800	\$3,625,760	\$3,831,840	\$4,469,160
\$4,767,000	\$4,767,000	\$5,500,000	\$7,059,480	\$7,333,333
\$7,666,667	\$7,830,000	\$7,839,960	\$8,113,930	\$13,125,000
\$13,486,300	\$27,093,019	\$27,504,630	\$30,560,700	\$30,603,448

Find the average of \$4,469,160 and \$4,767,000:

$$\begin{aligned} M &= \frac{\$4,469,160 + \$4,767,000}{2} \\ &= \boxed{\$4,618,080} \end{aligned}$$

Notice that this is a more typical salary within the group; exactly half of the players make more and half make less.

TRY IT

Find the median of the following data, representing the number of months an AIDS patient lives after taking a drug:

3	4	8	8	10	11	12	13	14	15
15	16	16	17	17	18	21	22	22	24
24	25	26	26	27	27	29	29	31	32
33	33	34	34	35	37	40	44	44	47

Finding the median manually means having to sort the data, which can be tedious. However, once we have it sorted, we simply have to find the center. Often this is easy to do, especially when the data is organized in rows, but there's another way to find the center of a list of n values; the center position, since it is midway between 1 and n , turns out to be the average of 1 and n .

Where's the Median

To find the median, we want to be able to quickly figure out what position to count to in the ordered data set. To do this, calculate

$$\frac{n+1}{2}$$

where n is the size of the data set.

- If n is odd, $\frac{n+1}{2}$ is a whole number. Count to that position and there you'll find the median.
- If n is even, $\frac{n+1}{2}$ is halfway between two whole numbers. Find the average of the values at those two positions.

Outliers: Let's revisit our small income data set:

\$30,000 \$45,000 \$50,000 \$52,000 \$1,000,000

The median is \$50,000, which is a more accurate measure of center than the mean, which is \$235,400. This illustrates how the mean is *sensitive* to outliers, whereas the median is *resistant* to outliers.

Mean and Median

When outliers are present, the median is a better measure of center. When outliers are absent, the mean can be used.

When trying to decide whether to use the mean or the median as the measure of the center of a data set, compare them to decide if they are drastically different. Of course, the best policy is to report both of them and compare them to determine whether the data set is symmetric or *skewed* (containing outliers).

Calculating Mean and Median from a Frequency Table

What if we're given a frequency table instead of raw data? For instance, suppose we want to calculate the average age of the players in our dataset, and we're given the frequency distribution (which we built in the last section):

Age	Frequency
20	2
21	4
22	4
23	2
24	4
25	2
26	2
27	1
28	2
29	3
30	1
31	1
36	2

(Notice that in this case, we removed the rows with frequencies of 0, since our focus at the moment is not on visualizing the data spread.)

We can, of course, simply reconstruct the dataset from the frequency table by listing two 20's, four 21's, and so on:

20, 20, 21, 21, 21, 21, 22, ...

At that point, we could calculate the mean and median as before.

However, if we have the frequency table, we can use a shortcut to calculate each answer.

MEAN

For the mean, notice that if we want to add up all the values

$$20 + 20 + 21 + 21 + 21 + 21 + 22 + \cdots$$

that's the same as multiplying 20 by 2, multiplying 21 by 4, and so on, and adding all these results.

Thus, we can simply multiply each value by its frequency and add these answers; that will give us the total sum more quickly then adding them one at a time. At that point, we can divide by the size of the dataset (30) to get the average.

$$\begin{aligned}\bar{x} &= \frac{20(2) + 21(4) + 22(4) + 23(2) + \cdots + 31(1) + 36(2)}{30} \\ &= 25.3\end{aligned}$$

MEDIAN

To find the center of the dataset, remember that its location can be found using

$$\frac{n+1}{2}.$$

Since we have 30 values, the center will be at

$$\frac{30+1}{2} = 15.5.$$

Since this is not a whole number, it means that the center will be between the 15th and 16th values.

Notice that the frequency table has the ages listed in order, and the frequencies help us count through the set. Simply add up the frequencies as you go: the first two positions hold 20's, the next four positions hold 21's, and so on, so we simply need to add up frequencies until we get to 15 and 16:

Age	Frequency	
20	2	2
21	4	2 + 4 = 6
22	4	6 + 4 = 10
23	2	10 + 2 = 12
24	4	12 + 4 = 16
25	2	
26	2	
27	1	
28	2	
29	3	
30	1	
31	1	
36	2	

Notice that both the 15th and 16th positions contain the value 24, so the median will be 24.

Note: this process for finding the mean and median from a frequency table doesn't work with a *grouped* frequency table, although a similar process can be used to *estimate* the mean and median in that case.

Weighted Mean

So far, in calculating the mean, we've assumed that all the values are equally significant. However, there are cases in which we are given *weights* for each value.

The most familiar example is probably that of calculating grades. For instance, suppose that a syllabus lists the following weights:

Assignment	Weight
Test 1	20%
Test 2	20%
Test 3	20%
Homework	15%
Project	10%
Final Exam	15%

Instead of listing percentages, of course, these assignments could also be given using a points system. The simplest way to do this would be to simply define a total of 100 points; in that case, the 20% allocated to each test would correspond to 20 points, and so on. To make grading easier, many people would instead use 1000 points, as shown:

Assignment	Points
Test 1	200
Test 2	200
Test 3	200
Homework	150
Project	100
Final Exam	150

In either case, once each assignment is graded, we can fill in a table like this:

Assignment	Score	Points
Test 1	85%	200
Test 2	92%	200
Test 3	87%	200
Homework	95%	150
Project	92%	100
Final Exam	91%	150

Notice how this looks more or less like a frequency table; the score on Test 1 accounts for 200 of the points available, so that test earned $(0.85)(200) = 170$ points. If we continue down the list, the final score is found by multiplying each percentage earned by the total number of points available, then dividing the final answer by 1000.

WEIGHTED AVERAGE

EXAMPLE 3

Find the final score of the student whose grades are listed below, using both the points system and the percentage system for defining weights.

Assignment	Score	Weight	Points
Test 1	85%	20%	200
Test 2	92%	20%	200
Test 3	87%	20%	200
Homework	95%	15%	150
Project	92%	10%	100
Final Exam	91%	15%	150

Using the points given, we can multiply each score by the available points and add these up:

$$\begin{aligned}
 \text{Points earned} &= (0.85)(200) + (0.92)(200) + (0.87)(200) + (0.95)(150) \\
 &\quad + (0.92)(100) + (0.91)(150) \\
 &= 899
 \end{aligned}$$

The student's final score, then, is the percentage of available points that they earned:

$$\text{Final Grade} = \frac{899}{1000} = 0.899 = \boxed{89.9\%}$$

Solution

Now, let's do the same thing using the percentage weights instead of possible points. Here, notice that the percentages are simply the number of points for each category divided by 1000. In other words, the last step of the calculation (dividing by 1000) is already done for us, so we simply need to multiply the percentage earned by the percentage weight:

$$\begin{aligned}\text{Final Grade} &= (0.85)(0.2) + (0.92)(0.2) + (0.87)(0.2) + (0.95)(0.15) \\ &\quad + (0.92)(0.1) + (0.91)(0.15) \\ &= 0.899 \\ &= \boxed{89.9\%}\end{aligned}$$

Mode

There is another, less used, measure of center: the mode.

Mode

The mode is the most frequently occurring value in a dataset.

There can be more than one mode in a data set as long as those values have the same frequency and the frequency is the highest. A data set with two modes is called **bimodal**.

EXAMPLE 4 MODE

Find the mode of the dataset summarized below, the ages of players in the NBA dataset.

Age	Frequency
20	2
21	4
22	4
23	2
24	4
25	2
26	2
27	1
28	2
29	3
30	1
31	1
36	2

Solution Since the data is already summarized with a frequency table, all we have to do is scan through and find the highest frequency (4) and check the ages that appear with that frequency. Thus, the modes (this dataset has three) are

$$\boxed{21, 22, 24}$$

TRY IT

The number of books checked out from the library from 25 students are as follows:

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

Find the mode.

The mode is not nearly as useful as the other two measures of center (mean and median), so we won't spend as much time on it as on the others.

Now we're ready to turn from measures of center to measures of spread, which we'll use to observe how much variety a dataset contains.

Range

The **range** is the simplest measure of spread. It is simply the distance between the smallest and largest data value in the set, which can be calculated by subtracting them.

Range

The range is calculated by finding the difference between the minimum and maximum value in a dataset:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

RANGE OF NBA PLAYERS' HEIGHTS

Find the range of heights for the NBA players listed in the dataset at the beginning of the chapter.

All we have to select from the list is the smallest value and the largest value. Scanning the list reveals that the minimum is 1.78 meters and the maximum is 2.13 meters. The range, then, is simply the difference between these two:

$$\begin{aligned}\text{Range} &= 2.13 - 1.78 \\ &= \boxed{0.35 \text{ m}}\end{aligned}$$

This relatively small range means that there is not much variation in the heights of the players in our sample.

How do we distinguish a small range from a large one? The key is to compare the range to a typical value. For instance (and this isn't the only way to do this), if we calculate the range as a percentage of the mean (the mean height is 1.99 m), we get

$$\frac{0.35}{1.99} = 18\%$$

On the other hand, the range of the points scored is much larger compared to its mean. The range of that list is 22.6, while the mean is 11.28, so the range is approximately 200% as large as the mean. By comparison, there is clearly much more variation within the points column than the players' heights.

Standard Deviation

The **standard deviation** is a number that measures how far a typical data value is from the mean. The standard deviation is always positive or zero. A small standard deviation means less spread in the data; a large standard deviation means more spread in the data.

First of all, the **deviation** of each data point x is its difference from the mean \bar{x} :

$$x - \bar{x}.$$

Each value in the data set has a deviation associated with it. If we want to find how far, on average, each data point is from the mean, it would make sense to take the average of the deviations. However, there's a problem with doing that: if we add up the deviations, we'll always get 0, because of the way that \bar{x} is calculated. Some of the deviations are positive, some are negative, and the positives cancel out the negatives.

To get around this, we square the deviations so that everything becomes positive. NOW, when we take an average² of these *squared* deviations, we get a meaningful number, instead of getting 0 every time. We call this "average" the variance:

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

²almost; see margin note

EXAMPLE 5

Solution

It is not quite an average, since we divide by $n - 1$ instead of n . The reasons for this are complicated, but they have to do with making the sample variance be what is called an unbiased estimator for the population variance.

The only problem now is that we've got an average of these squared things, so the units of our answer are not the same units we started with. In other words, if the data is given in units of inches, we've got a variance in square inches. To get an answer, we take the square root of the variance, and that's what we call the standard deviation.

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

To calculate the standard deviation,

1. Calculate the mean of the data values, \bar{x} .
2. Subtract the mean from each data value to find the deviations:

$$\text{deviation} = x - \bar{x}$$

3. Square each deviation: $(x - \bar{x})^2$
4. Take the sum of the squared deviations: $\Sigma(x - \bar{x})^2$
5. Divide that sum by n minus 1, where n is the number of data values: $\frac{\Sigma(x - \bar{x})^2}{n - 1}$
6. Take the square root of this quotient: $\sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

Standard Deviation

The standard deviation of a sample is given by

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

where n stands for the number of data values and x stands for each data value.

Don't worry; we'll use this formula once to calculate the standard deviation, and then let the calculator handle the tedious calculation for us.

EXAMPLE 6 AT THE MALL

Kari went shopping and bought five things. The prices are as follows:

\$20 \$4 \$15 \$9 \$3

Calculate the standard deviation for this data.

Solution

$$\bar{x} = (20 + 4 + 15 + 9 + 3)/5 = 10.20$$

The mean is \$10.20. We can use the table below to get the standard deviation.

