



# Probability & Statistics Notes

---

# One-way tables

## Individuals and variables

The simplest kind of information we'll work with in this course is a set of **individuals** with one or more properties, called **variables**. The individuals are the items in the data set and can be cases, things, people, etc.

In this table for “heights of people in the car,” the individuals are the people, and their heights are a variable because height is a property of each individual.

Person	Height
Wendy	5'6"
Michael	5'9"
Rachael	5'3"
Allen	5'11"

Together, individuals and variables are called **data**. When we organize data into a table like this one, we call it a **data table**. Here's an example of a data table about ice cream:



Flavor	Scoops sold	Contains chocolate?	Smooth or chunky?
Vanilla	300	No	Smooth
Chocolate	450	Yes	Smooth
Cookies & Cream	275	Yes	Chunky
Mint Chocolate Chip	315	Yes	Chunky
Fudge Brownie	375	Yes	Chunky
Rocky Road	250	Yes	Chunky

The individuals are the flavors: Vanilla, Chocolate, Cookies & Cream, etc., and the variables are their properties: Scoops sold, Contains chocolate, and Smooth or chunky.

Variables can be **categorical** or **quantitative**. In the table for “heights of people in the car,” there’s one quantitative variable: the height. In the table for “ice cream shop data for July,” there’s one quantitative variable (scoops), and two categorical variables (contains chocolate and smooth or chunky).

- **Categorical variables** are non-numerical variables. Categorical variables are also called “qualitative” variables. Their values aren’t represented with numbers. Whether or not the ice cream contains chocolate, and whether the ice cream is smooth or chunky are categorical variables. This is because the values there are words, not numbers.
- **Quantitative variables** are numerical variables. Their values are numbers. The height of the people in the car and the number of

scoops of ice cream sold are quantitative variables, because the values there are numbers.

We can divide quantitative variables into **discrete variables** and **continuous variables**.

- **Discrete variables** are those that we can obtain by counting. Therefore, they can take on only certain numerical values. For example, the number of scoops of ice cream sold is a discrete quantitative variable, because we can't really sell 8.3 or 5.23 scoops of ice cream.
- On the other hand, **continuous variables** may include data as decimals, fractions, or irrational numbers. For example, the height of the people in the car is a continuous variable.

## Levels of measurement

When we work with data, we need to understand the level of measurement, because not every statistical test can be used with every type of data set. There are four levels of data measurement: nominal, ordinal, interval, and ratio.

- Things like favorite food, colors, names, and “yes” or “no” responses have a **nominal** scale of measurement. Only categorical data can be measured with a nominal scale.
- Categorical data can also be **ordinal**. This type of data can be ordered. The top three national parks in California is an example of



ordinal data. Or, for example, when we ask a group of people about how they liked their trip, we may get responses like “awesome,” “good,” “satisfactory,” or “terrible,” which follow an order from best to worst.

- Data measured using an **interval** scale can be ordered like ordinal data. But interval data also gives us a known interval between measurements. For example, temperature is measured using an interval scale, because we can understand the exact interval of difference between 50 and 60 degrees.
- Data measured using a **ratio** scale is just like interval scale data, except that ratio scale data has a starting point, or absolute zero. Whereas interval data like temperature can have negative and positive values, things like time, height, and weight are examples of ratio scale data because those measures can never be negative.

When we have a single individual and a categorical or quantitative variable assigned to it, we can construct a one-way table. These kinds of tables are called “one-way” because they usually represent data for one individual only.

## Constructing one-way tables

When we construct a table, we want to think about whether we have more individuals or more variables. In the two tables we looked at, we had the individuals listed down the left-hand side, and the variables listed across the top.



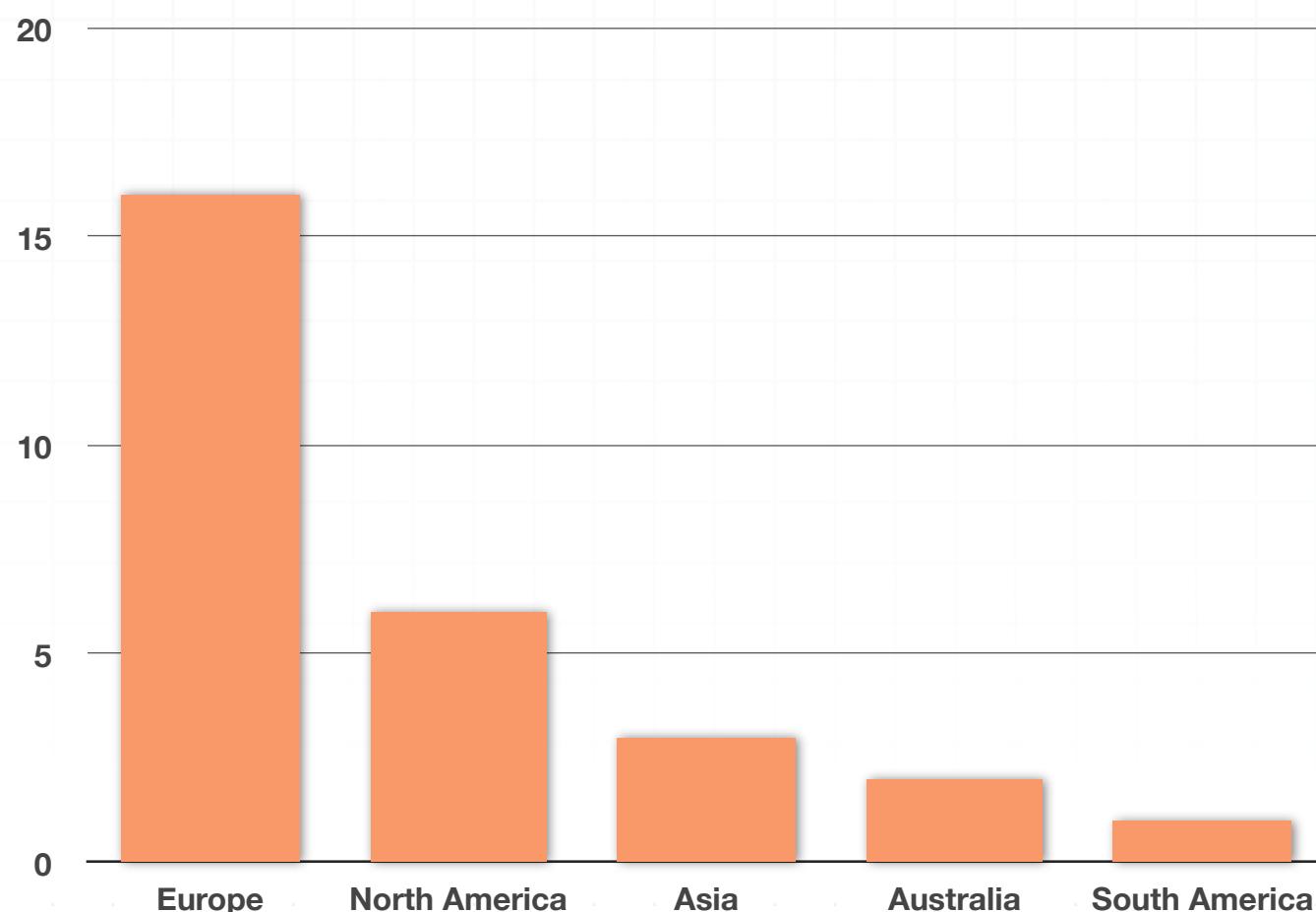
But if we have lots of variables but only a few individuals, it can be helpful to list the individuals across the top and the variables down the left side. For example, if we're comparing lots of information about two houses for sale in the same neighborhood, we might make a table like this:

	317 Spruce Rd	819 Lilac St
Price	\$299,000	\$349,000
Square footage	3652	3812
Price per square foot	\$82	\$92
Bedrooms	4	5
Bathrooms	2.5	3
Stories	3	2
Basement	Finished	Unfinished
Garage spaces	3	2
Lot acres	0.36	0.31
Grass backyard	Yes	yes
Year built	1974	2001
Property tax	\$2,356	\$2,595
Payment	\$1,120	\$1,045

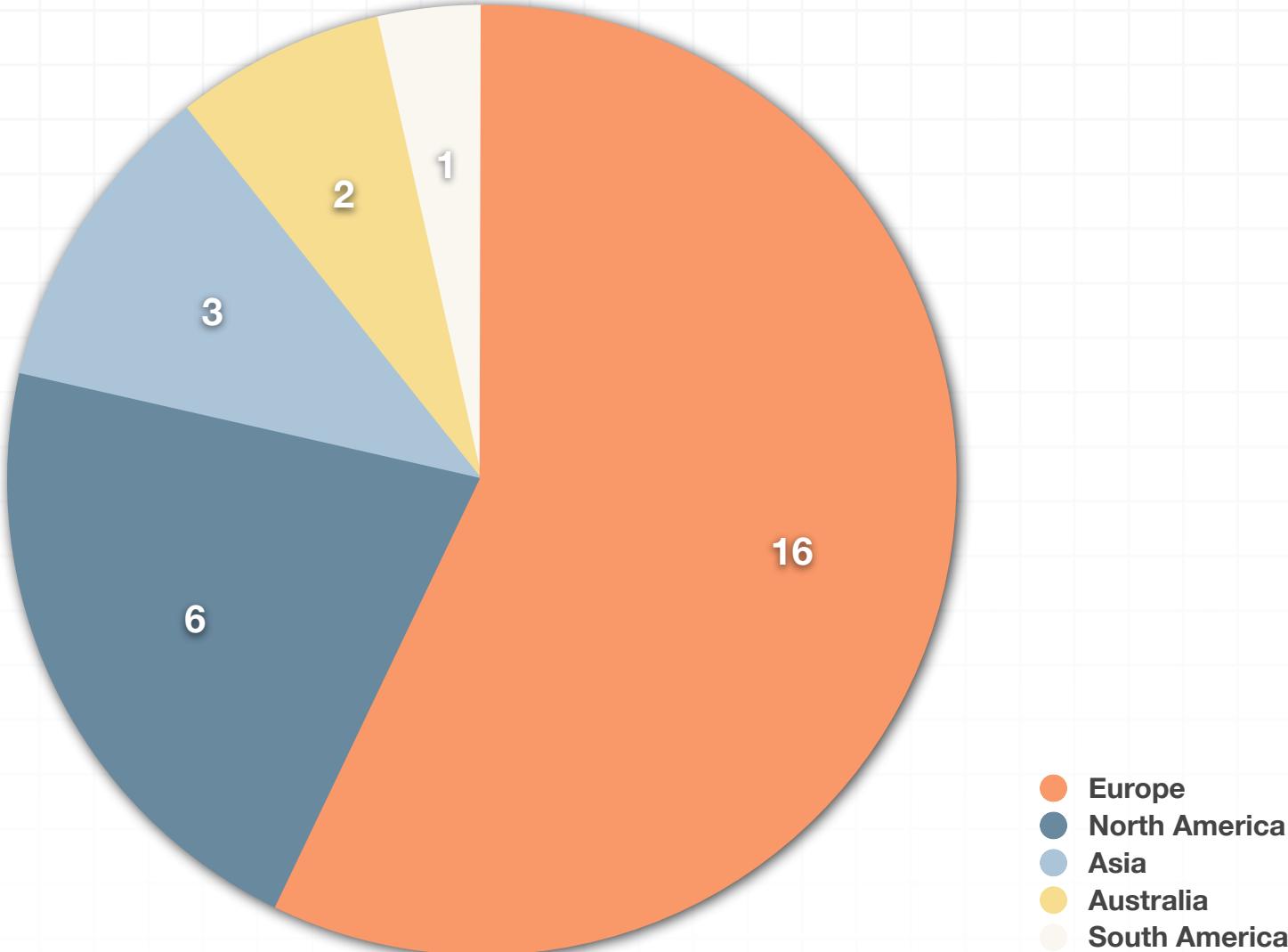
The individuals are the addresses listed across the top of the table, 317 Spruce Road and 819 Lilac Street. All the variables are listed down the left side of the table. Since there are so many more variables than individuals, listing the variables vertically makes the table fit better on paper than if we had tried to list all the variables horizontally.

# Bar graphs and pie charts

Bar graphs and pie charts are two of the simplest ways to summarize and represent data. In general, a **bar graph**, also called a **bar chart**, usually looks something like this:



And a **pie chart** usually looks something like this:



In this lesson, we'll start with data tables like the ones we looked at in the last lesson, and try to represent the data given in the tables in bar graphs and pie charts.

## Building data tables

Let's say we're given historical information about the host cities for the summer Olympic games, and we want to summarize this information into a simple data table.

Here is a list of host cities for the summer games, not including host cities for canceled games, from 1896 through 2016.

Games	Year	City, Country	Continent
I	1896	Athens, Greece	Europe
II	1900	Paris, France	Europe
III	1904	St. Louis, United States	North America
IV	1908	London, United Kingdom	Europe
V	1912	Stockholm, Sweden	Europe
VII	1920	Antwerp, Belgium	Europe
VIII	1924	Paris, France	Europe
IX	1928	Amsterdam, Netherlands	Europe
X	1932	Los Angeles, United States	North America
XI	1936	Berlin, Germany	Europe
XIV	1948	London, United Kingdom	Europe
XV	1952	Helsinki, Finland	Europe
XVI	1956	Melbourne, Australia	Australia
XVII	1960	Rome, Italy	Europe
XVIII	1964	Tokyo, Japan	Asia
XIX	1968	Mexico City, Mexico	North America
XX	1972	Munich, West Germany	Europe
XXI	1976	Montreal, Canada	North America
XXII	1980	Moscow, Soviet Union	Europe
XXIII	1984	Los Angeles, United States	North America
XXIV	1988	Seoul, South Korea	Asia
XXV	1992	Barcelona, Spain	Europe
XXVI	1996	Atlanta, United States	North America
XXVII	2000	Sydney, Australia	Australia
XXVIII	2004	Athens, Greece	Europe
XXIX	2008	Beijing, China	Asia
XXX	2012	London, United Kingdom	Europe



If we wanted to make a data table showing the number of times each continent has hosted the summer games, we could count this number for each continent from the data table, and create a summary table for count by continent:

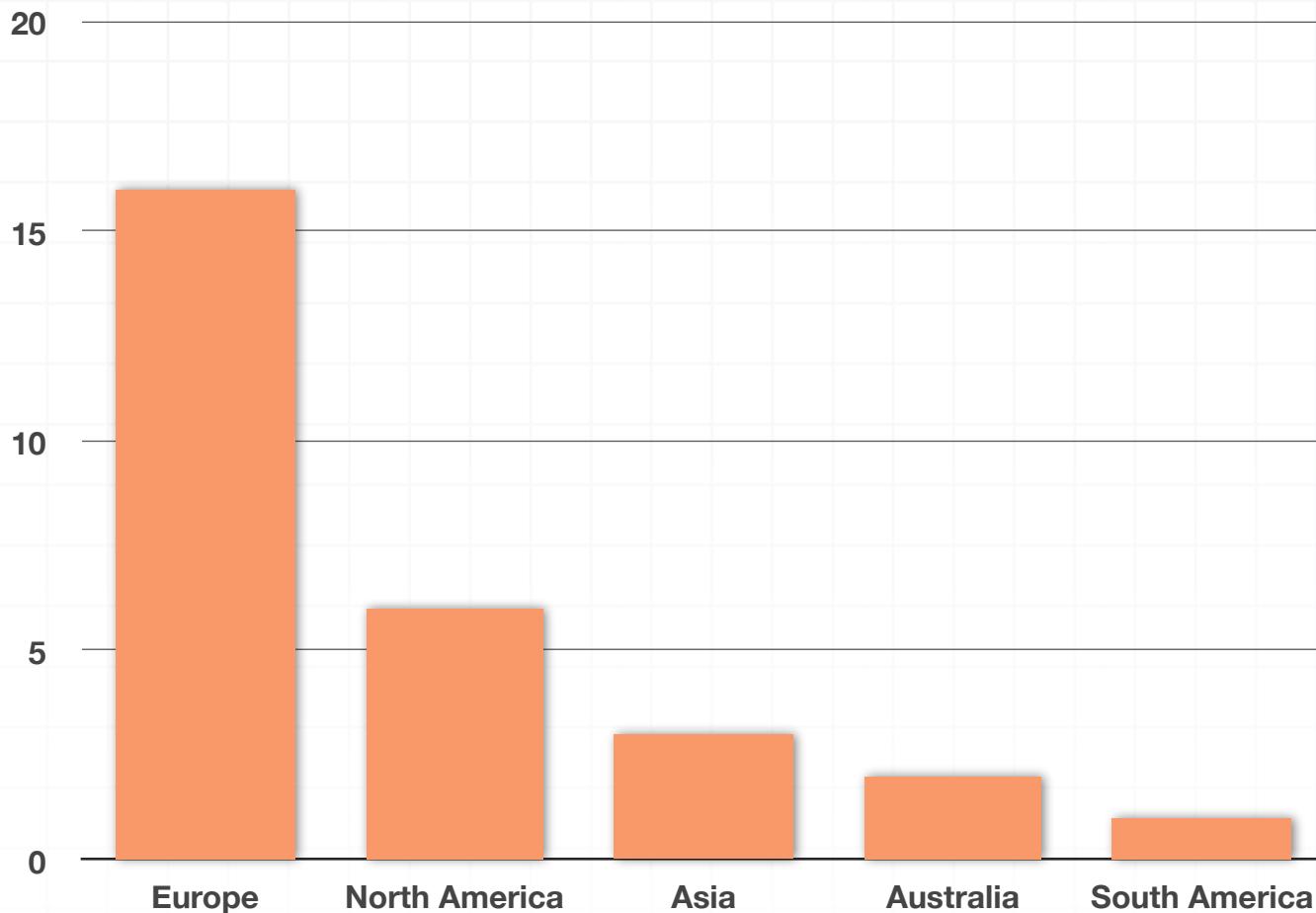
Continent	Count
Europe	16
North America	6
Asia	3
Australia	2
South America	1

The summary table is often called a **frequency table**, which shows the frequency or count of each individual. In the list of host cities of summer games, the individual, Europe, appeared 16 times, which is why the count, or frequency is 16.

## Building bar graphs

If we wanted to express the count of summer games by continent in a bar graph, it might look like this:



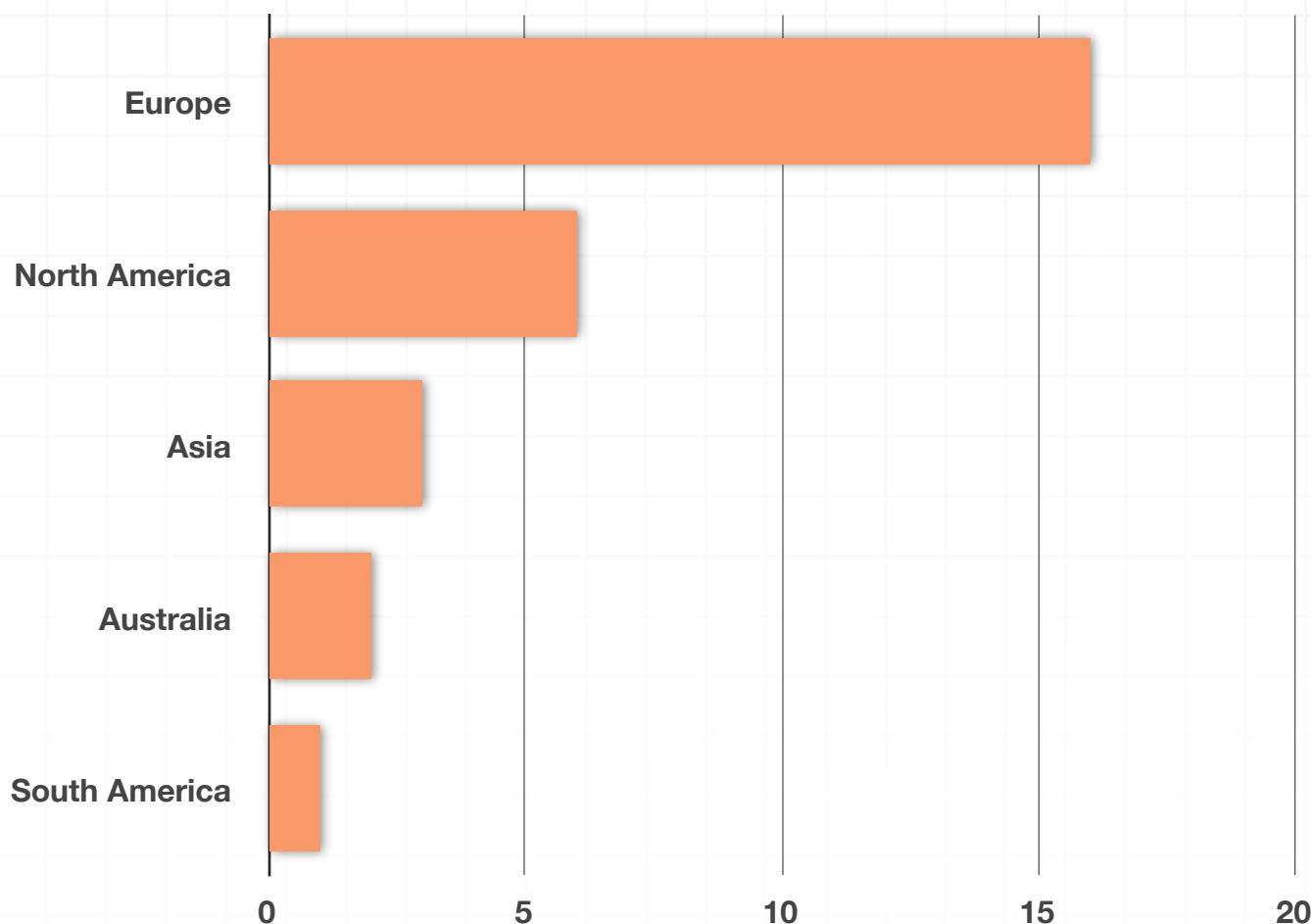


Notice that we have a list of the continents across the bottom of the bar graph, with the count of the number of times they've hosted the summer games up the left side. The continents are the **individuals**, and the count is a **quantitative variable**, because the count is a numeric property of each of the individuals.

The bar graph is a nice way to represent this data, because we can quickly get a visual picture of which continents have hosted the summer games most often.

Now we can quickly see that Europe has hosted more summer games by far than any other continent, North America has hosted the second-most number, and South America has hosted the summer games the fewest number of times. With this particular data set, since we know there are 7 continents, we could infer from the graph that Africa and Antarctica have never hosted the summer games.

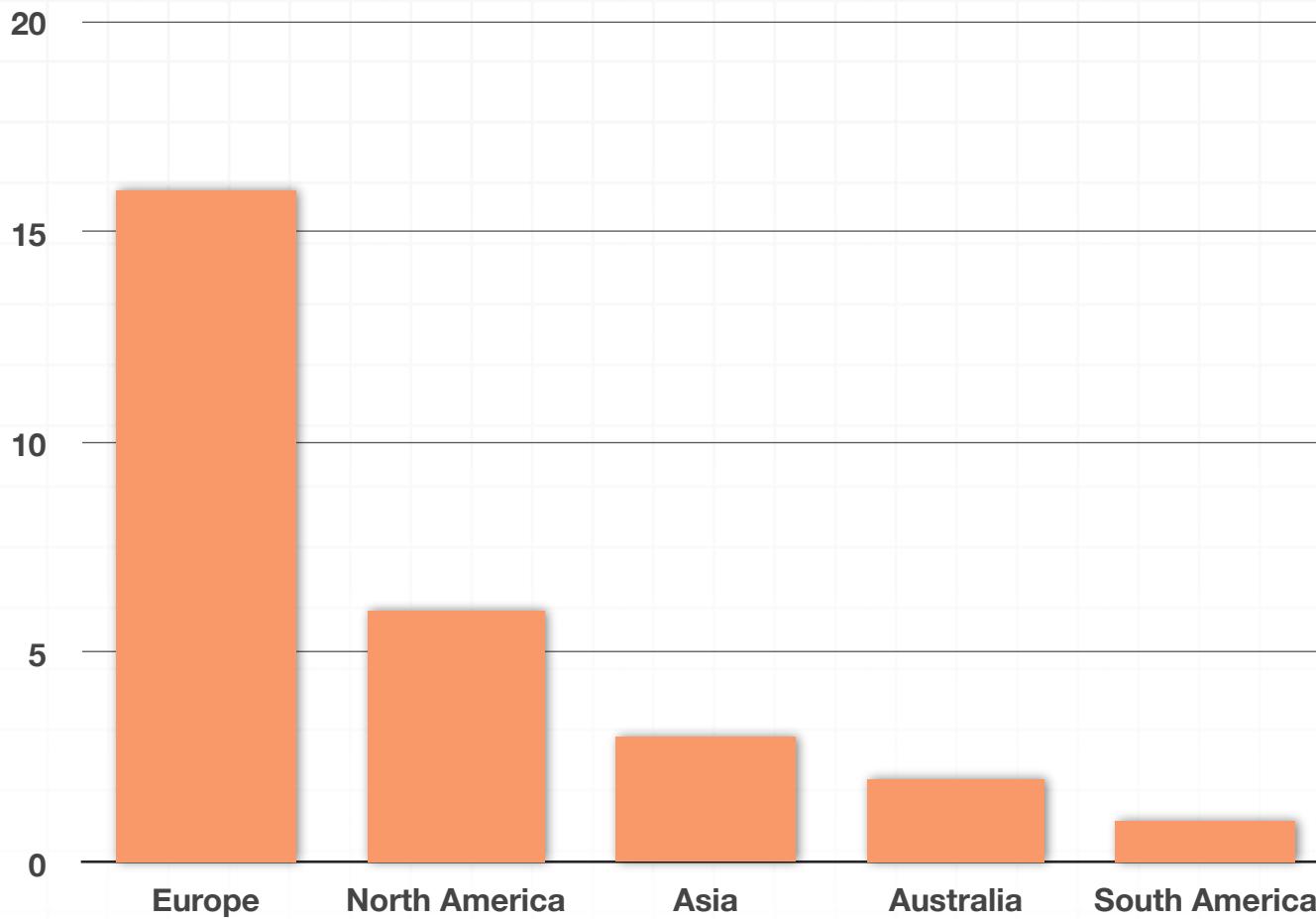
Bar graphs can also be built horizontally.