Data Value x	Deviations $(x - \bar{x})$	Deviations ² $(x - \bar{x})^2$
20	$20 - 10.20 = 9.8$	$(9.8)^2 = 96.04$
4	$4 - 10.20 = -6.2$	$(-6.2)^2 = 38.44$
15	$15 - 10.20 = 4.8$	$(4.8)^2 = 23.04$
9	$9 - 10.20 = -1.2$	$(-1.2)^2 = 1.44$
3	$3 - 10.20 = -7.2$	$(-7.2)^2 = 51.84$

Adding up all the values in the third column yields 210.8. The variance, s^2 , is equal to this sum divided by the total number of data values minus one.

$$s^2 = \frac{210.8}{5 - 1} = \$52.7.$$

The standard deviation s is equal to the square root of the variance.

$$s = \sqrt{52.7} = \boxed{\$7.26}$$

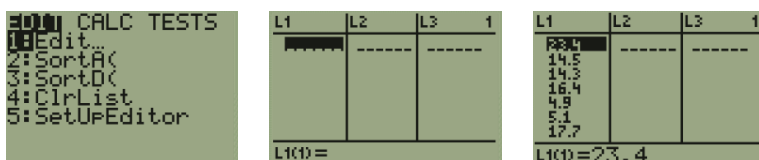
Using Your Calculator

For each of the statistics in this section, we've shown how to calculate them manually. Just as a reminder, we now know how to calculate each of the following for a dataset:

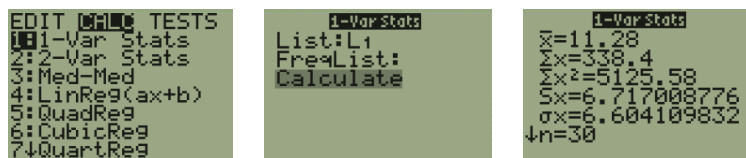
- Measures of Center:
 1. Average or Mean
 2. Median
 3. Weighted Mean
 4. Mode
- Measures of Spread:
 1. Range
 2. Standard Deviation

Most of these (with the exception of the mode and the range) can be found directly in a list of statistics that a TI calculator calls **1-Var Stats**. This list can be generated for a dataset entered in the calculator. The range can also be calculated indirectly from this list, but we'll have to stick to the manual method for finding the mode (thankfully, this is pretty simple).

To begin, enter a dataset as before; press the **STAT** button, then **ENTER** to enter the **Edit** menu. We'll use the points per game listed in the NBA dataset to illustrate the process.



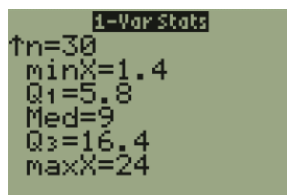
Next, press the **STAT** button again, and use the right arrow key to navigate over to the **CALC** menu. Select the first option, labeled 1: **1-Var Stats**. Since we entered the data in L1, we can leave the default options alone and simply select **Calculate**.



This list contains many statistics; notice that the first one is labeled \bar{x} ; this, of course, is the mean. The next is $\sum x$, or the sum of the values, and the one after that is the sum of the squares (we haven't used either of these values directly).

After that, notice the value of S_x ; this is the standard deviation for the sample. The next value is the population standard deviation, which we can ignore (if you take a statistics course, you can worry about the difference then).

Scrolling down using the arrow keys reveals a few more statistics, which we will investigate a bit more fully at the end of this section, when we discuss the Five Number Summary:



For now, notice that **minX** and **maxX**, the minimum and maximum values, can be used to calculate the range. Also, the median is listed, labeled **Med**.

FREQUENCY TABLES AND STATISTICS

Remember how we calculated the mean and median of a dataset using a frequency table earlier (and a weighted mean uses the same process)? We can also do this using the calculator; let's illustrate using the age data for the NBA players; refer to the earlier discussion of mean and median with a frequency table to see the distribution.

Now, when we enter the data, we'll enter the ages in L1 and the frequencies in L2, and this time, in the 1-Var Stats menu, change the FreqList option to L2 (press **2ND** **2** to write L2):

L1	L2	L3	Z
0	1	---	
1	4		
2	1		
3	1		
4	1		
5	1		
6	1		
7	1		
8	1		
9	1		
L2(1)=2			

1-Var Stats
List:L1
FreqList:L2
Calculate

1-Var Stats
 $\bar{x}=25.3$
 $\Sigma x=759$
 $\Sigma x^2=19737$
 $Sx=4.292334788$
 $\sigma x=4.220189569$
 $n=30$

The mean is 25.3, the same value we calculated manually, and the other values can be found in the list as well.

Using Excel

All of these statistics can be calculated using built-in Excel formulas; for each, enter the formula name, followed by open parentheses, then select the cells you want to use for the calculation, and close the parentheses.

Here are the formulas:

Statistic	Formula
-----------	---------

Mean	AVERAGE(cells)
------	----------------

Median	MEDIAN(cells)
--------	---------------

Mode	MODE.MULT(cells)
------	------------------

The MULT part means that this will return multiple values; results will spill over into the cells below the one with the formula.

Range	MAX(cells) - MIN(cells)
-------	-------------------------

As with the calculator, the range must be calculated indirectly.

Standard Deviation	STDEV.S(cells)
--------------------	----------------

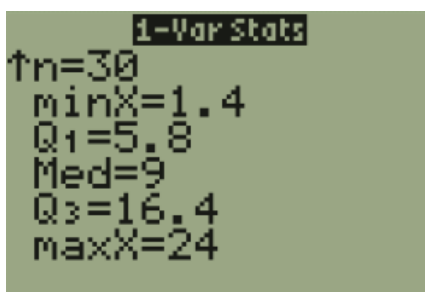
The S refers to the fact that this is the *sample* standard deviation; for the population standard deviation, use STDEV.P.

For example, the following spreadsheet shows the results for all of the numerical variables in the NBA dataset:

C32	=AVERAGE(C2:C31)								
	A	B	C	D	E	F	G	H	I
1	Name	Team	Number	Age	Position	Height	PPG	RB	Salary
2	Luguentz Dort	Thunder	5	21	SG	1.91	6.8	2.3	\$ 155,647.00
3	Troy Daniels	Nuggets	30	29	SG	1.93	4.3	1.1	\$ 384,541.00
4	Matt Thomas	Raptors	21	26	SG	1.93	4.9	1.5	\$ 898,310.00
5	Daniel Gafford	Bulls	12	21	PF	2.08	5.1	2.5	\$ 898,310.00
6	KZ Okpala	Heat	4	21	SF	2.03	1.4	1	\$ 898,310.00
7	Jalen Lecque	Suns	0	20	PG	1.93	2	0.4	\$ 898,310.00
8	Naz Reid	Timberwolves	11	21	C	2.06	9	4.1	\$ 898,310.00
9	Wes Iwundu	Magic	25	25	SF	1.98	5.8	2.5	\$ 1,618,420.00
10	J.J. Barea	Mavericks	5	36	PG	1.78	7.7	1.8	\$ 1,620,564.00
11	Bruno Caboclo	Rockets	5	24	SF	2.06	3	2	\$ 1,845,301.00
12	Skal Labissiere	Hawks	7	24	PF	2.08	5.8	5.1	\$ 2,338,847.00
13	Harry Giles III	Kings	20	22	C	2.11	6.9	4.1	\$ 2,578,800.00
14	Donovan Mitchell	Jazz	45	23	SG	1.85	24	4.4	\$ 3,625,760.00
15	PJ Washington	Hornets	25	22	PF	2.01	12.2	5.4	\$ 3,831,840.00
16	Rui Hachimura	Wizards	8	22	PF	2.03	13.5	6.1	\$ 4,469,160.00
17	Justin Holiday	Pacers	8	31	SF	1.98	8.3	3.3	\$ 4,767,000.00
18	Avery Bradley	Lakers	11	29	PG	1.91	8.6	2.3	\$ 4,767,000.00
19	Trey Lyles	Spurs	41	24	C	2.06	6.4	5.7	\$ 5,500,000.00
20	Josh Jackson	Grizzlies	20	23	SG	2.03	9	3.1	\$ 7,059,480.00
21	Langston Galloway	Pistons	9	28	SG	1.85	10.3	2.3	\$ 7,333,333.00
22	Joe Harris	Nets	12	28	SF	1.98	14.5	4.3	\$ 7,666,667.00
23	Jayson Tatum	Celtics	0	22	PF	2.03	23.4	7	\$ 7,830,000.00
24	RJ Barrett	Knicks	9	20	SG	1.98	14.3	5	\$ 7,839,960.00
25	Ben Simmons	76ers	25	24	PG	2.08	16.4	7.8	\$ 8,113,930.00
26	Jusuf Nurkic	Trail Blazers	27	26	C	2.13	17.6	10.3	\$ 13,125,000.00
27	JJ Redick	Pelicans	4	36	SG	1.91	15.3	2.5	\$ 13,486,300.00
28	Andre Drummond	Cavaliers	3	27	C	2.08	17.7	15.2	\$ 27,093,019.00
29	Andrew Wiggins	Warriors	22	25	SF	2.01	21.8	5.1	\$ 27,504,630.00
30	Paul George	Clippers	13	30	SG	2.03	21.5	5.7	\$ 30,560,700.00
31	Khris Middleton	Bucks	22	29	SF	2.01	20.9	6.2	\$ 30,603,448.00
32	Average		14.97	25.30		1.99	11.28	4.34	\$ 7,673,696.57
33	Median		11.50	24.00		2.01	9.00	4.10	\$ 4,618,080.00
34	Range		45.00	16.00		0.35	22.60	14.80	\$ 30,447,801.00
35	Standard Deviation		11.53	4.29		0.08	6.72	3.04	\$ 9,187,926.33
36	Mode		5	21		2.03	9	2.3	\$ 898,310.00
37			25	24			5.8	2.5	
38				22					

Five Number Summary and Boxplot

Let's go back to the bottom of the 1-Var Stats results from the calculator:



We've already discussed the median, and how it splits the data in half. In other words, half of the players averaged fewer than 9 points per game, and half scored above that. So between the minimum of 1.4 and the median, you'll find half of the players, and between the median and the maximum of 24, you'll find the other half.

What about those other two values (Q_1 and Q_3)? These, it turns out, continue the process of division in half. Specifically, Q_1 divides the lower half of players in half again (it is the median of the lower half) and Q_3 divides the upper half in half again.

Thus, these five numbers (the minimum, Q_1 , the median, Q_3 , and the maximum) split the data into quarters, and a quarter of the players fall into each of these divisions. This is where the Q notation comes from: Q_1 is called the **first quartile** and Q_3 is called the **third quartile**. The median is sometimes called the **second quartile**, and it can be denoted by Q_2 (instead of the M we used earlier).

Five Number Summary

The Five Number Summary lists the following values:

- Minimum
- First Quartile
- Median
- Third Quartile
- Maximum

These values divide the dataset into four divisions, and a quarter of the data values fall into each division.

The value in these five statistics (which together are called the *Five Number Summary*) is that they can give us a sense of where the data is clustered or spread out.

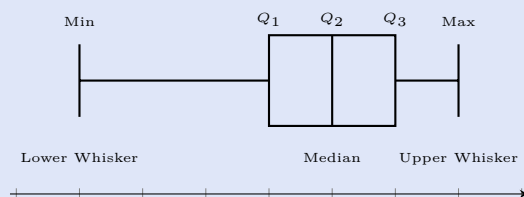
For instance, the first quarter of the data in the example above falls between 1.4 and 5.8 (a range of 4.4), the second quarter is between 5.8 and 9 (a range of 3.2), the third between 9 and 16.4 (7.4), and the last between 16.4 and 24 (7.6).

Notice how the ranges are tighter on the lower end; this indicates grouping on the low side, since values are packed in more closely. On the upper end, we have to cover a spread of 7 or 8 points to find the same number of players that 3 or 4 points covered on the lower end.

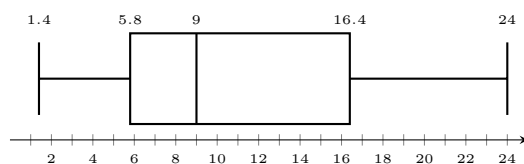
Since it can be hard to interpret these results when they're given as a list of number like this, we can draw a graph to visualize the Five Number Summary, called a *boxplot*.

Boxplot

A boxplot is a visual representation of a Five Number Summary. It consists of vertical lines marking each of the five values. The middle three (from the first quartile to the third quartile) are joined with a box, giving the plot its name. Lines are drawn from this box out to the minimum and maximum; these lines are called *whiskers*, so that these plots are sometimes called box-and-whisker plots.

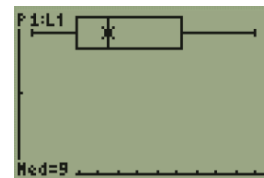
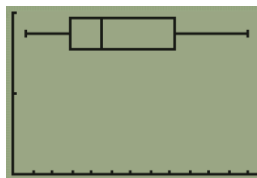
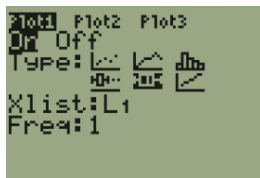


For instance, the boxplot for the players' scoring would look like the following:



Notice how the lower side is more scrunched together, and the upper side is more spread out; this is the same thing we observed from the values themselves, but it's easier to see on a picture.

The graphing calculator can also draw these; if you open the **STAT PLOT** menu and select one of the plots, you can see the option for a boxplot. Notice that there are two types; one separates *outliers* from the data, which we haven't discussed here, so we'll use the second version (without the dots). After adjusting the window, we can view the plot (note that to see the values of the bars, we can use the **TRACE** button).



Exercises 3.3

For problems 1–18, use the dataset shown in the table below. This is a sample of 15 players in Major League Baseball, chosen from the starting lineups of teams in 2019. The table shows the team, age, position, height, and salary for each player, as well as several statistics from that season. These include the number of games they played (G), their batting average (AVE) (the proportion of their at-bats for which they got a hit), and their home runs (HR).

Name	Team	Age	Height	G	AVE	HR	Salary
Cedric Mullins	Orioles	25	173 cm	22	.094	0	\$557,500
Tim Anderson	White Sox	26	185 cm	123	.335	18	\$1,400,000
Christin Stewart	Tigers	25	183 cm	104	.233	10	\$556,400
Alex Gordon	Royals	35	185 cm	150	.266	13	\$20,000,000
Jonathan Schoop	Twins	27	185 cm	121	.256	23	\$7,500,000
Marcus Semien	Athletics	29	183 cm	162	.285	33	\$5,900,000
Yandy Diaz	Rays	28	188 cm	79	.267	14	\$558,400
Randal Grichuk	Blue Jays	28	188 cm	151	.232	31	\$5,000,000
Josh Donaldson	Braves	33	185 cm	155	.259	37	\$23,000,000
Joey Votto	Reds	36	188 cm	142	.261	15	\$25,000,000
Cody Bellinger	Dodgers	24	193 cm	156	.305	47	\$605,000
Ryan Braun	Brewers	35	188 cm	144	.285	22	\$19,000,000
Maikel Franco	Phillies	27	185 cm	123	.234	17	\$5,200,000
Ian Kinsler	Padres	37	183 cm	87	.217	9	\$3,750,000
Marcell Ozuna	Cardinals	28	185 cm	130	.241	29	\$12,250,000

- Find the mean age of the players.
- Find the mean height of the players.
- Find the mean salary for the players.
- Find the median number of games played (G) for the players.
- Find the median number of home runs (HR) for the players.
- Find the median batting average (AVE) for the players.
- Find the mode(s) for the height of the players.
- Find the mode(s) for the ages of the players.
- For the players' ages, which is greater, the mean or the median?
- For the players' salaries, which is greater, the mean or the median?
- Find the range of the ages of the players.
- Find the range of the heights of the players.
- Find the range of the number of games played by the players.
- Find the standard deviation for the players' salaries.
- Find the standard deviation for the number of home runs hit by the players.
- Find the standard deviation for the players' batting averages.
- Calculate the Five Number Summary for the players' heights.
- Calculate the Five Number Summary for the number of games played by the players.
- Twenty-five randomly selected students were asked how many movies they had watched the previous week. The results are shown below.
- Twenty students were asked how many hours they worked per day. The table below shows the results.

Number of Movies	Frequency
0	5
1	9
2	6
3	4
4	1

Find the mean and median for the number of movies watched per student.

Number of Hours	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

Find the mean and median for the number of hours worked per student.

21. Find the weighted average of the student's grades listed below, using the given percentage value for each category.

Assignment	Score	Weight
Test 1	81	30%
Test 2	88	30%
Homework	92	10%
Quizzes	89	10%
Final Exam	84	20%

23. In a neighborhood donut shop, one type of donut has 530 calories, three types of donuts have 330 calories, four types of donuts have 320 calories, seven types of donuts have 410 calories, and five types of donuts have 380 calories. Find the mean and median calories of the donuts.

22. Find the weighted average of the student's grades listed below, using the given point value for each category.

Assignment	Score	Points
Tests	86	200
Projects	94	100
Homework	90	50
Final Exam	83	150

24. In a recent issue of *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted 2 days, 36 lasted 3 days, 18 lasted 4 days, 19 lasted 5 days, 4 lasted 6 days, 1 lasted 7 days, 1 lasted 8 days, and 1 lasted 9 days. Find the mean and median length (in days) of an engineering conference.

SECTION 3.4 Linear Regression



How can we predict the value of a home? If someone wants to sell their house, they have to pick an asking price; where does that come from?

The value of a home depends on many factors: location, size, amount of land, when it was built; the list goes on and on. However, what if we isolate just *one* of these factors and try to see how it impacts the price?

Let's say we decide to focus on size, in square feet. What we need, then, is a list of homes for which we can find the size and the price, so that we can compare them. Here's a small (fictional) dataset:

Size (sq. ft.)	Selling Price (\$1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Notice that the price is listed in thousands of dollars, so the price of the first home in the list is \$400,000.

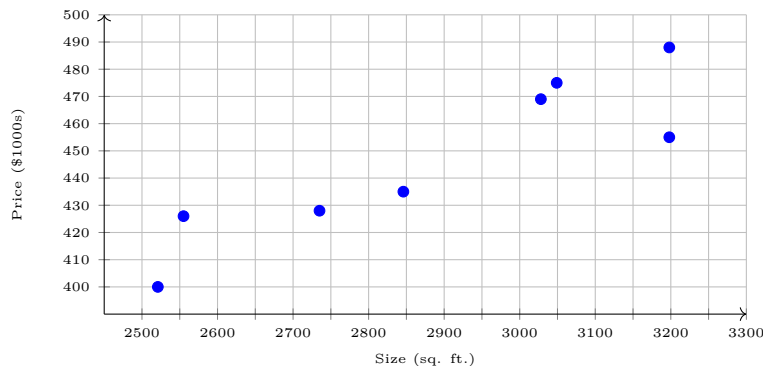
You may be able to scan down the list and recognize that as the size increases, the price increases as well; this seems natural enough. However, we'd like to be more specific about this relationship—that's what we'll learn how to do in this section.

When we're investigating a relationship between two variables—size and price here—it's always good to start with a scatterplot so that we can visualize the connection.

Recall: when we discussed scatterplots in the section on visualizing data, remember that the order of the variables is significant. We need to decide which one *determines* the other, and call it x (the other will be y). Since it makes more sense to say that the size of a house determines its price rather than the other way around, let's call the size x and the price y .

This will also be important later, when we start making predictions. We'll want to predict the *price* of a house with a given *size*; in general terms, we'll always predict y using a given value of x .

Here's the scatterplot:



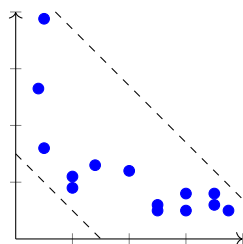
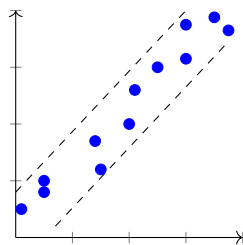
Although the trend is that larger houses cost more, notice the last two points on the right: clearly the price doesn't depend *only* on the size, since the last two houses have the same square footage, but one is priced over \$30,000 higher.

First of all, notice that the trend we expected is visible: as you move to the right (as square footage increases), the points rise vertically (higher prices). Second, notice how it looks like the points are more or less grouped around a straight line.

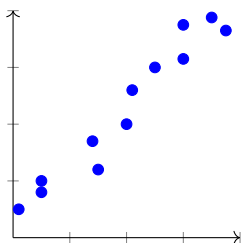
Association

- Two variables have a **linear association** if the scatter plot shows the data clustering around a straight line.
- Two variables have a **positive association** if larger values of one variable are linked with larger values of the other variable.
- Two variables have a **negative association** if larger values of one variable are linked with smaller values of the other variable.

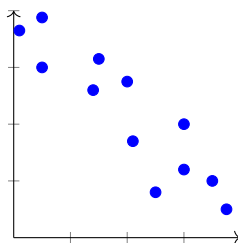
One way to visualize this is to think about using a pair of parallel lines to enclose all the points on the graph. The closer you can draw these lines, the more tightly the points are clustered along a straight line, and thus the linear association is stronger.



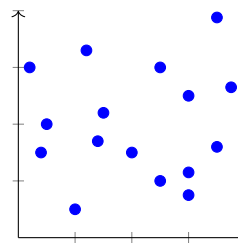
Here are a few examples of positive and negative association:



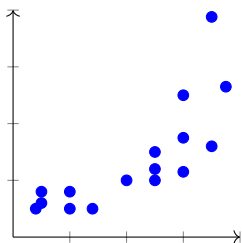
Positive linear



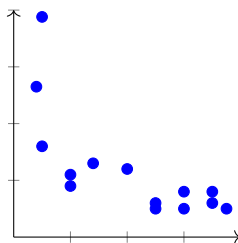
Negative linear



No association



Positive nonlinear



Negative nonlinear

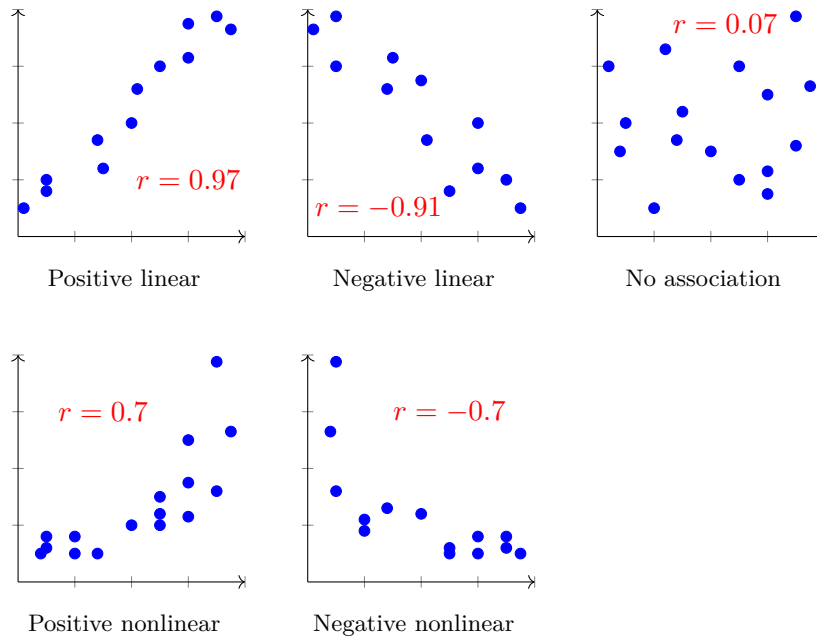
At the moment, whether the association is linear or nonlinear seems like a subjective question; just by looking at the graphs, we can often convince ourselves one way or the other. We need a specific way to measure how strong a linear association is; that's why we'll start by looking for **correlation**, we'll calculate a *correlation coefficient* which will answer exactly this question.

Once we're convinced that there is a strong linear relationship, the next question is, what exactly *is* the relationship? For this, we'll build a linear equation that we can use to make new predictions, like if we wanted to predict the value of a house, and we could measure its square footage.

Correlation

To measure the strength of a linear relationship (in other words, *how linear is this relationship?*), we calculate a single number, called the correlation coefficient; we label it r .

Before we see how to calculate it, let's look at a few examples. Let's go back to those examples with different associations, and this time we'll show the value of r for each dataset.

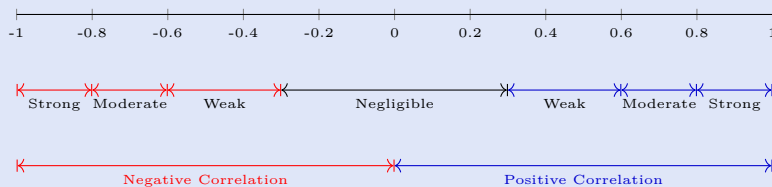


From even this small set of examples, we can observe the following features of the correlation coefficient.

Properties of the Correlation Coefficient

1. The sign of r describes the direction of the association (positive r , positive association, and vice versa).
2. The closer r is to 1 or -1 , the more linear the association is.
3. The closer r is to 0, the less linear association there is.