When we build a vertical bar graph, it's common to sort the data largest to smallest, so that the tallest bars appear on the left, in descending order down to the smallest bars on the right. When we build a horizontal bar graph, it's common to put the largest bars at the top and the smallest bars at the bottom.

## Reading bar graphs

If we only have the bar graph, and no data table to work with, we may only be able to get approximate values from the bar graph. Using this bar graph again,



we see that the vertical axis isn't marked off at every increment, only at every increment of 5. So based on how far up the bar extends for Europe, for example, we only know with absolute certainty that Europe has hosted between 15 and 20 times. We could probably guess that they've hosted about 16 or 17 summer games, but we might not feel absolutely sure.

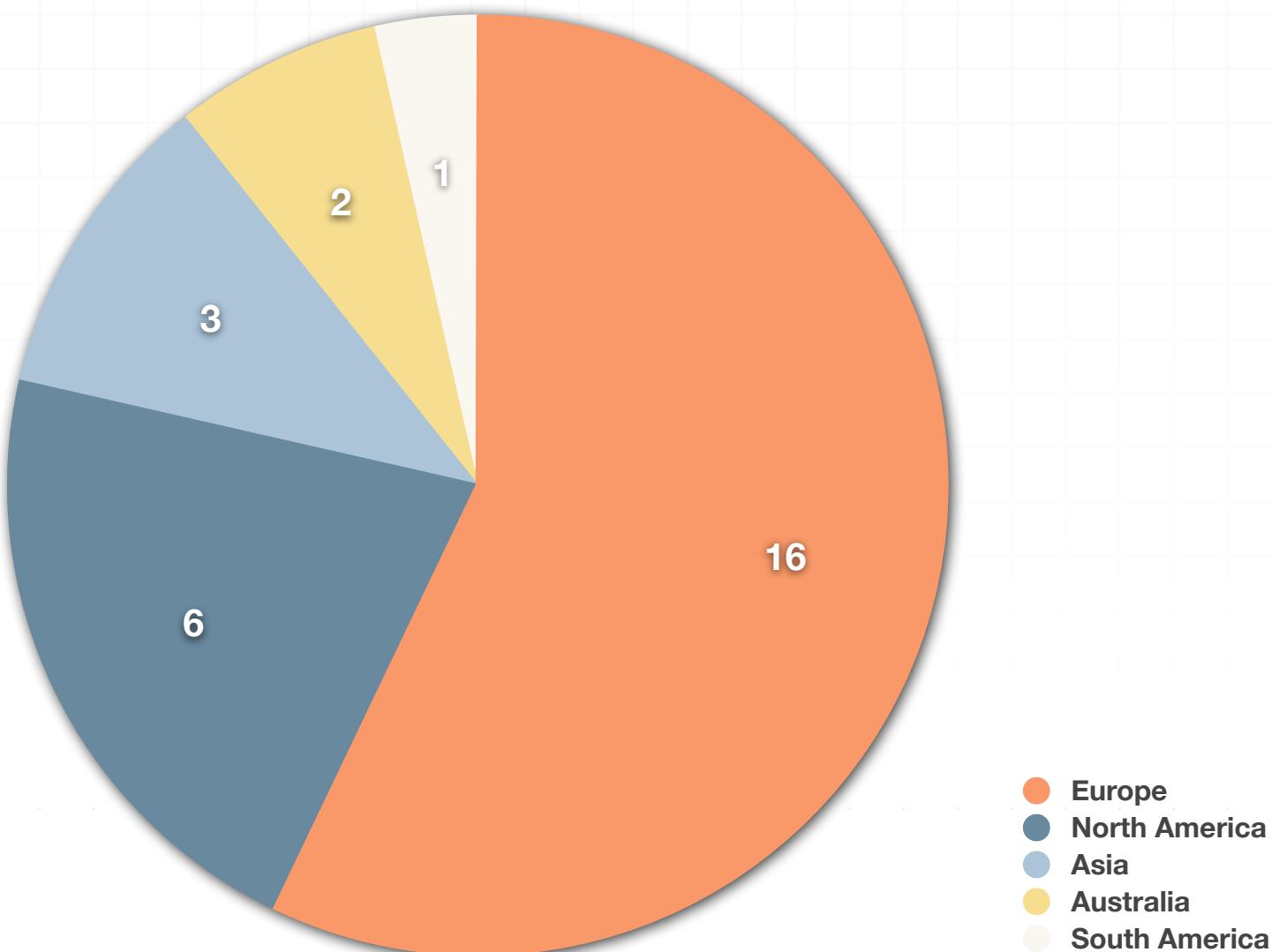
At quick glance we would know that Australia has probably hosted only 2 or 3 times.

It's worth making the point that bar graphs aren't always great at expressing exact values, but they're excellent at giving us a quick visual picture of data.

## Building pie charts

This same data can be displayed in a pie chart. It's helpful in a pie chart, though not necessary, to make each individual a different color so that it's easy to see the distinction between sections.

Like the bar graphs, it's customary to put the largest slices next to each other, in order all the way down to the smallest sections.



### Example

Create a bar graph and pie chart that shows the number of times each continent has hosted the winter Olympic games. Use the data table to first create a summary table, then build the bar graph and pie chart.

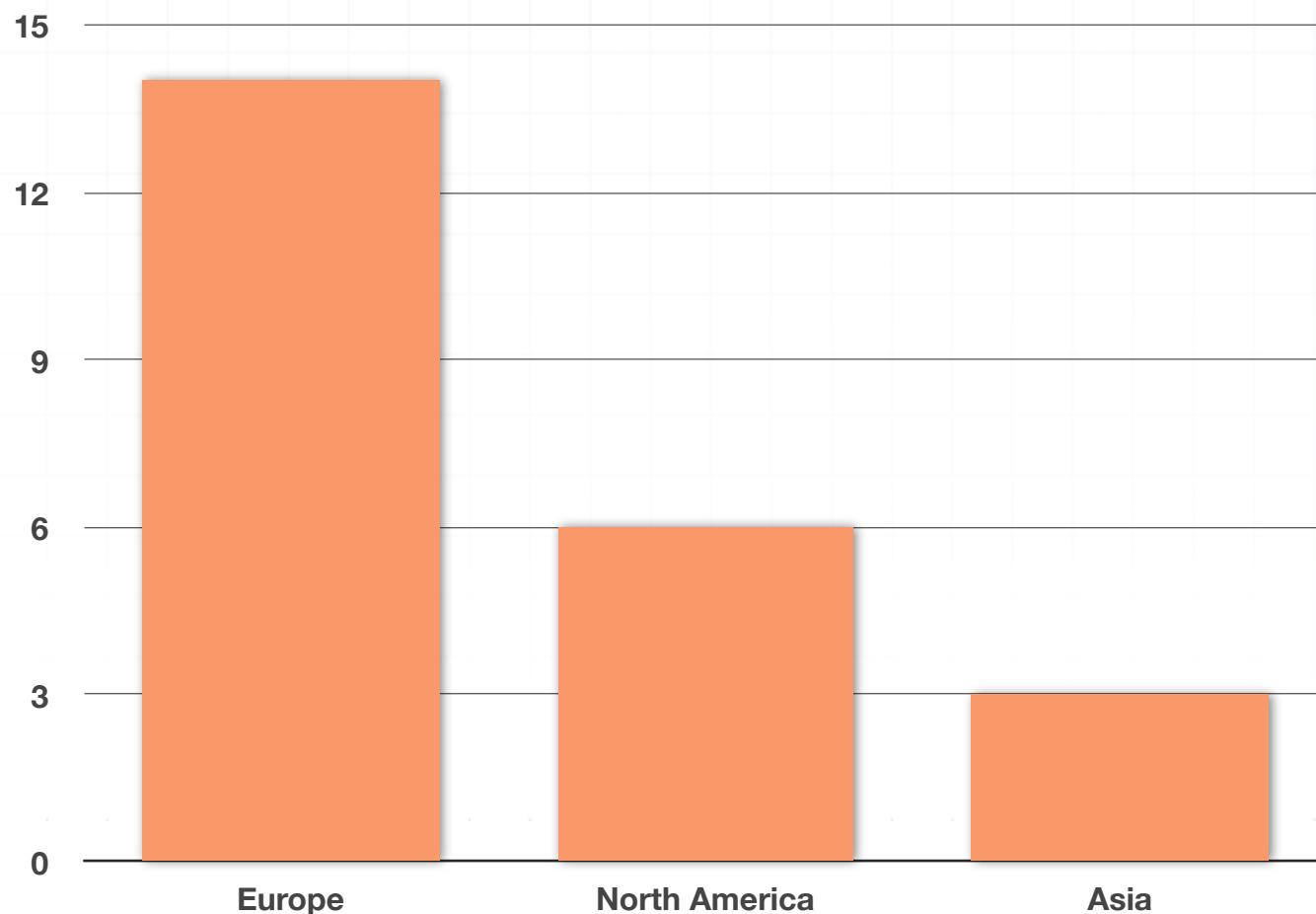
**Host cities for the winter Olympic games, not including host cities for canceled games, from 1924 through 2018**

<b>Games</b>	<b>Year</b>	<b>City, Country</b>	<b>Continent</b>
I	1924	Chamonix, France	Europe
II	1928	St. Moritz, Switzerland	Europe
III	1932	Lake Placid, United States	North America
IV	1936	Garmisch-Partenkirchen, Germany	Europe
V	1948	St. Moritz, Switzerland	Europe
VI	1952	Oslo, Norway	Europe
VII	1956	Cortina d'Ampezzo, Italy	Europe
VIII	1960	Squaw Valley, United States	North America
IX	1964	Innsbruck, Austria	Europe
X	1968	Grenoble, France	Europe
XI	1972	Sapporo, Japan	Asia
XII	1976	Innsbruck, Austria	Europe
XIII	1980	Lake Placid, United States	North America
XIV	1984	Sarajevo, Yugoslavia	Europe
XV	1988	Calgary, Canada	North America
XVI	1992	Albertville, France	Europe
XVII	1994	Lillehammer, Norway	Europe
XVIII	1998	Nagano, Japan	Asia
XIX	2002	Salt Lake City, United States	North America
XX	2006	Turin, Italy	Europe
XXI	2010	Vancouver, Canada	North America
XXII	2014	Sochi, Russia	Europe
XXIII	2018	Pyeongchang, South Korea	Asia

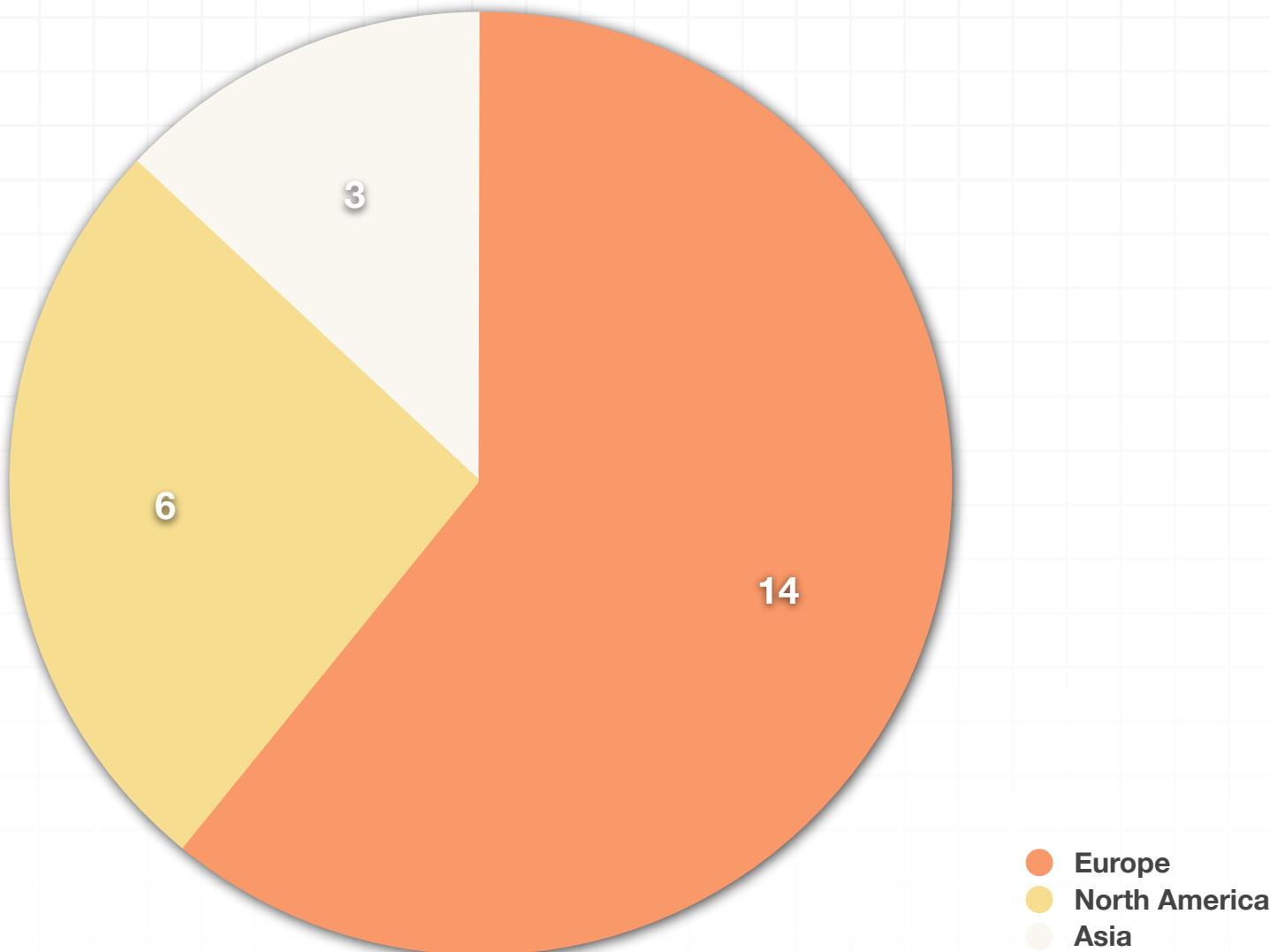
If we try to make a frequency table first of this information, we get