As a general rule of thumb, a value greater than 0.8 or less than -0.8 (r is always between -1 and 1) corresponds to a strong linear relationship.



There is nothing set in stone about the value of 0.8 marking the difference between strong and moderate linear correlation; these values are simply ones commonly agreed upon within the statistical community (and there is not a universal standard, either). You may find others who use 0.7 as the breakpoint; for consistency in this book, however, we will use 0.8. Of course, if we calculate r to be something like 0.78, that's close enough to 0.8 to call it a strong linear relationship.

EXAMPLE 1 **INTERPRETING A CORRELATION COEFFICIENT**

If a dataset comparing two variables yields an r -value of -0.83 , what does that tell us?

Solution

First of all, since it is negative, the correlation is negative, meaning that as one variable increases, the other decreases.

Second, since this value is close to -1 , and in the range we defined as “strong” correlation, there is a strongly linear relationship between these two variables.

Calculating r The formula for the correlation coefficient is quite complicated. We will show it here simply to be thorough, but we will allow the calculator to handle the computation for us.

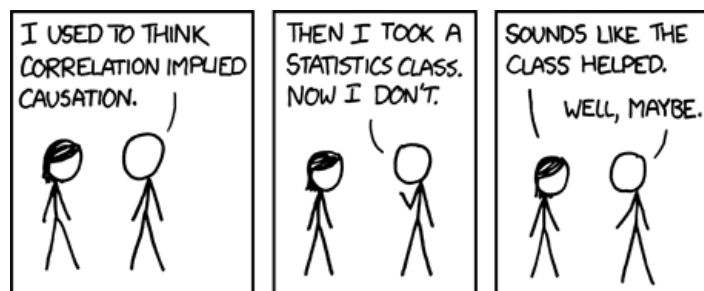
$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

Note: It doesn’t matter which variable is x and which is y ; the correlation coefficient is the same either way.

Again, we won’t bother using this formula; it’s tedious to use, and it doesn’t add anything to our understanding.

Correlation Does Not Imply Causation

Just because two variables are highly correlated, that doesn’t mean that one causes the other. In the example of the house sizes and their price, there IS a causal link, but you can’t assume that in every case where there’s a strong correlation.



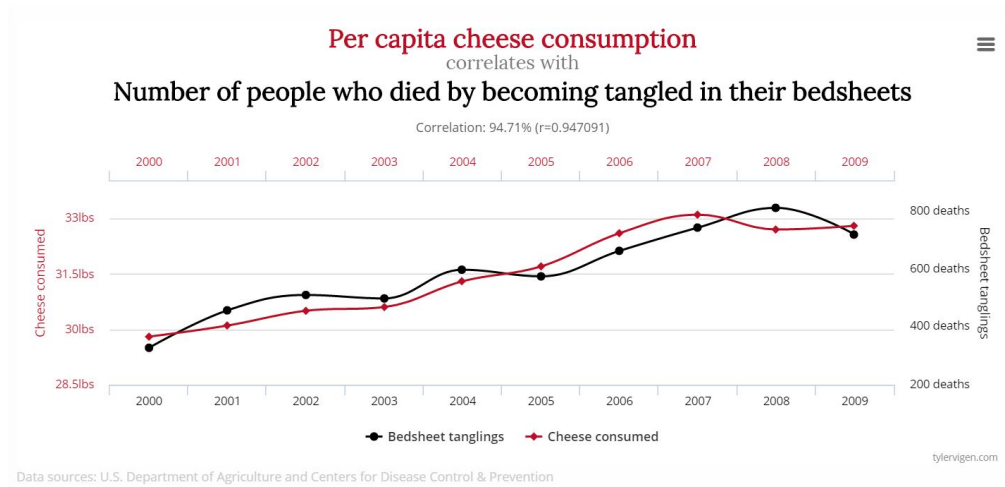
“Correlation doesn’t imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing ‘look over there.’ ”

xkcd.com

For instance, the number of injuries sustained in a swimming pool is correlated with the sales of ice cream cones. Do ice cream cones cause injuries, or vice versa? Of course not; it’s just that both of them are much more common in warmer weather.

In that case, we call the weather a **confounder**, a third variable that is related to the two we’re interested in. If we don’t consider this third variable, it can fool us into thinking that the other two cause each other.

Even if there isn't a confounder, sometimes two variables can be related by coincidence. There's a book and website by Tyler Vigen (tylervigen.com) devoted to showing such correlations. For example:



Regression

We now have the first part of the puzzle: we can measure *correlation*, so we can tell whether or not there is a strong linear relationship. Once we've decided that there is, we turn our attention to the problem of **regression**, finding a linear equation to describe this relationship.

Let's go back to the data involving house sizes and prices:

Size (sq. ft.)	Selling Price (\$1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

We haven't shown this calculation yet (we'll see it when we start using the calculator in a bit), but for this dataset,

$$r = 0.9.$$

Thus, we know that there is a strong *positive* linear relationship, meaning that as one variable increases, so does the other; larger houses cost more, and vice versa.

Okay, so there is a linear relationship here, but what is it, exactly? We want to find a linear relationship of the form

$$\hat{y} = ax + b$$

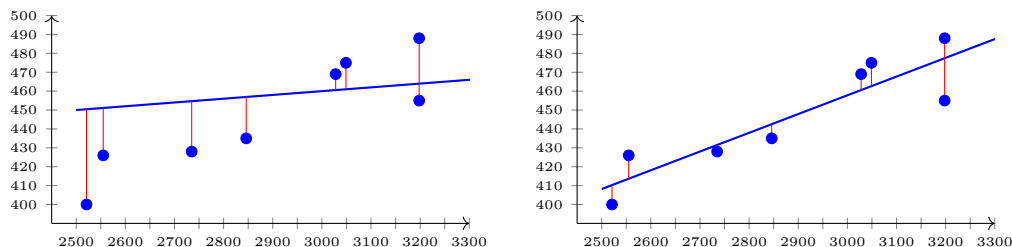
using *linear regression*.

First, a note about notation: \hat{y} , which we read as “*y* hat,” refers to *predicted y*-values based on given *x* values. In other words, the first house, with 2521 square feet, is selling for \$400,000, so the *x* value of 2521 corresponds to an *actual y* value in the dataset of \$400,000.

However, once we define an equation, we can plug 2521 into that equation for *x*, and the equation will *predict* a price that goes with that square footage. That price will probably not be exactly \$400,000, but it will be close. The difference between the *actual* value *y* and the *predicted* value \hat{y} is called the **error** or **residual** for that *x* value:

$$\text{Residual} = y - \hat{y}$$

For any equation we define, we can calculate the residuals for all the points in our dataset. Here's the key: *we want to minimize the residuals*, so we can keep trying different equations, moving the prediction line around, until we're satisfied:



The line on the right is a better fit for the data, because the residuals are smaller, on the whole.

Why Least Squares?

Since some residuals will be positive and others will be negative, just minimizing the total or sum of the residuals isn't good enough, because we could draw a line that doesn't fit the data very well, but simply balances the positive and negative errors. It turns out that if we square all the residuals, we get all positive answers, and minimizing this result leads to the line of best fit.

How do we get the *best* line, though? This ideal line is called the **line of best fit** or the **least-squares regression line**. The derivation of the formula is much too complicated to show here, but the formula is shown below. Note that for actual calculations, we'll simply use the calculator, so we won't use this formula, other than one quick example.

Equation of the Regression Line

Given ordered pairs (x, y) with sample means \bar{x} and \bar{y} , sample standard deviations s_x and s_y , and correlation coefficient r , the equation of the least-squares regression line for predicting y from x is

$$\hat{y} = ax + b$$

where the slope and intercept are given by

- **Slope:** $a = r \frac{s_y}{s_x}$
- **Intercept:** $b = \bar{y} - a\bar{x}$

EXAMPLE 2

REGRESSION WITH HOUSE PRICES

Construct the least-squares regression line for the house price data given below, knowing that $r = 0.9$.

Size (sq. ft.)	Selling Price (\$1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Solution

First, we need to calculate the mean and standard deviation for each of the variables. To do this, we can enter the data as shown into the calculator and use **2-Var Stats** (we could also enter one column at once and use **1-Var Stats** twice, but this is quicker).

L1	L2	L3	1
2521	400		
2555	426		
2735	428		
2846	435		
3028	469		
3049	475		
3198	488		
3198	455		

2-Var Stats
Xlist:L1
Ylist:L2
FreqList:
Calculate

2-Var Stats
$\bar{x}=2891.25$
$\Sigma x=23130$
$\Sigma x^2=67383000$
$s_x=269.4935727$
$\sigma_x=252.0881542$
$n=8$

Scanning through the results, we collect the following:

$$\bar{x} = 2891, \bar{y} = 447$$

$$s_x = 269.5, s_y = 29.7$$

Notice that in order to find b in the regression equation, we must know a , so we'll start by calculating a :

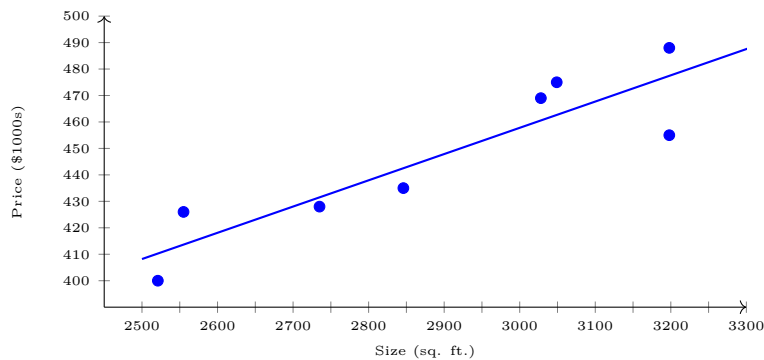
$$\begin{aligned} a &= r \frac{s_y}{s_x} \\ &= (0.9) \left(\frac{29.7}{269.5} \right) \\ &= 0.099 \end{aligned}$$

Now we can find b :

$$\begin{aligned} b &= \bar{y} - a\bar{x} \\ &= 447 - (0.099)(2891) \\ &= 160.8 \end{aligned}$$

Therefore, the regression line (graphed below) is given by the equation

$$\hat{y} = 0.099x + 160.8$$



We've now computed the regression line manually exactly once, and from now on, we'll rely on the calculator to do it for us, because using the formula doesn't give us any extra insight.

Regression with the Calculator

Once you've entered the data into the calculator, with the x values in the first list and y values in the second list, press the **STAT** button and scroll to the right to access the **CALC** menu (where we found **1-Var Stats** earlier). There, look for the fourth option, labeled **4:LinReg(ax+b)**. After selecting it, you can leave all the default options alone, as long as you entered x and y in the usual order. Select **Calculate** to see the results.

```

EDIT  CALC  TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

```

```

LinReg(ax+b)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:
Calculate

```

```

LinReg
y=ax+b
a=.0991979543
b=160.1939146
r²=.8110655049
r=.9005917526

```

The results shown here are for the house price example; notice that the answers we calculated manually are a bit different due to rounding the answers along the way, but the differences are small.

Note: when you first follow this process, you may not see the r value in the results. If you don't, there's a setting you can change to display it. To do this, press **2ND** **0** to access the full catalog, then use the arrow keys to scroll down (the options are listed in alphabetical order) to an option called **DiagnosticOn**. Press **ENTER** twice to turn this on, and then repeat the steps above; now r should appear with the other results.

EXAMPLE 3 REGRESSION LINE

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200.

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Find the equation of the least-squares regression line that can be used to predict a student's performance on the final exam based on their third test score.

Solution

Note: To find r with the calculator, we calculate the equation of the line of best fit. However, if we then find that r is too close to 0 for a good linear association, we will discard the linear equation. The calculator will calculate the linear equation whether it is relevant or not; it's up to us to interpret the results.

First, enter the data in the calculator, then use the **LinReg** function to calculate the regression line.

```

EDIT  ▢ 2ND  ▢ TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

```

```

LinReg(ax+b)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:
Calculate

```

```

LinReg
y=ax+b
a=4.827394209
b=-173.513363
r²=.4396931104
r=.663093591

```

The correlation coefficient is

$$r = 0.66$$

so there is a moderate linear relationship, and it is a positive association, as we might expect.

The regression line has the equation

$$\hat{y} = 4.83x - 173.51$$

TRY IT

Consider the following data set.

x	9	5	7	13	-8	-2	6	-10
y	3	3	31	36	0	3	-2	-14

1. Calculate r , the correlation coefficient. Interpret this value.
2. Compute the least-squares regression line for this data set.

Making Predictions

One of the primary reasons for computing a regression equation is to use it to make predictions about new x values. For instance, using the house price dataset, we could assess the value of a new home on the market by measuring its square footage.

PREDICTING HOUSE VALUES

EXAMPLE 4

Using a sample of houses on the market, we found the following regression equation to predict the price y based on the square footage x :

$$\hat{y} = 0.099x + 160.8$$

Use this equation to predict the price of homes with the following square footage values:

- (a) 2700 square feet
- (b) 4500 square feet

Which prediction do you expect to be more reliable?

Solution

- The predicted price for a home with 2700 square feet is

$$\begin{aligned}\hat{y} &= 0.099(2700) + 160.8 \\ &= 428.1\end{aligned}$$

Recall that prices are listed in thousands of dollars, so this corresponds to \$428, 100.

- For 4500 square feet:

$$\begin{aligned}\hat{y} &= 0.099(4500) + 160.8 \\ &= 606.3\end{aligned}$$

so the predicted price is \$606, 300.

Now, which of these is likely to be more reliable? The key is that the first house falls within the range of sizes that are listed in our sample; this is called an **interpolation**. On the other hand, our sample doesn't include any houses as large as the second one, so that represents an **extrapolation**. In general, interpolations are more reliable, because the trend may be different outside the range for which we have data. Maybe really large houses are more expensive than expected because they are built with top-quality materials.

Using the data for students' test scores shown below, predict the final exam score for students who scored 60 and 80 on the third exam.

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

TRY IT

Note: Not every x value that you can plug into the regression equation is a meaningful one. For instance, you could try predicting the final exam score of a student who got a 90 on the third test (even though the third test scores can only go up to 80), or the price of a home with negative square footage, and each equation will dutifully give you a value. Just note that that value is meaningless; you need to use common sense when making predictions.

EXAMPLE 5 NFL QUARTERBACKS

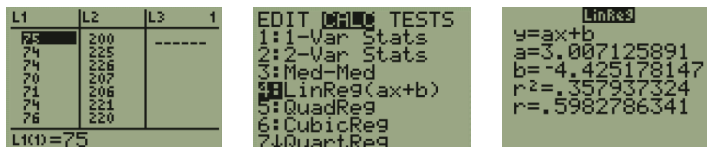
The following table lists the heights (in inches) and weights (in pounds) of 10 NFL quarterbacks in the 2019 season.

Name	Height	Weight
Lamar Jackson	75	200
Patrick Mahomes	74	225
Dak Prescott	74	226
Kyler Murray	70	207
Russell Wilson	71	206
Deshaun Watson	74	221
Matt Ryan	76	220
Josh Allen	77	233
Drew Brees	72	209
Aaron Rodgers	74	225

- Calculate the correlation coefficient for this data.
- Is there are strong linear relationship?
- Compute the regression line for predicting weight from height.
- Graph the data and the regression equation.
- Predict the weight of a quarterback who is 73 inches tall.
- Does Drew Brees weigh more or less than the weight predicted by the regression line, based on his height?

Solution

First, enter the data in the calculator and use the **LinReg** option; this will give us the answers for r and the regression equation.



- The value of the correlation coefficient is

$$r = 0.6$$

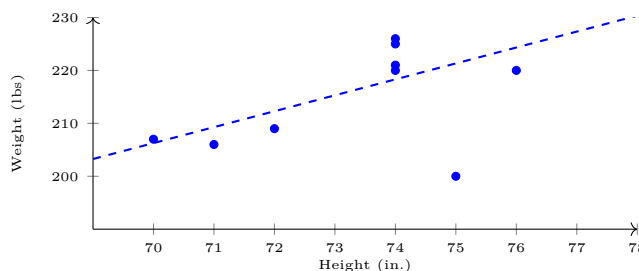
Notice in the graph below that there is a clear outlier; Lamar Jackson is one of the tallest quarterbacks, but also one of the lightest. If we remove him from the sample, the correlation coefficient jumps to 0.88, greatly strengthening the linear relationship.

This illustrates how *sensitive to outliers* the correlation coefficient is.

- The relationship isn't strong ($r \not\geq 0.8$), but there is a moderate linear relationship ($r \geq 0.6$), so we'll continue.
- The regression equation is

$$\hat{y} = 3.01x - 4.43$$

- The graph is shown below (note that some of the points overlap).



- (e) Simply substitute 73 for x :

$$\hat{y} = 3.01(73) - 4.43$$
$$= \boxed{215.3 \text{ pounds}}$$

- (f) Since Drew Brees is 72 inches tall, we can use the regression equation to predict his weight:

$$\hat{y} = 3.01(72) - 4.43$$
$$= 212.3 \text{ pounds}$$

A QB of that height is predicted to weight 212 pounds, so since Drew Brees only weighs 209 pounds, he weighs less than predicted.

A blood pressure measurement consists of two numbers: the systolic pressure, which is the maximum pressure taken when the heart is contracting, and the diastolic pressure, which is the minimum pressure taken at the beginning of the heartbeat. Blood pressures were measured (in millimeters of mercury, mmHg) for a sample of 16 adults.

Systolic	134	115	113	123	119	118	130	116
Diastolic	87	83	77	77	69	88	76	70
Systolic	133	112	107	110	108	105	157	154
Diastolic	91	75	71	74	69	66	103	94

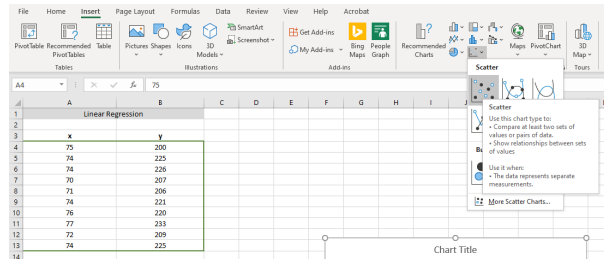
- Calculate r , the correlation coefficient.
- Do you think there is a strong linear association?
- Compute the regression line for predicting the diastolic pressure from the systolic pressure.
- Predict the diastolic pressure for a patient whose systolic pressure is 125 mmHg.

Using Excel

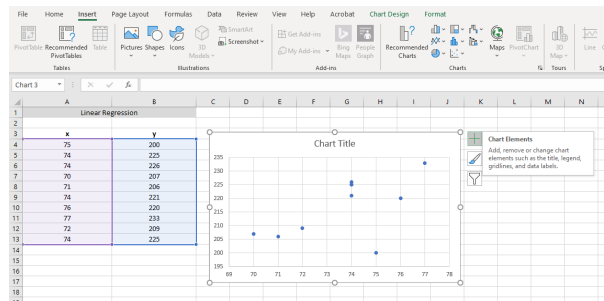
We can also use Excel to calculate regression equations. To begin, enter the data as shown (we'll use the data from the QB height/weight example):

	A	B	C	D
1	Linear Regression			
2				
3	x	y		
4	75	200		
5	74	225		
6	74	226		
7	70	207		
8	71	206		
9	74	221		
10	76	220		
11	77	233		
12	72	209		
13	74	225		
14				
15				

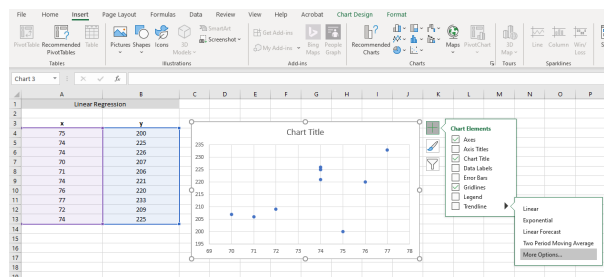
Next, we need to insert a scatterplot of these data points; select the Insert menu at the top of the screen, then select the scatterplot option under the Charts section (select the first type of scatterplot under that submenu).



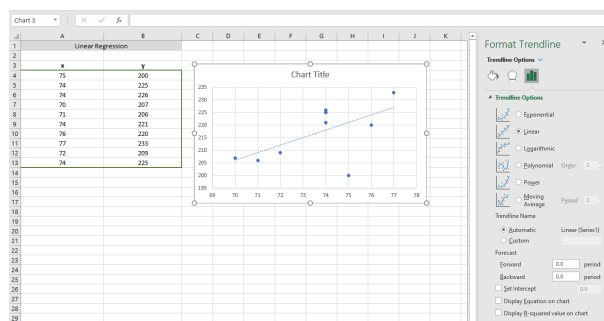
This creates a graph similar to the one that we drew earlier for this same example; we could add a title to the chart and the axes, but we won't bother for this example.



To add a regression line, click on the plus symbol at the upper right (when the chart is selected). One of the options is to add a trendline. If you check that box, the line will be drawn, but by default, Excel won't show the equation of the line, which is what we're after, so we need to select "More Options" after clicking on the arrow next to the trendline option.

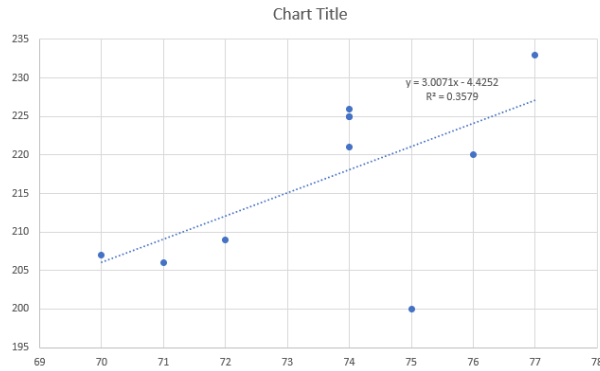


This brings up the following menu. Notice that there are different types of equations we can use to model the data, including several that we'll encounter later in the chapter. For now, though, leave the linear option selected, but check the box at the bottom of the menu that says "Display Equation on chart."



We can also select the option to display the R-squared value on the chart (more on this in a moment).

This leads to the following:



The equation, as calculated by Excel, is

$$\hat{y} = 3.0071x - 4.4252$$

which, after rounding, is identical to the one obtained by the calculator (this is no surprise, since they're using the same formulas).

Note: Excel does not directly report the value of r , but it can show the value R^2 , which is simply the square of the correlation coefficient, so we can calculate the value of r with one step:

$$r = \sqrt{R^2} = \sqrt{0.3579} = \boxed{0.598}$$

After rounding, this is the same as the value we found earlier (0.6).

Exercises 3.4

1. If the value of r is 0.91 for a dataset comparing two variables, what does that tell us?
2. If the value of r is -0.43 for a dataset comparing two variables, what does that tell us?
3. Compute the equation of the regression line for a dataset that has the following statistics:
4. Compute the equation of the regression line for a dataset that has the following statistics:

$$\bar{x} = 5, \quad s_x = 2, \quad \bar{y} = 1350, \quad s_y = 100, \quad r = 0.70$$

$$\bar{x} = 152, \quad s_x = 24.5, \quad \bar{y} = 26, \quad s_y = 2.7, \quad r = -0.82$$

For problems 5–8, use the dataset shown in the table below. This is a sample of 15 players in Major League Baseball, chosen from the starting lineups of teams in 2019. The table shows the team, age, position, height, and salary for each player, as well as several statistics from that season. These include the number of games they played (G), their batting average (AVE) (the proportion of their at-bats for which they got a hit), and their home runs (HR).