Continent	Count
Europe	14
North America	6
Asia	3

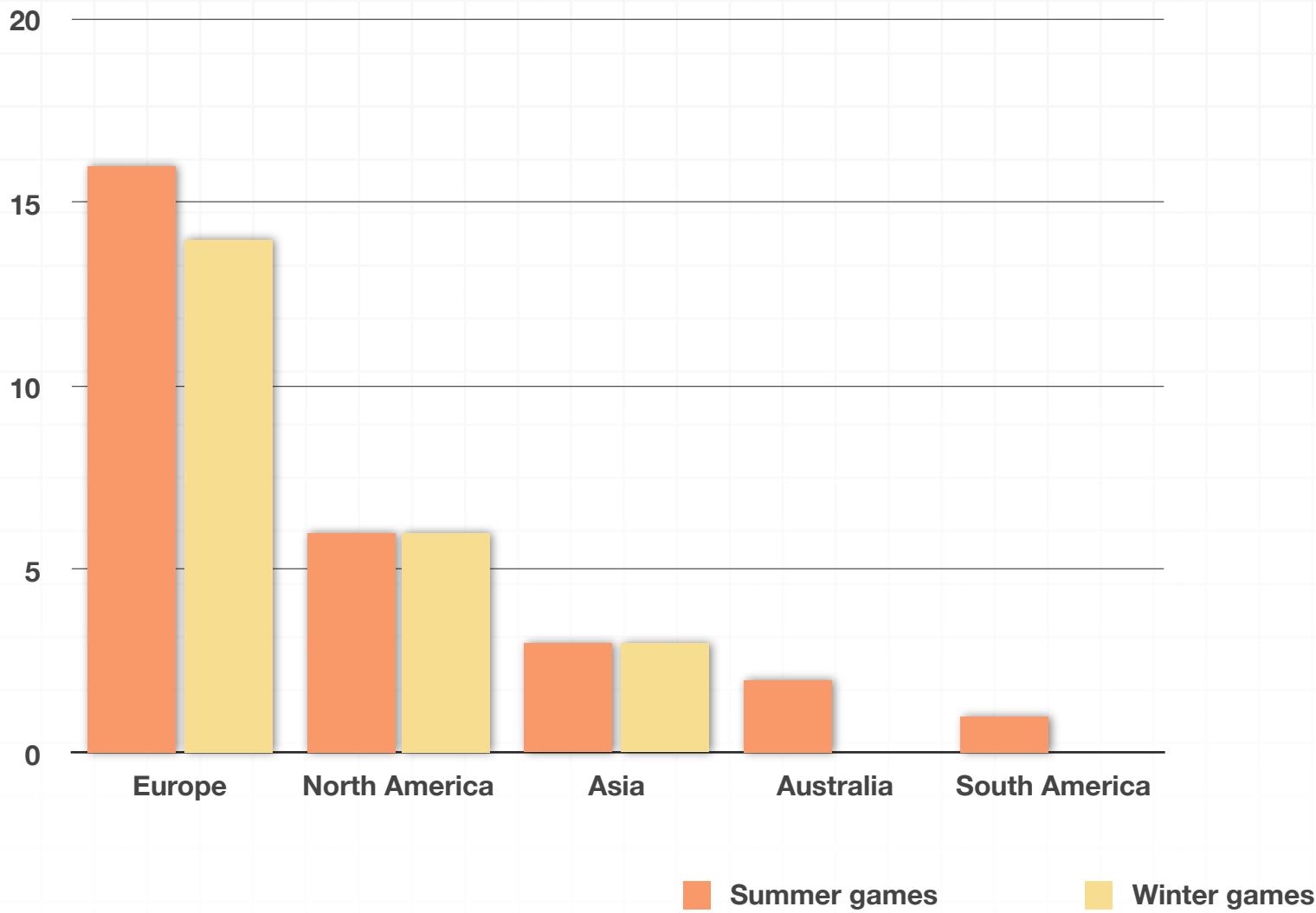
Then we can make a bar graph,



or a pie chart.



Bar graphs are also great for showing multiple variables for the same individuals, side by side. Now that we've created bar graphs for host continents for both the summer and winter Olympic games, let's bring them together:



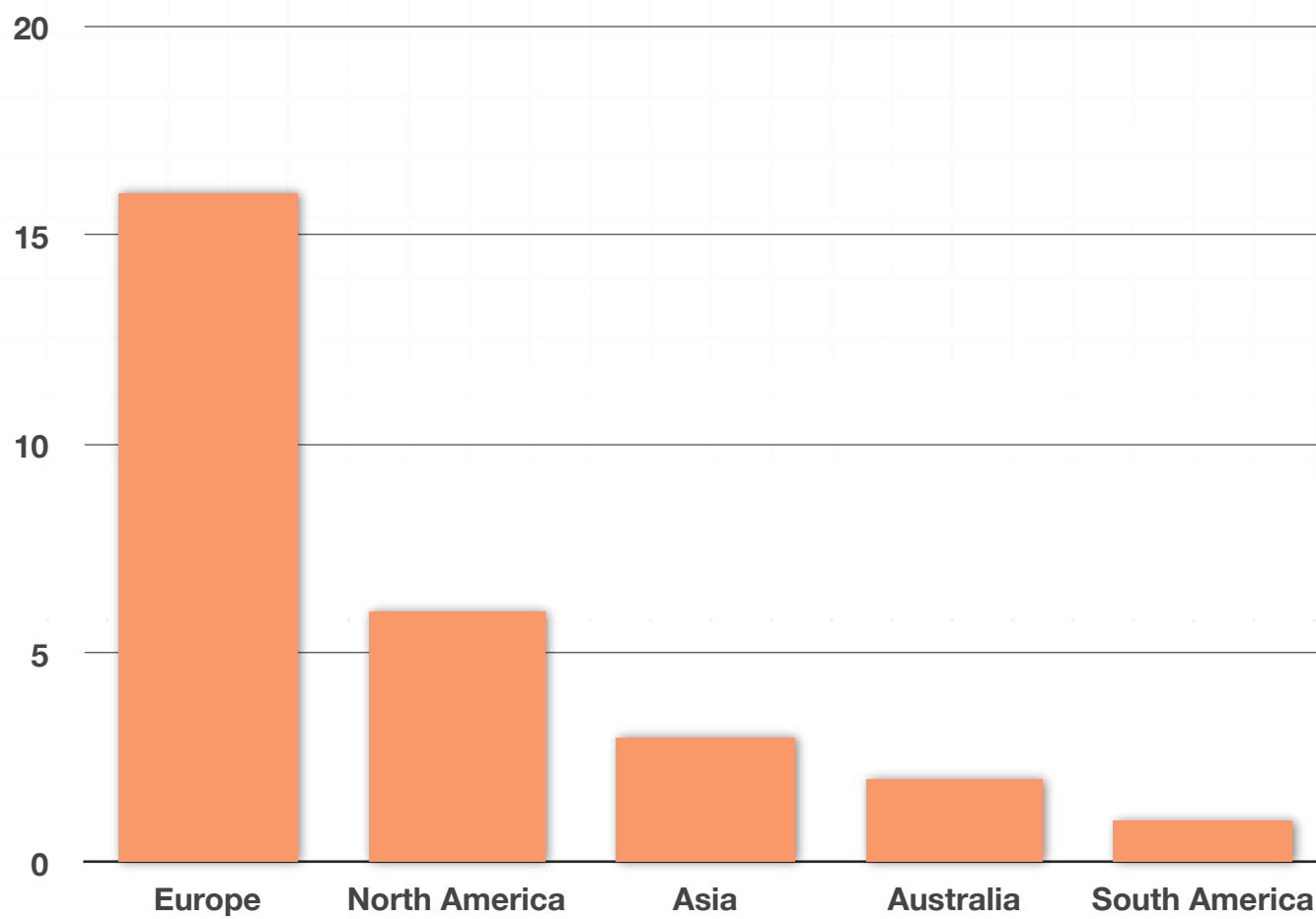
This kind of side-by-side bar graph allows us to quickly see what we already knew from the previous bar graphs, like the fact that Europe has hosted more summer games and more winter games than any other continent.

But we get more information from this, too, like the fact that Europe has hosted more summer games than it has winter games, or that North America has hosted an equal number of summer and winter games.

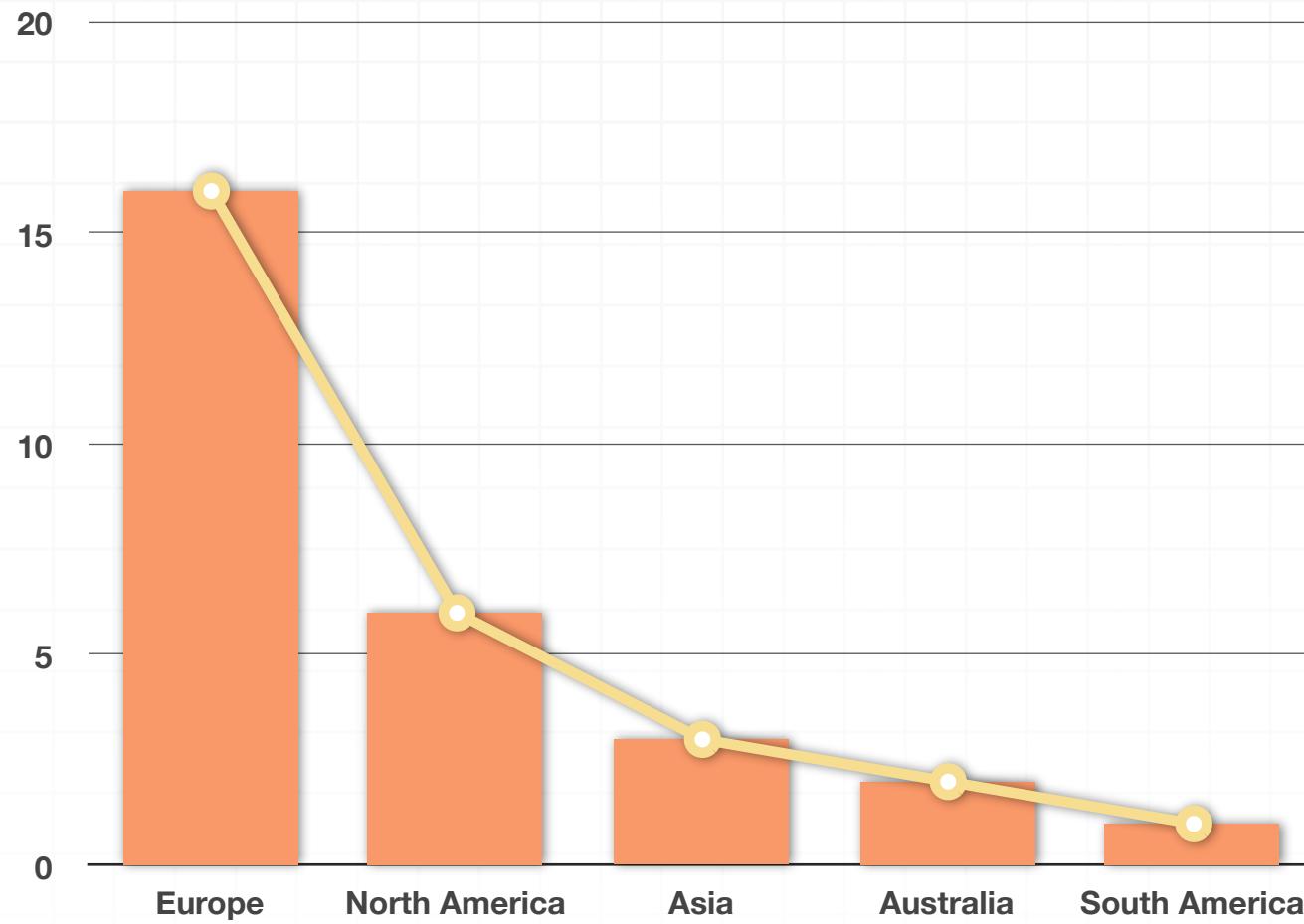
# Line graphs and ogives

**Line graphs** are really similar to bar graphs. In fact, to turn a bar graph into a line graph, all we have to do is connect the middle of the top of each bar to the middle of the top of the bar beside it with a straight line, and we'll form the line graph.

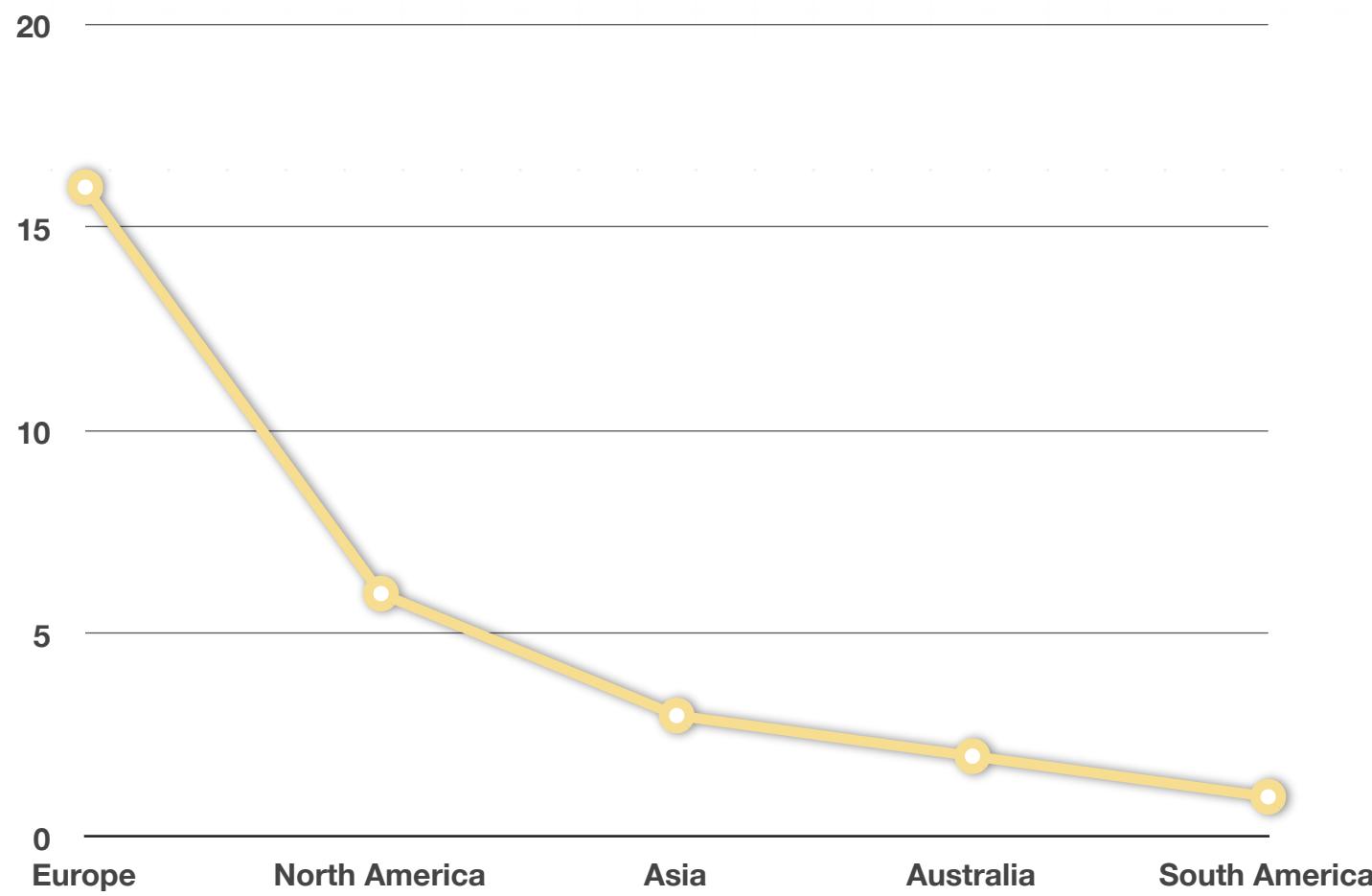
For example, we could take a bar graph for the summer Olympics from the previous section.



Then we can use straight lines to connect the top of each bar.



Then we can remove the bars, and we have a line graph that represents the same data.

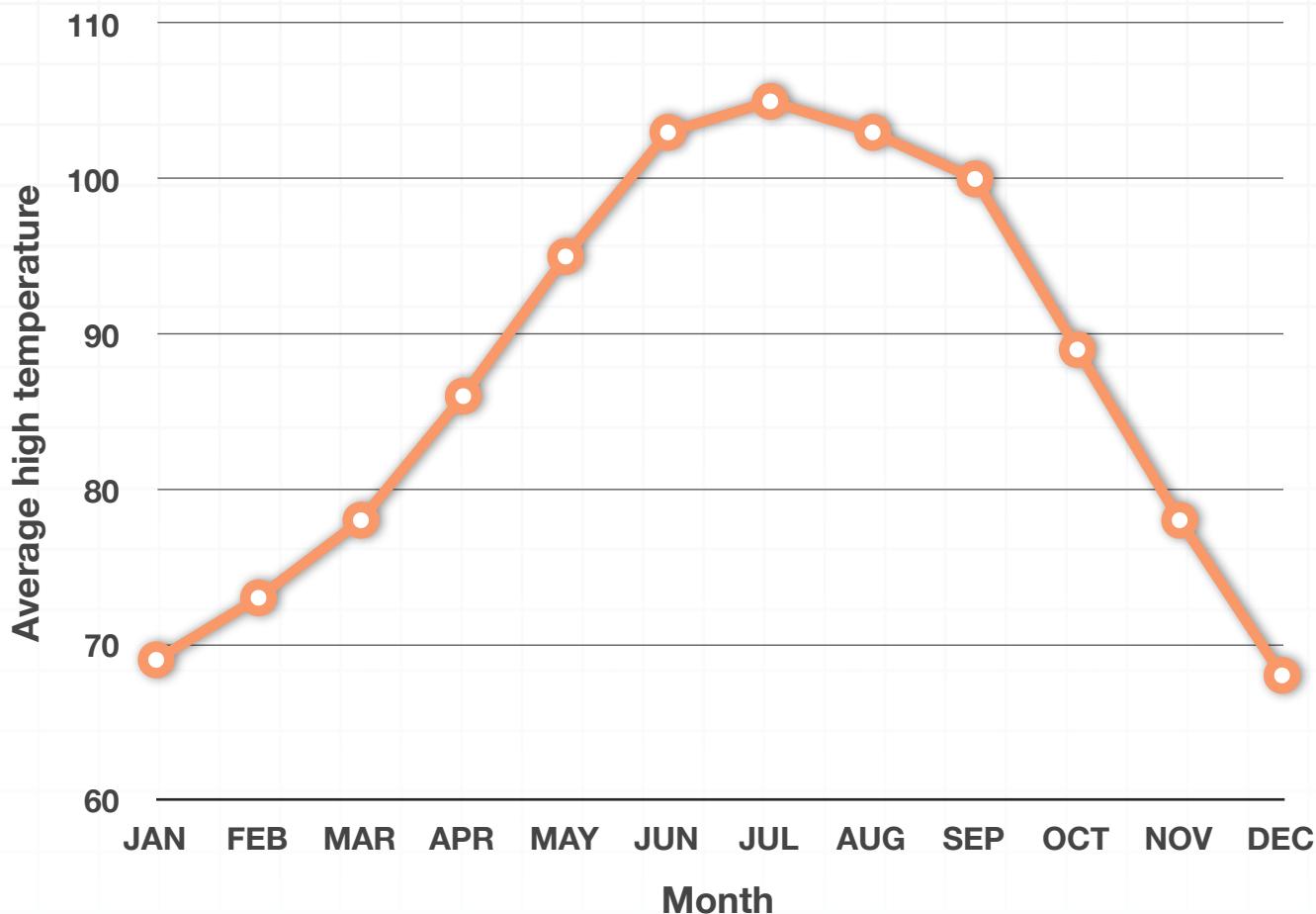


While this example is helpful for illustrating how we can transition from a bar graph to a line graph, and how the two graph types are related, in actuality, a line graph is a bad choice for this data. Line graphs are used to show changes over time or a connection between individuals, but this data is comparing things between different groups. That means the bar graph is a better choice than a line graph for this data in particular.

Let's look at an example with data that's much better suited for a line graph, like the monthly average high temperature in Scottsdale, Arizona over the course of a year. The data is

Month	Average high
JAN	69
FEB	73
MAR	78
APR	86
MAY	95
JUN	103
JUL	105
AUG	103
SEP	100
OCT	89
NOV	78
DEC	68

The historical average high temperature in degrees Fahrenheit is given for each month. If we plot these as a line graph, we get



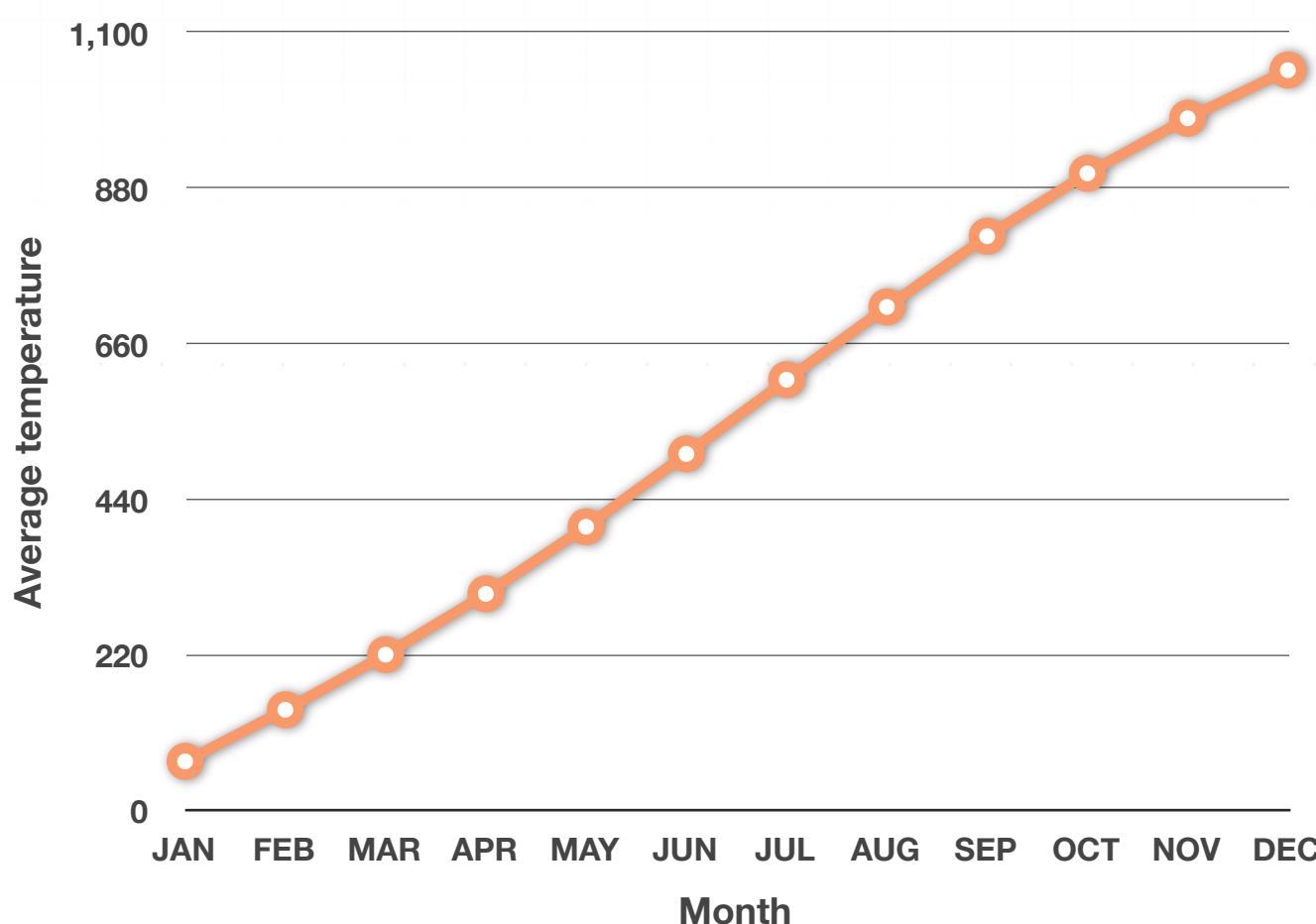
Notice that the slope of each line segment gives us an idea of how much change occurred between two data points. For example, there's a much larger change in average high temperature between September and October than there is between August and September, because the slope of the line connecting September to October is much steeper, indicating a larger decrease in temperature, than the slope of the line connecting August to September, which is much flatter, and therefore indicates a smaller decrease in temperature.