Name	Team	Age	Height	G	AVE	HR	Salary
Cedric Mullins	Orioles	25	173 cm	22	.094	0	\$557,500
Tim Anderson	White Sox	26	185 cm	123	.335	18	\$1,400,000
Christin Stewart	Tigers	25	183 cm	104	.233	10	\$556,400
Alex Gordon	Royals	35	185 cm	150	.266	13	\$20,000,000
Jonathan Schoop	Twins	27	185 cm	121	.256	23	\$7,500,000
Marcus Semien	Athletics	29	183 cm	162	.285	33	\$5,900,000
Yandy Diaz	Rays	28	188 cm	79	.267	14	\$558,400
Randal Grichuk	Blue Jays	28	188 cm	151	.232	31	\$5,000,000
Josh Donaldson	Braves	33	185 cm	155	.259	37	\$23,000,000
Joey Votto	Reds	36	188 cm	142	.261	15	\$25,000,000
Cody Bellinger	Dodgers	24	193 cm	156	.305	47	\$605,000
Ryan Braun	Brewers	35	188 cm	144	.285	22	\$19,000,000
Maikel Franco	Phillies	27	185 cm	123	.234	17	\$5,200,000
Ian Kinsler	Padres	37	183 cm	87	.217	9	\$3,750,000
Marcell Ozuna	Cardinals	28	185 cm	130	.241	29	\$12,250,000

5. Suppose we want to try to predict a player's salary based on the number of home runs they hit (HR).
 - (a) Before doing any calculations, does it seem likely that there will be a strong association between these two variables? If so, which direction do you expect for the association?
 - (b) Calculate the value of the correlation coefficient, r , using a calculator.
 - (c) Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
 - (d) Find the equation of the regression line for this association.
(Note: this may not be meaningful, depending on the value of r , but we can still use it for practice.)
 - (e) Ignoring the possibility that the regression line may not be a good fit for the data, use this regression line to predict the salary of a player who hits 21 home runs. Then predict the salary of a player who hits 70 home runs. Which prediction is likely to be more accurate?
6. Suppose we want to try to predict a player's height based on their age.
 - (a) Before doing any calculations, does it seem likely that there will be a strong association between these two variables? If so, which direction do you expect for the association?
 - (b) Calculate the value of the correlation coefficient, r , using a calculator.
 - (c) Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
 - (d) Find the equation of the regression line for this association.
(Note: this may not be meaningful, depending on the value of r , but we can still use it for practice.)
 - (e) Ignoring the possibility that the regression line may not be a good fit for the data, use this regression line to predict the height of a player who is 26 years old.

7. Suppose we want to try to predict the number of home runs that a player hits (HR) based on the number of games they play (G).

- Before doing any calculations, does it seem likely that there will be a strong association between these two variables? If so, which direction do you expect for the association?
- Calculate the value of the correlation coefficient, r , using a calculator.
- Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
- Find the equation of the regression line for this association.
(Note: this may not be meaningful, depending on the value of r , but we can still use it for practice.)
- Ignoring the possibility that the regression line may not be a good fit for the data, use this regression line to predict how many home runs a player will hit if he plays 100 games.

8. Suppose we want to try to predict a player's batting average based on the number of home runs they hit.

- Before doing any calculations, does it seem likely that there will be a strong association between these two variables? If so, which direction do you expect for the association?
- Calculate the value of the correlation coefficient, r , using a calculator.
- Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
- Find the equation of the regression line for this association.
(Note: this may not be meaningful, depending on the value of r , but we can still use it for practice.)
- Ignoring the possibility that the regression line may not be a good fit for the data, use this regression line to predict the batting average of a player who hits 12 home runs. Then predict the batting average of a player who hits 58 home runs. Which prediction is likely to be more accurate?

9. The data set below shows the GMAT scores for five MBA students and the students' grade point averages (GPA) upon graduation.

GMAT	660	580	480	710	600
GPA	3.7	3.0	3.2	4.0	3.5

- Calculate r , the correlation coefficient between these two variables.
- Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
- Compute the regression line for predicting GPA from GMAT score.
- Predict the GPA of a student who gets a score of 500 on the GMAT.
- Does the student with a GMAT score of 580 have a higher or lower GPA than the one predicted by the regression line?

10. The data set below shows the mileage and selling prices of eight used cars of the same model.

Mileage	21,000	34,000	41,000	43,000	65,000	72,000	76,000	84,000
Price	\$16,000	\$11,000	\$13,000	\$14,000	\$10,000	\$12,000	\$7,000	\$7,000

- Calculate r , the correlation coefficient between these two variables.
- Interpret the value of r ; specifically, describe the direction of the trend and the strength of the linear association.
- Compute the regression line for predicting price from mileage.
- Predict the price of a car with 30,000 miles.
- Does the car with 43,000 miles on it have a higher or lower price than the one predicted by the regression line?

SECTION 3.5 The Normal Distribution

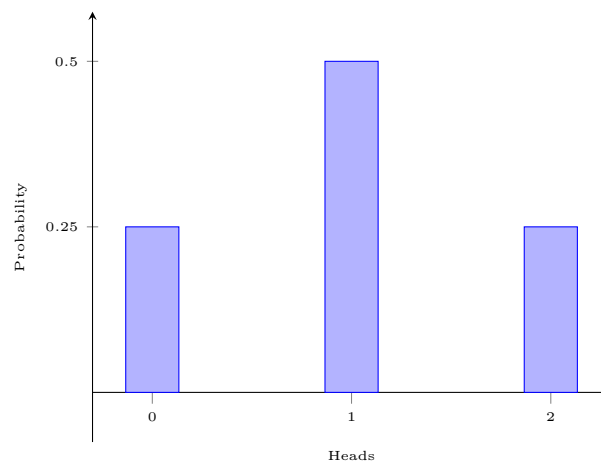


In 1654, a French writer who called himself Chevalier de Méré (it must have sounded better than his real name, Antoine Gombaud), wrote a letter to two mathematicians, Blaise Pascal and Pierre de Fermat. He had a question about a game of chance, and the mathematicians leaped at the gambler's challenge. Between the two of them, they began to develop the theory of probability, and gambling was never the same again.

One of the problems that Pascal and Fermat considered was related to flipping a coin repeatedly. If you flip a coin once, what's the probability that it comes up heads? Since there are two possibilities, and each is equally likely, the answer is $1/2$ or 50%. Now, what if you flip a coin twice? How many heads could you get? Here are all the possibilities:

HH HT TH TT

Notice that you could get a total of 2 heads (1 out of 4), 1 head (2 out of 4), or 0 heads (1 out of 4). We can graph this:

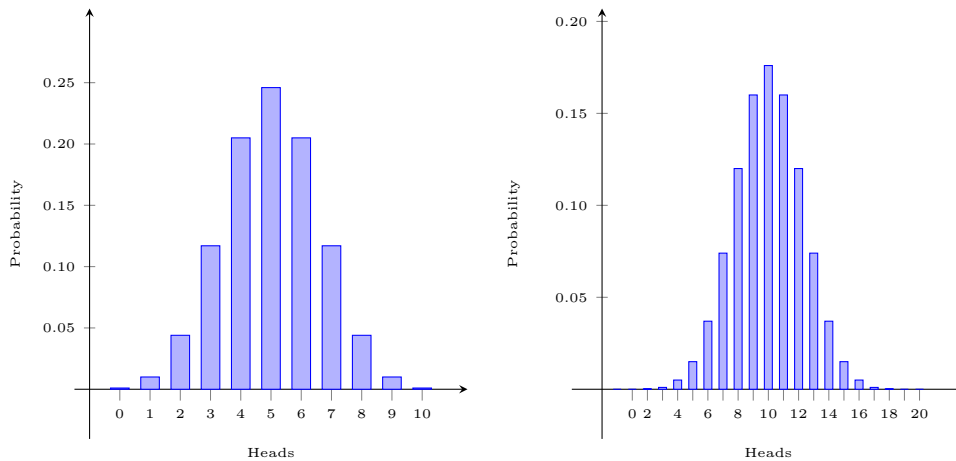


It turns out that, as the number of coin flips increases, an interesting pattern emerges. However, the results always have the following two features present:

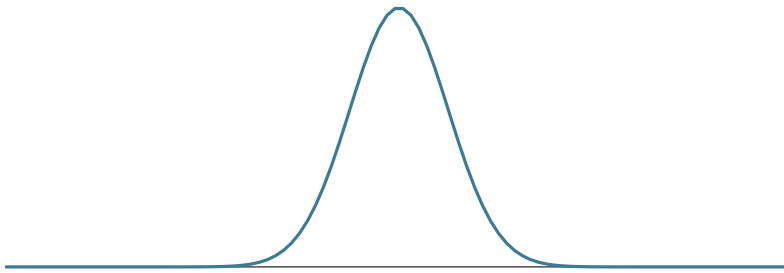
1. The probabilities are symmetric (notice above how the probability of 2 heads and 0 heads are equal).
2. The highest probability occurs at the center, and decreases from there out to the extremes.

The second feature makes sense, because it would be really unusual to get a long unbroken string of heads or tails; a mixture is more common, and the likeliest result would be to get an even number of each.

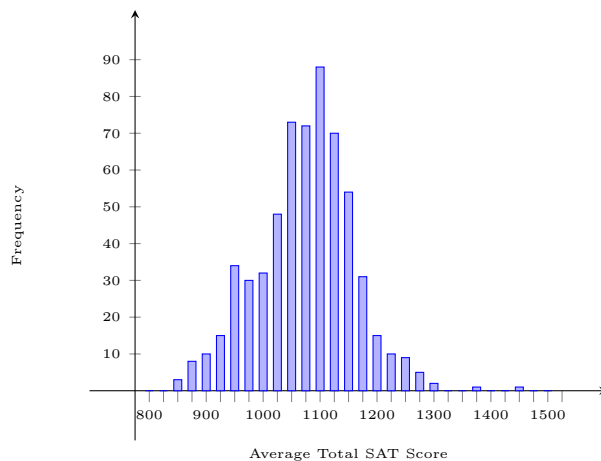
Here are the graphs for larger numbers of trials (flipping the coin 10 or 20 times):



The shape of this graph starts to look smoother and smoother for higher numbers, but it always keeps this particular shape:



Here's the surprising part: this curve starts to pop up over and over again. For instance, the North Carolina public school system released the average SAT score for each high school in the state. The results are shown in the histogram below.



In fact, almost any measurement of natural phenomena starts to exhibit this familiar pattern. If you measure people's height, shoe size, ear length, or IQ, or the amount of milk that a cow produces, or how long a natural pregnancy lasts, you'll see the results follow this same pattern: there's a value that's the most common, and the further you get from that value, the less common the result. It's actually more specific than that, and there's a particular mathematical pattern underlying all of this.



C.F. Gauss

It didn't take long for mathematicians to notice this pattern, and they gave this special curve several names: it is often called the *Gaussian curve* after Carl Friedrich Gauss, one of the most prolific mathematicians of all time. It is also known as the **normal curve**³ or the **normal distribution**, and sometimes the term *bell curve* is used in reference to its shape, how it looks similar to a bell.

Different Bell Curves

If you look back at the comparison between flipping a coin 10 times and 20 times, you can see that the curve, while it has the same overall form, has a different specific shape in each case: one is shorter and wider, and the other is taller and narrower.

In other words, we can observe that different normal curves have different amounts of *spread*. Because of that, it may not surprise you to learn that we can describe this spread using one of its measures that we discussed earlier in this chapter: the *standard deviation*.

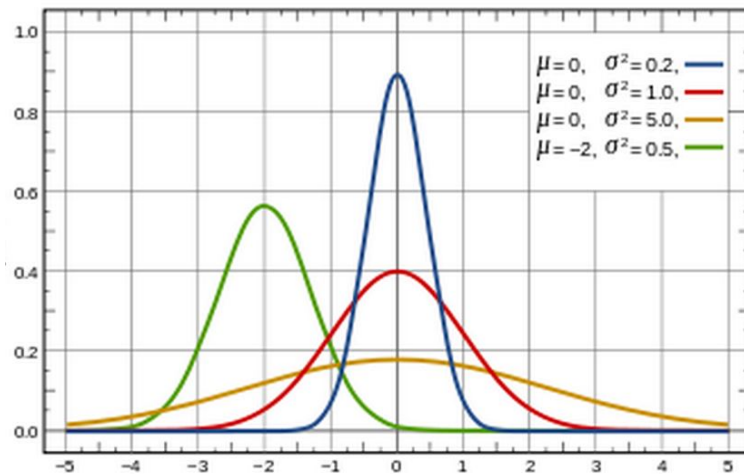
Also, if you look closer at those two side-by-side bar charts, you'll notice that the center occurs at different locations: in the experiment with 10 coin flips, the center is at 5, and the center is at 10 for the other. It turns out, again with little surprise, that we can describe this center using the *mean*.

Normal Curve Parameters

Every normal curve is determined by two values: its center and spread. Specifically, the center is defined by the **mean**, and the spread is defined by the **standard deviation**.

Knowing these two values tells us everything we need to know about a normal curve; they completely define its shape.

Here are a few examples of different normal curves:



Notice how three of them have the same mean; they are centered at the same point. They all have different standard deviations; the one with the smallest standard deviation is the narrowest (and thus the tallest), and the one with the largest standard deviation is the most spread out (and thus the shortest).