A line graph is a fantastic way to display this kind of data, mostly because it shows clearly how a data point changes over time. We inherently understand that average high temperature fluctuates over the course of a year, with (in the northern hemisphere) the highest temperatures usually occurring in the summer months, and the lowest temperatures usually occurring the winter months.

## Ogives

An **ogive** is a special kind of line graph. This kind of graph looks just like a line graph, but think of an ogive as an “accumulated” line graph. Just like other types of graphs, an ogive does well at representing some kinds of data, and less well at representing others.

For example, it doesn’t make sense to show accumulated temperature over time. Taking the average monthly temperature in Scottsdale, Arizona, if we use an ogive, we would show the temperature in January as 69, the temperature in February as  $69 + 73$ , the temperature in March as  $69 + 73 + 78$  etc. The ogive would look like this:



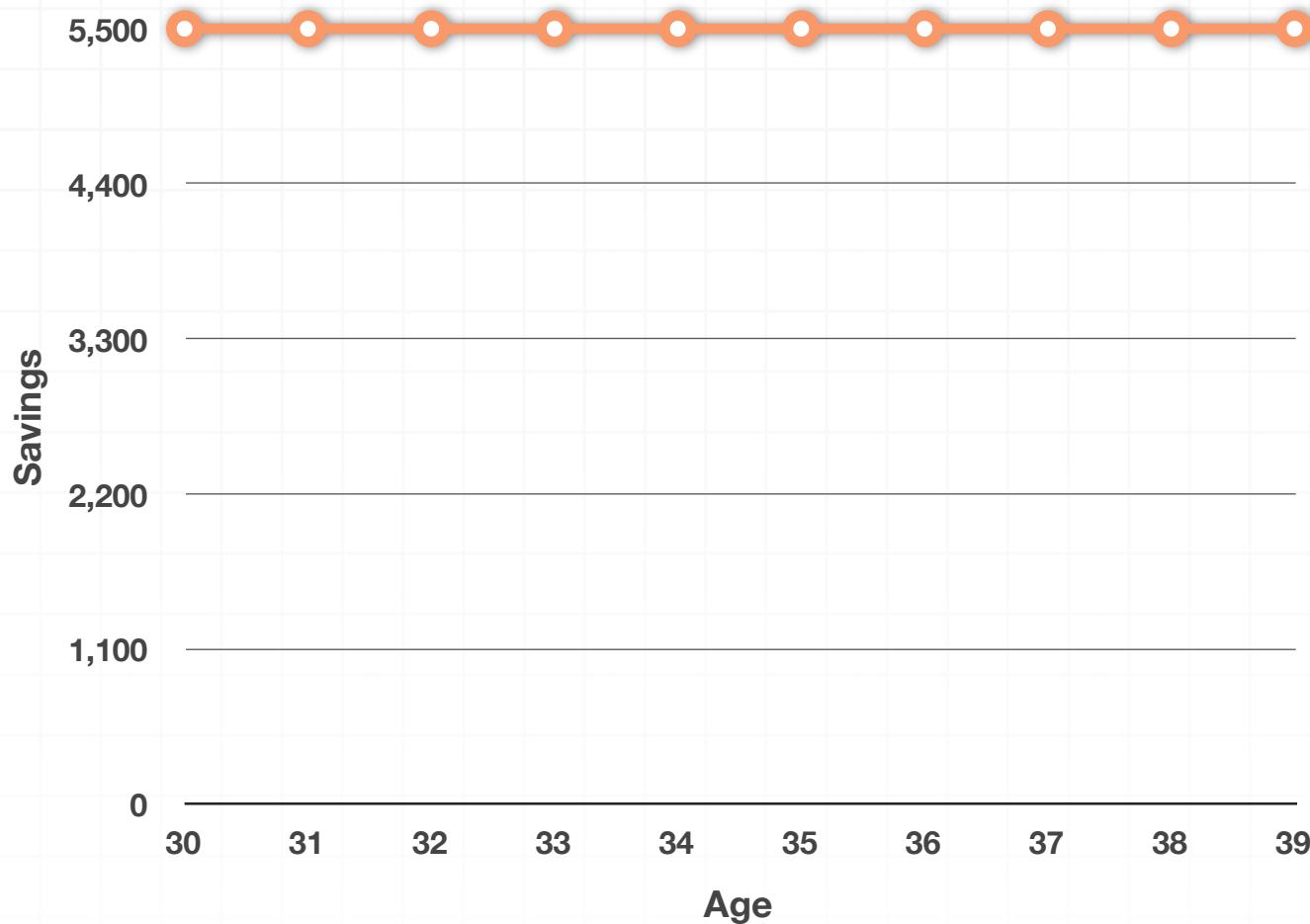
But it doesn't make sense to have a temperature of almost 1,100 degrees in December. Temperature is not a very good data point to represent with an ogive.

But an ogive would be perfect for showing something like accumulated savings over time. For example, let's say we invest \$5,500 each year into a Roth IRA retirement savings account, from age 30 until age 39.

Age	Investment
30	\$5,500
31	\$5,500
32	\$5,500
33	\$5,500
34	\$5,500
35	\$5,500
36	\$5,500
37	\$5,500
38	\$5,500
39	\$5,500

Graphing the invested amount each year in a line graph gives

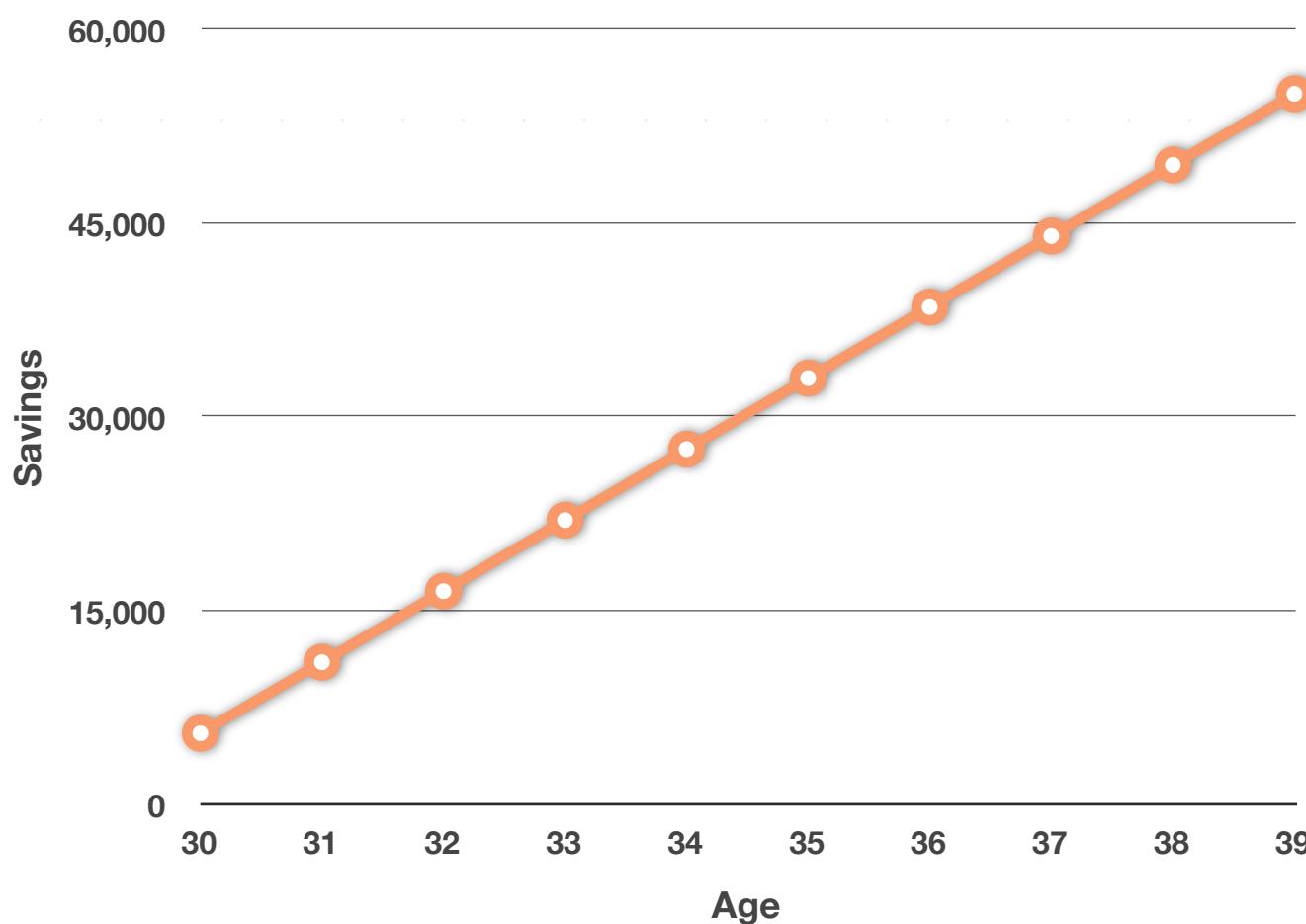




While this is valuable information, we might be even more interested in the accumulated amount over time. If we find the accumulated total after each year, the data is

Age	Investment	Total
30	\$5,500	\$5,500
31	\$5,500	\$11,000
32	\$5,500	\$16,500
33	\$5,500	\$22,000
34	\$5,500	\$27,500
35	\$5,500	\$33,000
36	\$5,500	\$38,500
37	\$5,500	\$44,000
38	\$5,500	\$49,500
39	\$5,500	\$55,000

Graphing the accumulated total each year could be much more valuable, because it gives us visibility into how our savings will grow over time.



From this ogive, we can quickly see that we'll have something between \$50,000 and \$60,000 once we've made our investment at age 39. Or we can see that we'll cross the \$30,000 mark once we make our investment at age 35.



# Two-way tables

Remember that one-way tables were made using **variable** data given for **individuals**. For example, people and their heights, or ice cream flavors and the number of scoops sold.

In this section we're transitioning to talk about two-way tables, which can be constructed from data that's dependent on two categorical variables.

Sometimes we talk about this kind of data in terms of **independent variables** and **dependent variables**.

In the case of data in one-way tables, we had one independent variable, called the individuals, and one or more dependent variables, called the variables. In the case of data in two-way tables, we have two independent categories on which the variables are dependent.

The difference between both types is easiest to visualize by comparing a data table for each one.

## Two-way tables

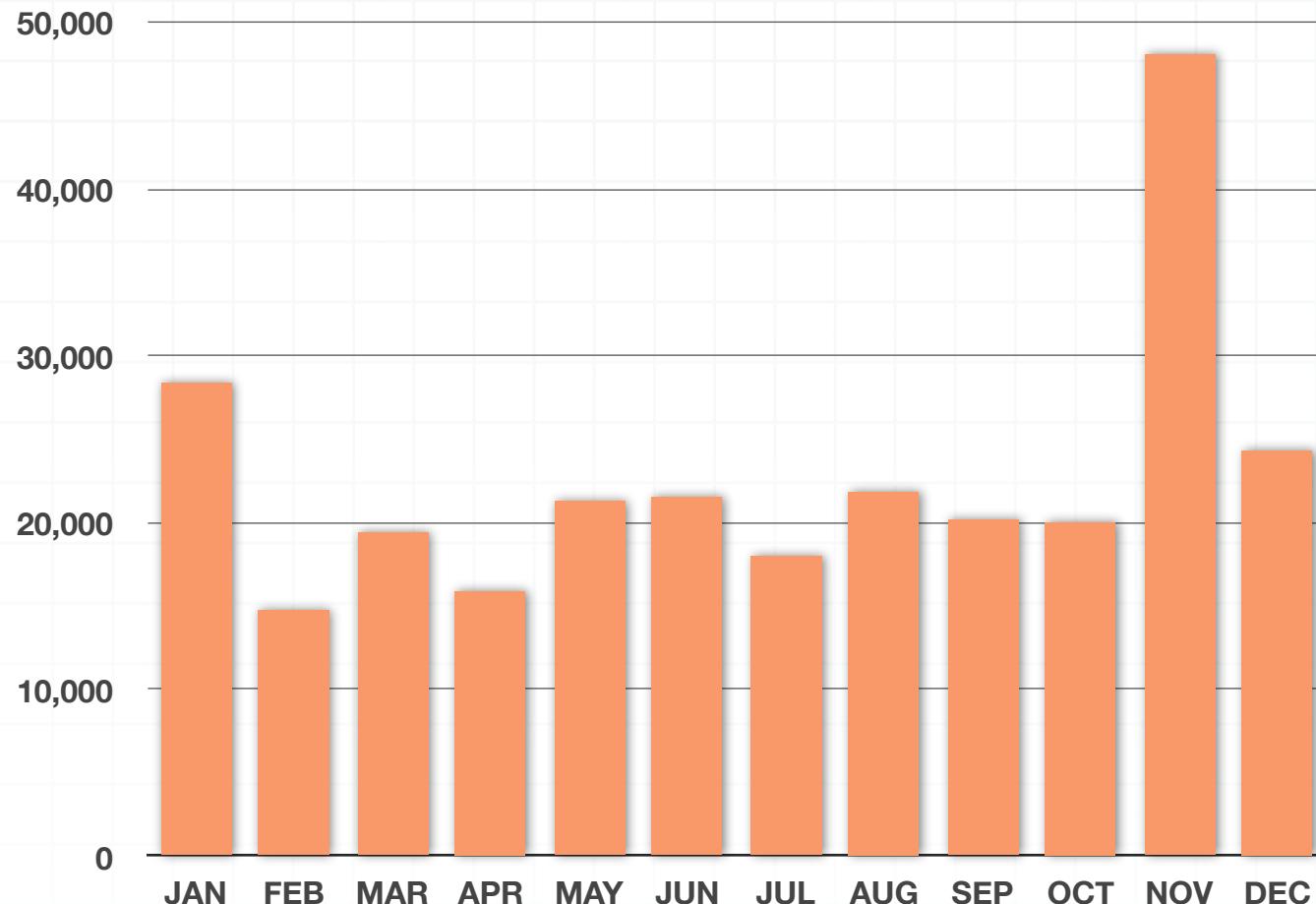
For example, let's say that an artist started a business selling his paintings at the beginning of 2014. By the end of 2017, he's been in business for four years, and his business has grown every year. If we add up the revenue earned all four years in January, all four years in February, all four years in March, etc., we could summarize his total earnings by month in a **one-way table**:



Month	Revenue
JAN	\$28,361
FEB	\$14,744
MAR	\$19,407
APR	\$15,891
MAY	\$21,277
JUN	\$21,530
JUL	\$17,990
AUG	\$21,838
SEP	\$20,174
OCT	\$20,025
NOV	\$48,055
DEC	\$24,318

In this one-way table, we have the months down the left side. Those are the individuals (also called the independent variable). The revenue is the variable (also called the dependent variable). In this chart, the revenue the artist earned from selling his paintings only depends on one thing: the month. We can only ask questions like “How much has the artist earned in February?” or “How much has the artist earned in September?”

Let's represent this as a bar graph.



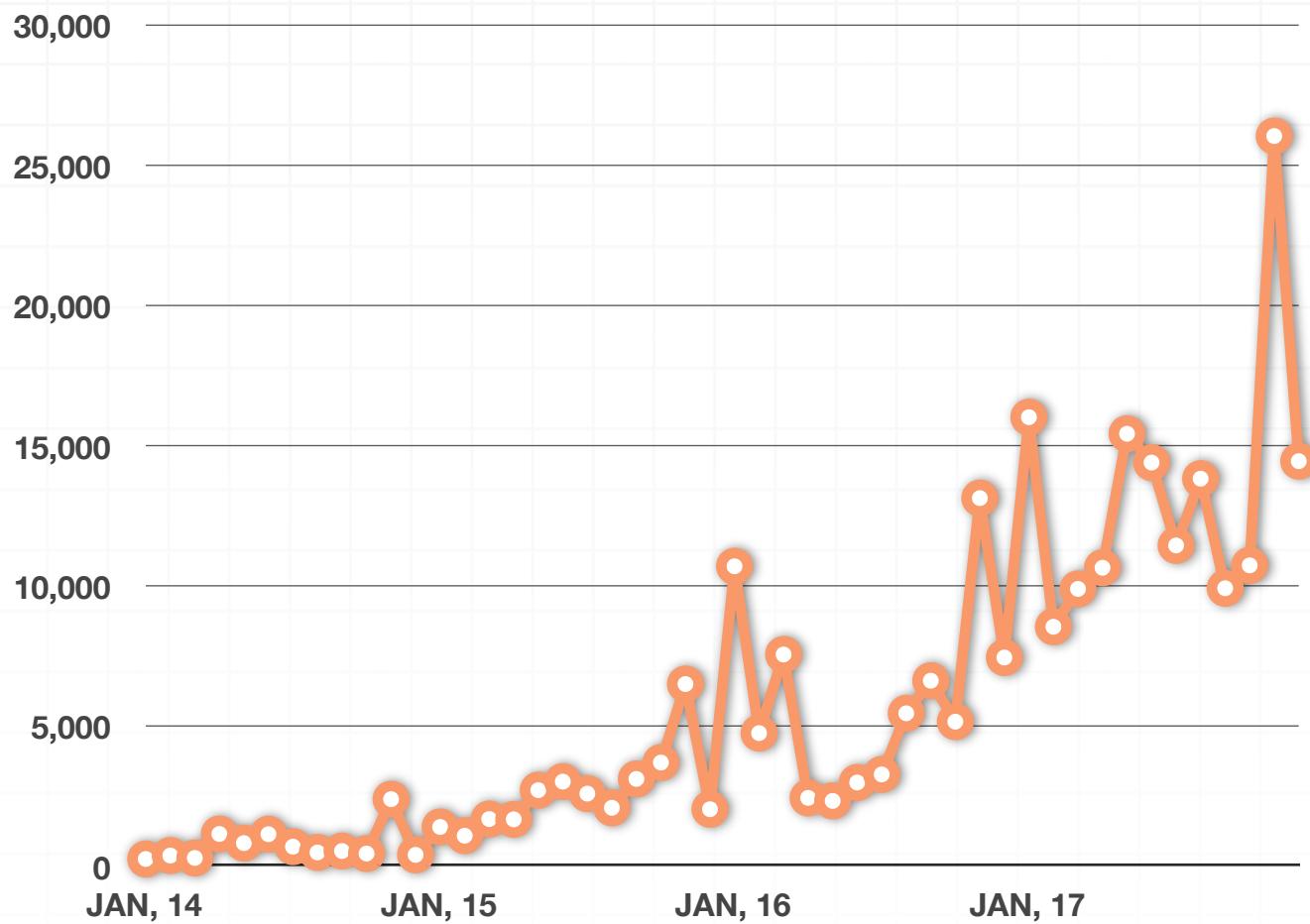
This certainly gives us information about which months tend to be best for the artist. November and January seem to be best. But since the data is summarized, adding the revenue for January for all four years into one figure, and for February, for March, for April, etc., this one-way data doesn't give us any insight into how the artist's business has grown over time.

So to get more visibility, we could instead break apart the revenue information and show each year separately in a **two-way table**:

Month	2014	2015	2016	2017
JAN	\$254	\$1,396	\$10,696	\$16,015
FEB	\$377	\$1,084	\$4,745	\$8,538
MAR	\$291	\$1,679	\$7,549	\$9,888
APR	\$1,146	\$1,668	\$2,434	\$10,643
MAY	\$820	\$2,708	\$2,326	\$15,423
JUN	\$1,138	\$3,014	\$2,982	\$14,396
JUL	\$694	\$2,586	\$3,270	\$11,440
AUG	\$486	\$2,080	\$5,451	\$13,821
SEP	\$538	\$3,109	\$6,614	\$9,913
OCT	\$448	\$3,695	\$5,153	\$10,729
NOV	\$2,387	\$6,495	\$13,128	\$26,045
DEC	\$401	\$2,030	\$7,441	\$14,446

Notice how the data in the two-way table is now dependent on two independent things, not just one. In the one-way table, if someone asked us how much revenue the artist earned, we'd reply "For which month?" But in the two-way table, if someone asked us how much revenue the artist earned, we'd reply "For which year and for which month?" In the two-way table, each value depends on two things: the month and the year. Whereas in the one-way table, each value depended on only one thing: the month.

We could graph the data in the two-way table in lots of different ways, but let's do it as a line graph that starts in January, 2014 and ends in December, 2017.



By giving the artist's revenue data in a two-way table, we get so much more insight into his business.

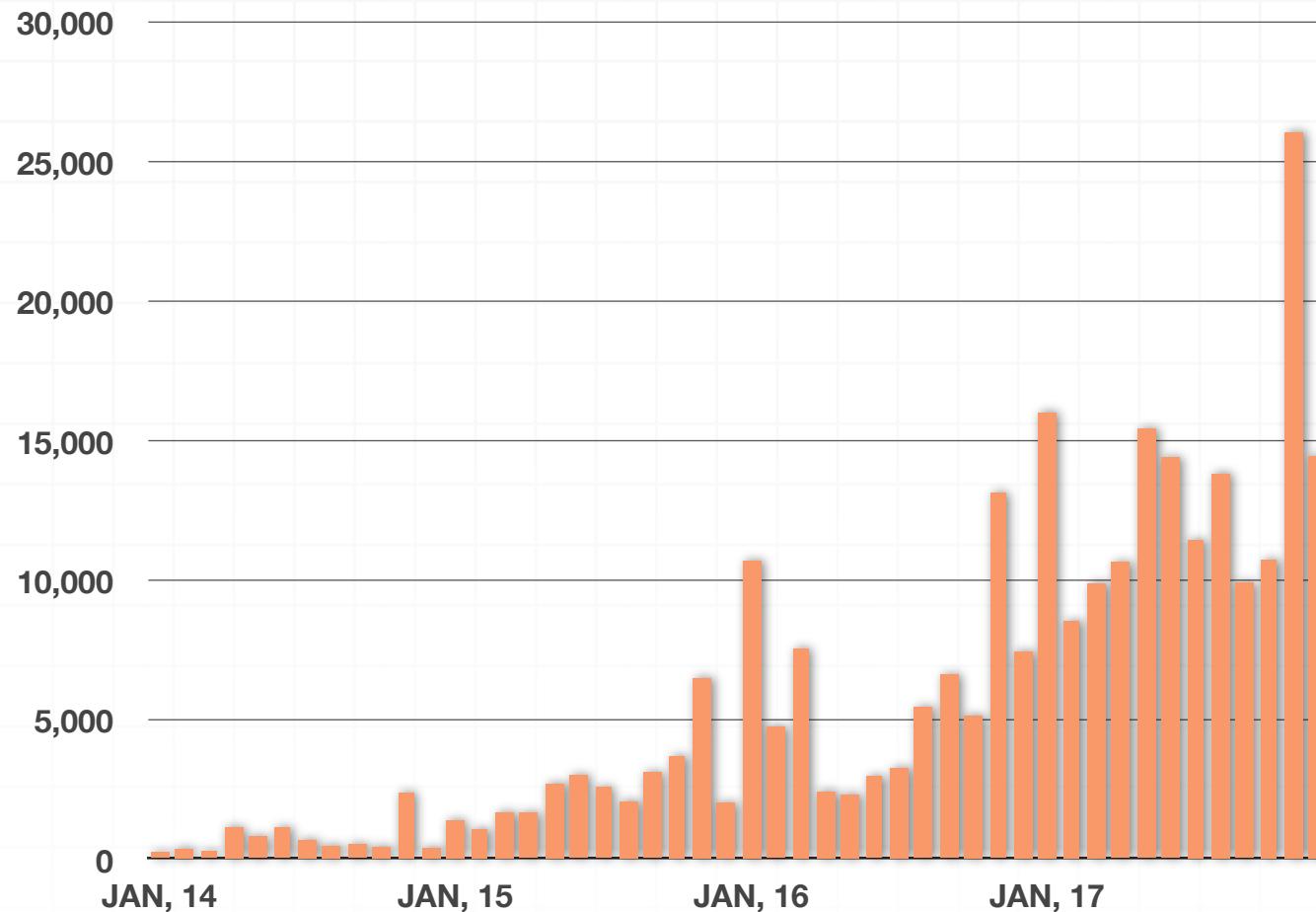
## Showing totals

In two-way tables, we can also summarize the data by row and by column, plus give a total for the entire table in the lower right-hand corner.