The Empirical Rule

The beauty of the normal distribution is that there is a predictable pattern for how likely a certain value is. For instance, in the example of the NC SAT scores, we could predict how likely it is for a particular school's average score to be between 900 and 1000. Or if we knew the mean and standard deviation for IQ scores, we could tell approximately how many people have an IQ over 120.

³The term *normal curve* was popularized by the great statistician Karl Pearson, who said this in 1920:

Many years ago [in 1893] I called the Laplace-Gaussian curve the *normal curve*, which name, while it avoids the international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another *abnormal*.

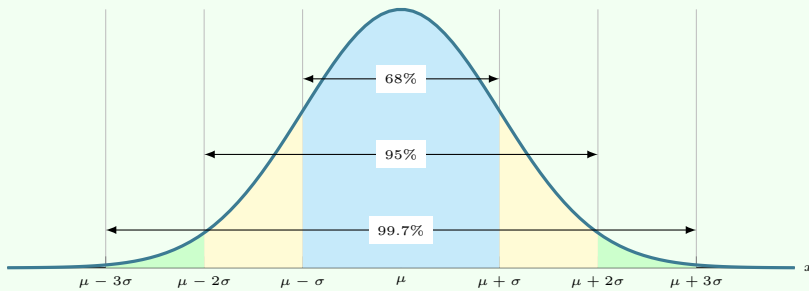
To do this, we'll use what's called the **Empirical Rule**, which is a simplified version of a much larger concept, but it's enough for us to work with. The Empirical Rule predicts what percentage of the data will fall within one step, two steps, or three steps of the center (mean), where the step is the size of the standard deviation.

The Empirical Rule

Approximately 68% of the data is within **one** standard deviation of the mean.

Approximately 95% of the data is within **two** standard deviations of the mean.

Approximately 99.7% of the data is within **three** standard deviations of the mean.



Note that this diagram uses μ for the population mean (as opposed to \bar{x} for the sample mean) and σ for the population standard deviation (as opposed to s for the sample standard deviation).

As you can see, another name for the Empirical Rule is the “68-95-99.7 Rule.” Let’s try an example using this rule.

COLLEGE ENTRANCE EXAMS

EXAMPLE 1

The scores on a college entrance exam are normally distributed with a mean of 52 points and a standard deviation of 11 points. About 95% of the values lie between what two scores?

We know that 95% of the data is within two standard deviations of the mean.

Solution

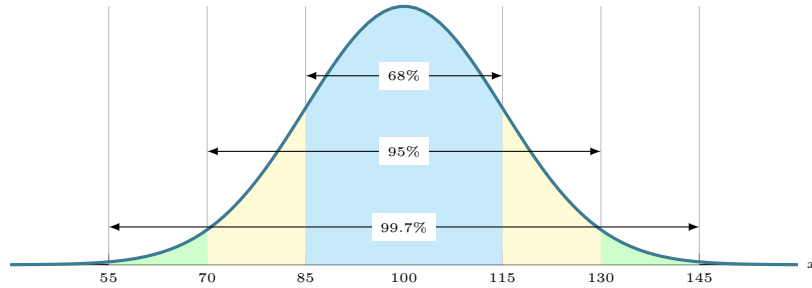
$$\begin{aligned} \text{Scores} &= \text{mean} \pm (2)(\text{standard deviation}) \\ &= 52 \pm (2)(11) \\ &= 52 \pm 22 \\ &= \boxed{(30, 74)} \end{aligned}$$

Hence, 95% of the values fall between a score of 30 and a score of 74.

EXAMPLE 2 THE INTELLIGENCE QUOTIENT

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Use the Empirical Rule to find the data that is within one, two, and three standard deviations of the mean.

Solution



1. 68% of the data is within one standard deviation of the mean.

$$\begin{aligned}\text{IQ} &= \text{mean} \pm (1)(\text{standard deviation}) \\ &= 100 \pm (1)(15) \\ &= \boxed{(85, 115)}\end{aligned}$$

Thus, 68% of people have an IQ between 85 and 115.

2. 95% of the data is within two standard deviations of the mean.

$$\begin{aligned}\text{IQ} &= \text{mean} \pm (2)(\text{standard deviation}) \\ &= 100 \pm (2)(15) \\ &= 100 \pm 30 \\ &= \boxed{(70, 130)}\end{aligned}$$

Thus, 95% of people have an IQ between 70 and 130.

3. 99.7% of the data is within three standard deviations of the mean.

$$\begin{aligned}\text{IQ} &= \text{mean} \pm (3)(\text{standard deviation}) \\ &= 100 \pm (3)(15) \\ &= 100 \pm 45 \\ &= \boxed{(55, 145)}\end{aligned}$$

Thus, 99.7% of people have an IQ between 55 and 145.

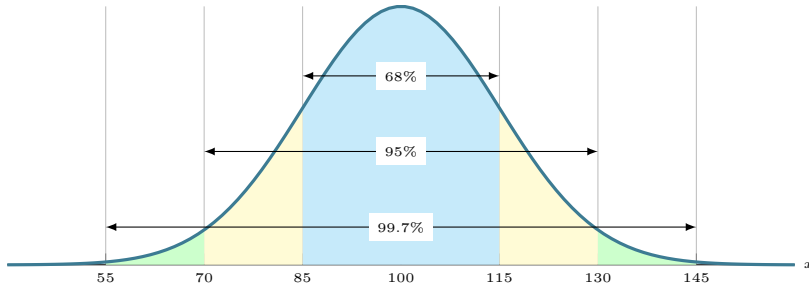
Again, this rule gives a way to decide whether a data point is unusual or not. An IQ of over 130 is very unusual, and an IQ of over 145 is even more so.

Since 99.7% have IQs in the range from 55 to 145, only 0.3% of people have IQs outside that range. Since the bell curve is symmetric, half of those, or 0.15% of people (15 people out of 1000) have IQs over 145.

TRY IT

The mean height of boys 15 to 18-years old from Chile is 170 cm with a standard deviation of 6 cm. Male heights are known to be normally distributed. Using the Empirical Rule, find the range of heights that contain approximately 68%, 95%, and 99.7% of the data.

Let's go back to the figure from that last example:

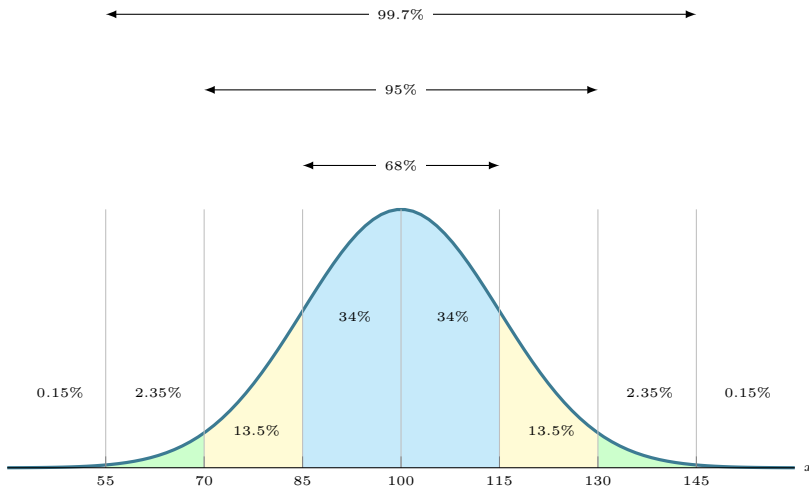


What if we want to find the percentage of the data that falls in some other range? Like what about the percentage of IQs that fall between 100 and 115? Or above 85? Between 70 and 115?

All of this can be done with a little clever analysis of the figure above. We just need to divide it up into segments that are each one standard deviation (15 IQ points) wide and figure out what percentage of the data is contained in each slice.

First of all, notice that the center region (between 85 and 115) contains 68% of the data. Because the graph is symmetric, we can conclude that each half of that contains 34%.

Next, the two yellow regions together contain $95\% - 68\% = 27\%$, so each region contains half of that, or 13.5%. Similarly, the two green regions account for $99.7\% - 95\% = 4.7\%$, so each of them contains 2.35% of the data. Finally, the tails outside the green account for the remaining 0.3% of the data, so each side contains 0.15%.



The important point is not to memorize these percentages, but rather to understand how we figured them out. If you can follow and recreate that process, all you'll have to memorize is the 68–95–99.7 part, and you can reproduce a picture like that one in a minute or two of quick thought. Once you can do that, you can answer questions like the following one.

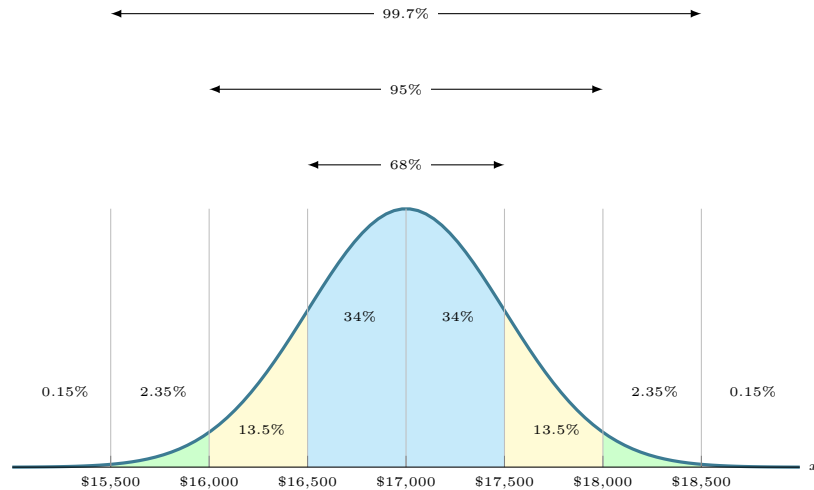
EXAMPLE 3 CAR SALES

Suppose you know that the prices paid for cars are normally distributed with a mean of \$17,000 and a standard deviation of \$500. Use the 68–95–99.7 Rule to find the percentage of buyers who paid

- | | |
|-----------------------------------|-----------------------------------|
| (a) between \$16,500 and \$17,500 | (b) between \$17,500 and \$18,000 |
| (c) between \$16,000 and \$17,000 | (d) between \$16,500 and \$18,000 |
| (e) below \$16,000 | (f) above \$18,500 |

Solution

We can use the same process that was just described to build the following diagram, using the given mean and standard deviation.



You should be able to use the figure above to reason out that

- the percentage of buyers who spent between \$16,500 and \$17,500 was 68%.
- the percentage of buyers who spent between \$17,500 and \$18,000 was 13.5%.
- the percentage of buyers who spent between \$16,000 and \$17,000 was 47.5%.
- the percentage of buyers who spent between \$16,500 and \$18,000 was 81.5%.
- the percentage of buyers who spent below \$16,000 was 2.5%.
- the percentage of buyers who spent above \$18,500 was 0.15%.

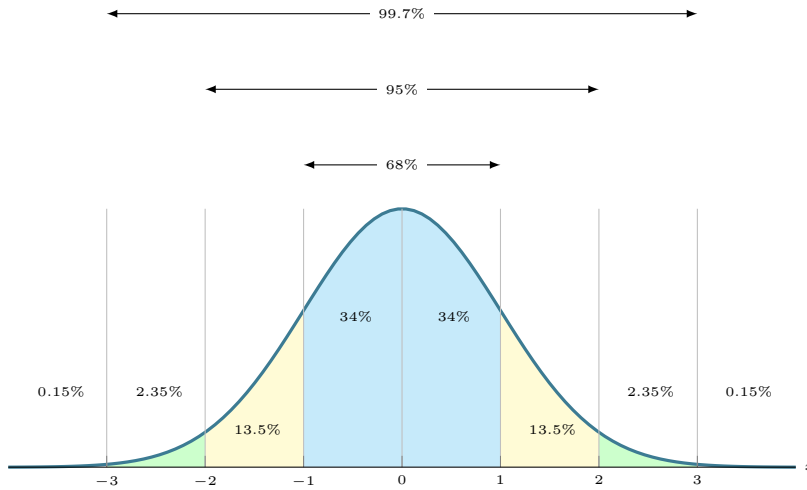
TRY IT

The mean height of boys 15 to 18-years old from Chile is 170 cm with a standard deviation of 6 cm. Male heights are known to be normally distributed. Using the Empirical Rule, find

- the percentage of boys with heights between 158 and 176.
- the percentage of boys with heights above 188.
- the percentage of boys with heights below 164.

Z-Scores

We've seen that the Empirical Rule is consistent no matter what the mean and standard deviation are; the proportion of the data in each section is always the same. We can step back and describe the location of any part of the curve by specifying how many steps (standard deviations) it is to the right or left of the center. The following graph shows this, with positive values representing steps above the mean, and negative values below.



Because of this consistency, we will give these values (-3 , -2 , etc.) a special name; we will call them **z-scores**. A z-score is simply how many standard deviations a point is above or below the mean. For instance, in the IQ example, an IQ score of 130 would have a z-score of 2, and in the car price example, a price of \$16,000 would have a z-score of -2 .

One purpose of z-scores is to put numbers on an equal footing, in terms of how unusual they are. For instance, consider two standardized tests, the SAT and the ACT. These tests use completely different scales for their results, so how could you compare the scores on the two? The answer is, you can *scale the results to put them on equal footing* by finding where the score on each test falls *in relation to the average*.

In other words, when you get a score back for one of these tests, instead of asking “which number is larger,” since that’s irrelevant, you’d ask, “how many standard deviations above the mean was each score?”

Before we do examples, we need a quick formula for calculating z-scores: remember that z will simply count the number of standard deviations above or below the mean that correspond to a particular value in the dataset. So then, to find the z-score for a data point in a data set with a known mean and standard deviation, we just need to find its distance from the mean, and then divide that by the standard deviation to find out how many steps it will take from the mean to reach it.

Z-Scores

If x is a data value in a data set with mean \bar{x} and standard deviation s , the z-score that corresponds to that data value is

$$\text{z-score} = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$$

$$z = \frac{x - \bar{x}}{s}$$

EXAMPLE 4 FEMALE HEIGHTS

Female adult height is normally distributed with a mean of 65 in. and a standard deviation of 3.5 in.

Find the z -scores of the following heights:

(a) 58 in.

(b) 71 in.

Solution

(a) The z -score corresponding to 58 in. is

$$z = \frac{58 - 65}{3.5} = \boxed{-2}$$

(b) The z -score corresponding to 71 in. is

$$z = \frac{71 - 65}{3.5} = \boxed{1.71}$$

Thus, a woman at 71 in. tall is 1.71 standard deviations above the mean, while a woman at 58 in. is 2 standard deviations below the mean.

It should be clear that the 58 in. tall woman is more unusual, since her height is farther from the center.

TRY IT

Scores on the SAT and ACT are normally distributed:

Test	Mean	Std. Deviation
SAT	500	100
ACT	18	6

You score 550 on the SAT and 24 on the ACT. On which test did you have a better score, relative to everyone else who took the test?

We can also do a bit of algebra to work in the opposite direction, if we start with a z -score.

EXAMPLE 5 WORKING BACKWARD FROM Z-SCORES

Scores on an IQ test are normally distributed with a mean of 100 and a standard deviation of 15. Find the IQ score that corresponds to each of the following z -scores.

(a) -1.5

(b) 2.05

Solution

Recall that $z = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$

(a) If the z -score is -1.5 :

$$-1.5 = \frac{\text{IQ} - 100}{15} \rightarrow -22.5 = \text{IQ} - 100 \rightarrow \text{IQ} = \boxed{77.5}$$

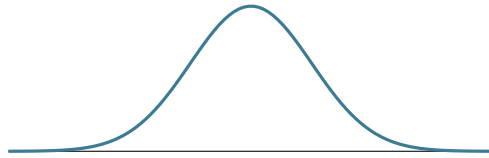
(b) If the z -score is 2.05 :

$$2.05 = \frac{\text{IQ} - 100}{15} \rightarrow 30.75 = \text{IQ} - 100 \rightarrow \text{IQ} = \boxed{130.75}$$

The Normal Distribution and Polls

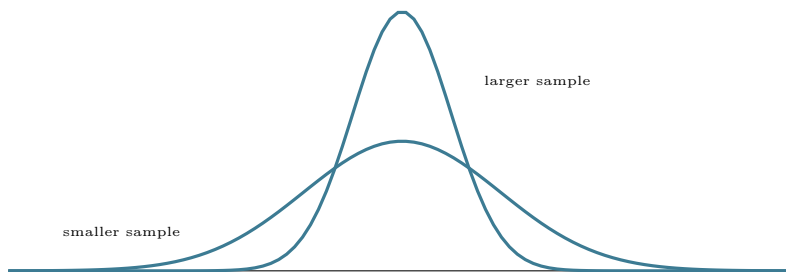
Suppose you're tasked with conducting a straw poll to predict the victor in a close Senate race between Candidate Smith and Candidate Jones. You randomly poll 500 people and ask them who they plan to vote for, and 52% of them respond Smith and 48% Jones. Good so far, but you begin to wonder: is this really an accurate representation of the population? You picked a good sample, but is there any way to put a number on how certain you are that your results are a valid predictor of what the population will do?

The answer is based on the Normal Distribution. The idea is this: if we took another sample and polled them, and then another sample, and another and another, and repeated this process many times over, the results of our poll would begin to look like a normal distribution.



In other words, most of those polls we conducted would look similar to each other, and they would be grouped together. There would be a few polls that would have drastically lower percentages for Smith, and a few would have drastically higher percentages—simply due to the inherent variability of a sample that we can never fully eliminate—but most of them would be clustered around the true percentage of the population that plan to vote for Smith. In other words, the wrong polls would be rare, and the polls that are more right would be more common.

In fact, it turns out that we can specifically define this distribution as a normal distribution, which tells us that we're—for instance—95% confident that the results we got when we took the first poll were within two standard deviations of the mean. Now then, if we make our sample larger, it turns out that the standard deviation on this normal distribution gets smaller, so the results are more precise.



Margin of Error This allows us to define something called the margin of error. You may have seen or heard this term in the context of polls, especially political polls. The margin of error is an inevitable part of using a sample to predict what a larger population will do, and it only depends on n , the size of the sample (strangely enough, it doesn't depend on the size of the population). The larger the sample size, the smaller the margin of error will be, and thus the more precise the results of the poll will be.

Margin of Error

If the sample size of a poll is n , there is at least a 95% chance that the sample percentage lies within

$$\frac{1}{\sqrt{n}} \times 100\%$$

of the population percent.

The margin of error with a 95% confidence level is $\pm \frac{1}{\sqrt{n}} \times 100\%$.

writing $\times 100\%$ simply means that we convert the decimal value $1/\sqrt{n}$ to a percentage

Beware, though, that you don't take for granted that the margin of error is the only thing to worry about; we've already seen that there are other sources of error, like poor sampling. Also, we didn't even talk about other sources of bias, like self-interest or word choice.

EXAMPLE 6 MARGIN OF ERROR

What is the margin of error on a poll with a sample size of 1000 people?

Solution

The margin of error is

$$\pm \frac{1}{\sqrt{1000}} \times 100\% = \boxed{\pm 3.16\%}$$

A margin of error of about 3% (which is common for many political polls) corresponds to a sample size of 1000.

TRY IT

What is the margin of error on a poll with a sample size of 1800 people?

What if we want to work backwards: can we find the sample size needed for a particular margin of error?

EXAMPLE 7 SAMPLE SIZE

If you want a poll to have a margin of error of 2% or less, what's the minimum sample size you should use?

Solution

Remember that the margin of error is

$$\frac{1}{\sqrt{n}} \times 100\%$$

Thus, for a margin of error of 2%:

$$2\% = \frac{1}{\sqrt{n}} \times 100\%$$

If we divide both sides by 100% (i.e. convert percentages to decimals on both sides of the equation), we get

$$0.02 = \frac{1}{\sqrt{n}}$$

Now, since the unknown we're solving for is in the denominator, we'll multiply both sides by \sqrt{n} , and then divide both sides by 0.02 (it turns out this is equivalent to flipping both sides upside down):

$$\begin{aligned} 0.02\sqrt{n} &= 1 \\ \sqrt{n} &= \frac{1}{0.02} \\ &= 50 \end{aligned}$$

Finally, to get rid of the square root, we simply need to square both sides:

$$\begin{aligned} \sqrt{n} &= 50 \\ n &= 50^2 \\ &= \boxed{2500} \end{aligned}$$

In order for a poll to have a margin of error of 2% or less, at least 2500 people need to be polled. Once again, the larger the sample, the smaller the margin of error.

TRY IT

If you want a poll to have a margin of error of 4.5% or less, what's the minimum sample size you should use?

Exercises 3.5

- The heights of American adult males are normally distributed with a mean of 177 cm and a standard deviation of 7.4 cm. Find the range of heights that contain approximately
 - 68% of the data
 - 95% of the data
 - 99.7% of the data
- Suppose that babies' weights are normally distributed with a mean of 3.23 kg and a standard deviation of 0.87 kg. Find the range of weights that contain approximately
 - 68% of the data
 - 95% of the data
 - 99.7% of the data
- Suppose that the scores on a statewide standardized test are normally distributed with a mean of 72 and a standard deviation of 4. Estimate the percentage of scores that were
 - between 68 and 76.
 - above 76.
 - below 64.
 - between 68 and 84.
- GMAT scores are approximately normally distributed with a mean of 547 and a standard deviation of 95. Estimate the percentage of scores that were
 - between 262 and 832.
 - above 642.
 - below 262.
 - between 262 and 452.
- The selling prices for homes in a certain community are normally distributed with a mean of \$321,000 and a standard deviation of \$38,000. Estimate the percentage of homes in this community with selling prices
 - between \$283,000 and \$397,000.
 - between \$245,000 and \$435,000.
 - below \$245,000.
 - above \$245,000.
- The widths of platinum samples manufactured at a factory are normally distributed, with a mean of 1.1 cm and a standard deviation of 0.2 cm. Find the z -scores that correspond to each of the following widths.
 - 1.5 cm
 - 0.94 cm
- The average height of American adult males is 177 cm, with a standard deviation of 7.4 cm. Meanwhile, the average height of Indian males is 165 cm, with a standard deviation of 6.7 cm. Which is taller relative to his nationality, a 175-cm American man or a 162-cm Indian man?
- Water usages in American showers are normally distributed, with the average shower using 17.2 gallons, and a standard deviation of 2.5 gallons. Estimate the percentage of showers that used
 - more than 22.2 gallons.
 - less than 14.7 gallons.
 - between 12.2 and 22.2 gallons.
 - between 9.7 and 19.7 gallons.
- Suppose that wedding costs in the Caribbean are normally distributed with a mean of \$7,500 and a standard deviation of \$975. Estimate the percentage of Caribbean weddings that cost
 - between \$6525 and \$9450.
 - above \$9450.
 - below \$6525.
 - between \$4575 and \$10,425.
- Suppose that the time that a new roof will last before needing to be replaced follows a normal distribution, with a mean of 25 years and a standard deviation of 5 years. Estimate the percentage of roofs that last
 - longer than 30 years.
 - between 20 and 40 years.
 - less than 20 years.
 - more than 40 years.
- The average resting heart rate of a population is 88 beats per minute, with a standard deviation of 12 bpm. Find the z -scores that correspond to each of the following heart rates.
 - 120 bpm
 - 71 bpm
- Kyle and Ryan take entrance exams at two different universities. Kyle scores a 430 on an exam with a mean of 385 and a standard deviation of 70, while Ryan scores a 31 on an exam with a mean of 28 and a standard deviation of 4.5. Which do you think is more likely to be accepted at their university of choice?

13. A doctor measured serum HDL levels in her patients, and found that they were normally distributed with a mean of 63.4 and a standard deviation of 3.8. Find the serum HDL levels that correspond to the following z -scores.

(a) $z = -0.85$

(b) $z = 1.33$

15. What is the margin of error for a poll with a sample size of 2000 people?

17. If you want a poll to have a margin of error of 2.5%, how large will your sample have to be?

14. If the distribution of weight of newborn babies is approximately normal, with a mean of 3.23 kilograms and a standard deviation of 0.87 kilograms, find the weights that correspond to the following z -scores.

(a) $z = 2.20$

(b) $z = -1.73$

16. What is the margin of error for a poll with a sample size of 150 people?

18. If you want a poll to have a margin of error of 1%, how large will your sample have to be?