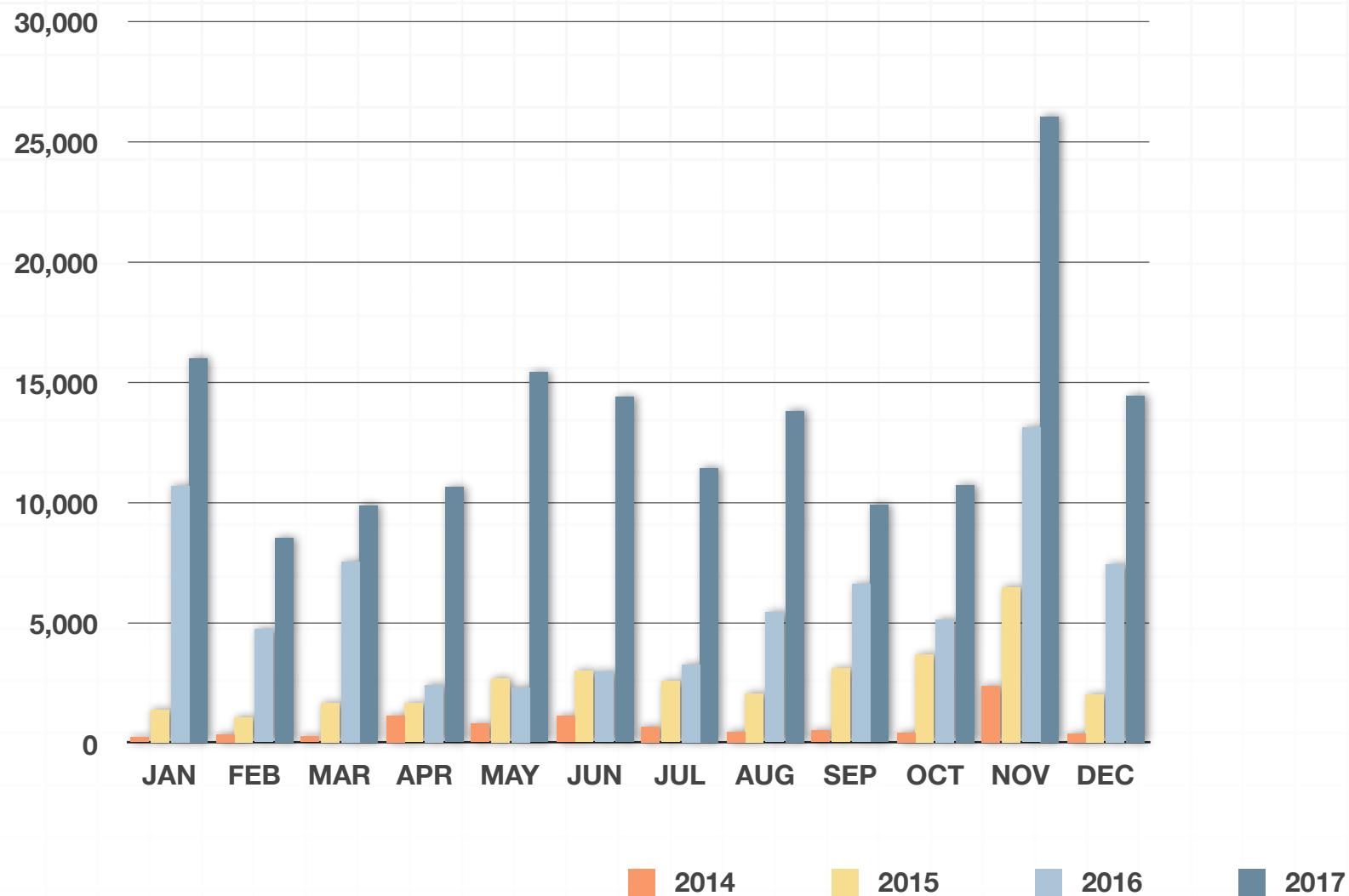
Month	2014	2015	2016	2017	Total
JAN	\$254	\$1,396	\$10,696	\$16,015	<b>\$28,361</b>
FEB	\$377	\$1,084	\$4,745	\$8,538	<b>\$14,744</b>
MAR	\$291	\$1,679	\$7,549	\$9,888	<b>\$19,407</b>
APR	\$1,146	\$1,668	\$2,434	\$10,643	<b>\$15,891</b>
MAY	\$820	\$2,708	\$2,326	\$15,423	<b>\$21,277</b>
JUN	\$1,138	\$3,014	\$2,982	\$14,396	<b>\$21,530</b>
JUL	\$694	\$2,586	\$3,270	\$11,440	<b>\$17,990</b>
AUG	\$486	\$2,080	\$5,451	\$13,821	<b>\$21,838</b>
SEP	\$538	\$3,109	\$6,614	\$9,913	<b>\$20,174</b>
OCT	\$448	\$3,695	\$5,153	\$10,729	<b>\$20,025</b>
NOV	\$2,387	\$6,495	\$13,128	\$26,045	<b>\$48,055</b>
DEC	\$401	\$2,030	\$7,441	\$14,446	<b>\$24,318</b>
<b>Total</b>	<b>\$8,980</b>	<b>\$31,544</b>	<b>\$71,789</b>	<b>\$161,297</b>	<b>\$273,610</b>

From the two-way table, we can now see the revenue totals for each month down the right side of the table (these match the totals in the one-way table), but we can also see the revenue totals for each year along the bottom. We can still see which month is best for the artist (November), but we can also see which year was best for the artist (2017).

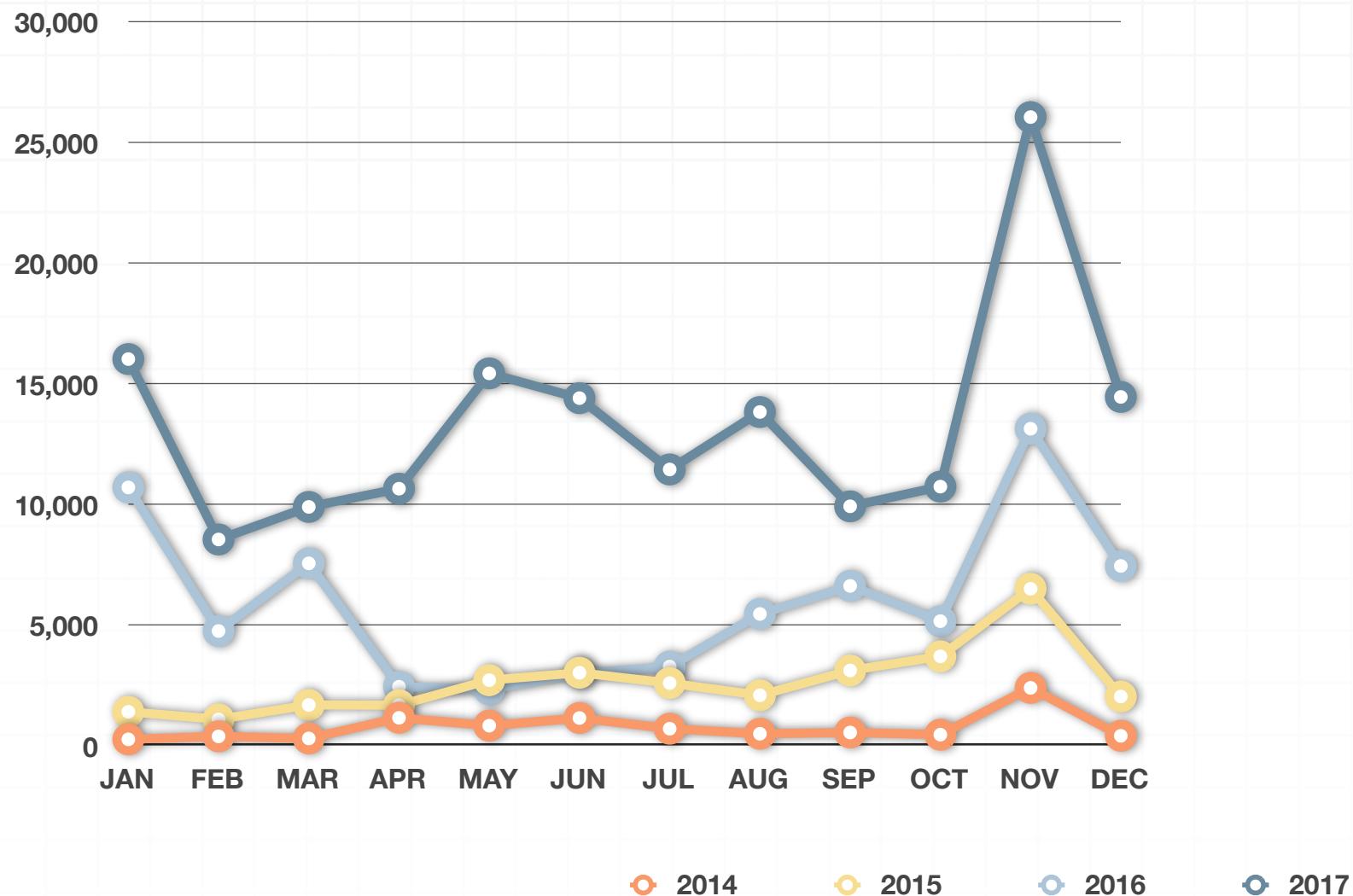
We did a line graph to represent all of this data over time, but we could have done a bar graph, instead.



We could also create a **comparison bar graph**, illustrating the revenue totals with side-by-side color coded bars for each year,



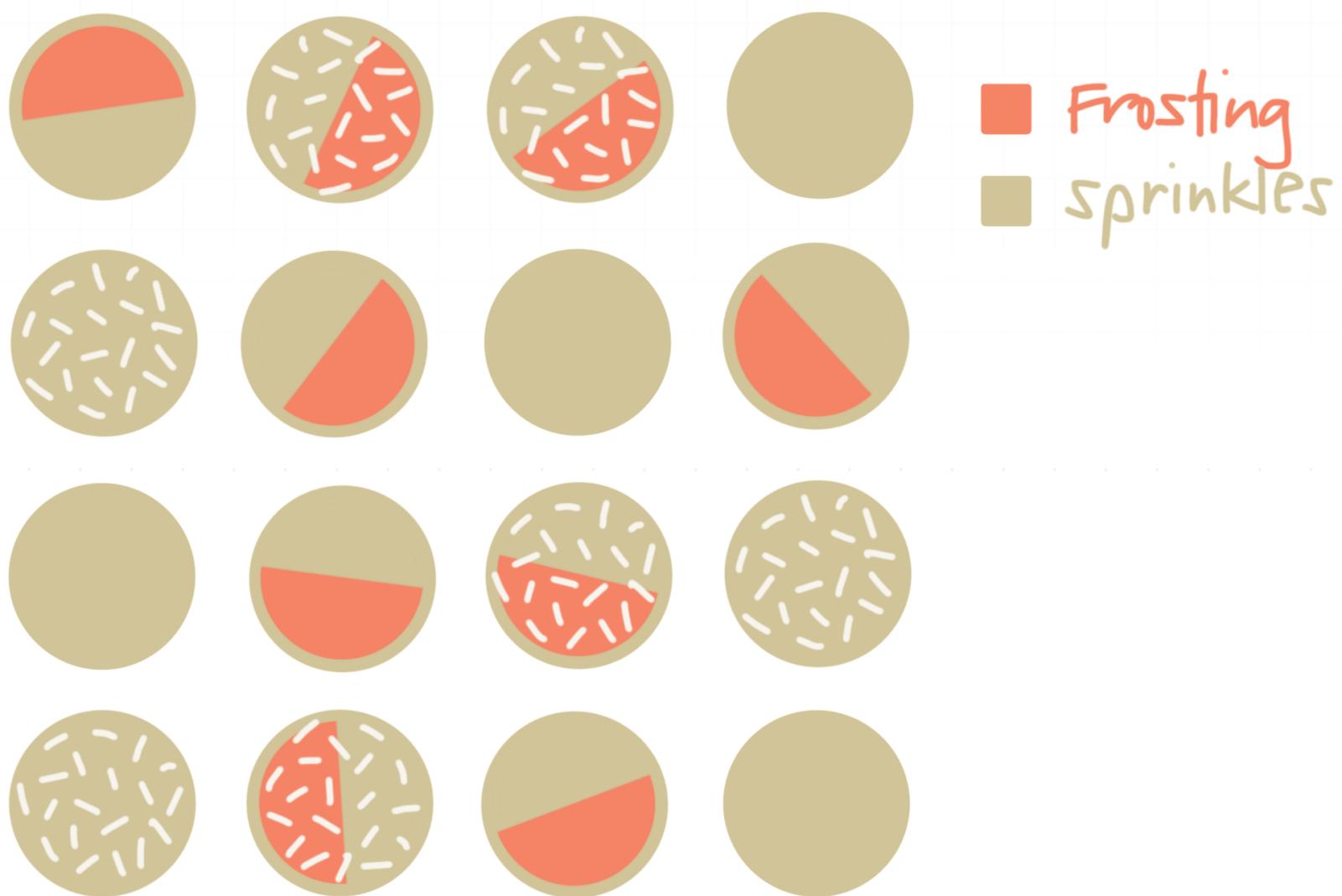
or a **comparison line graph**, depicting color-coded line graphs for each year.



# Venn diagrams

We've seen how we can express data in a **two-way table**, and then translate that data into bar and line graphs. But we can also express data from a two-way table in a different visualization called a **Venn diagram**.

Let's say we're baking sugar cookies for our family for Christmas. Like any good baker would, we want to categorize our cookies by making a chart that describes how many of each kind of cookie we have. Here's a picture of all of our sugar cookies:



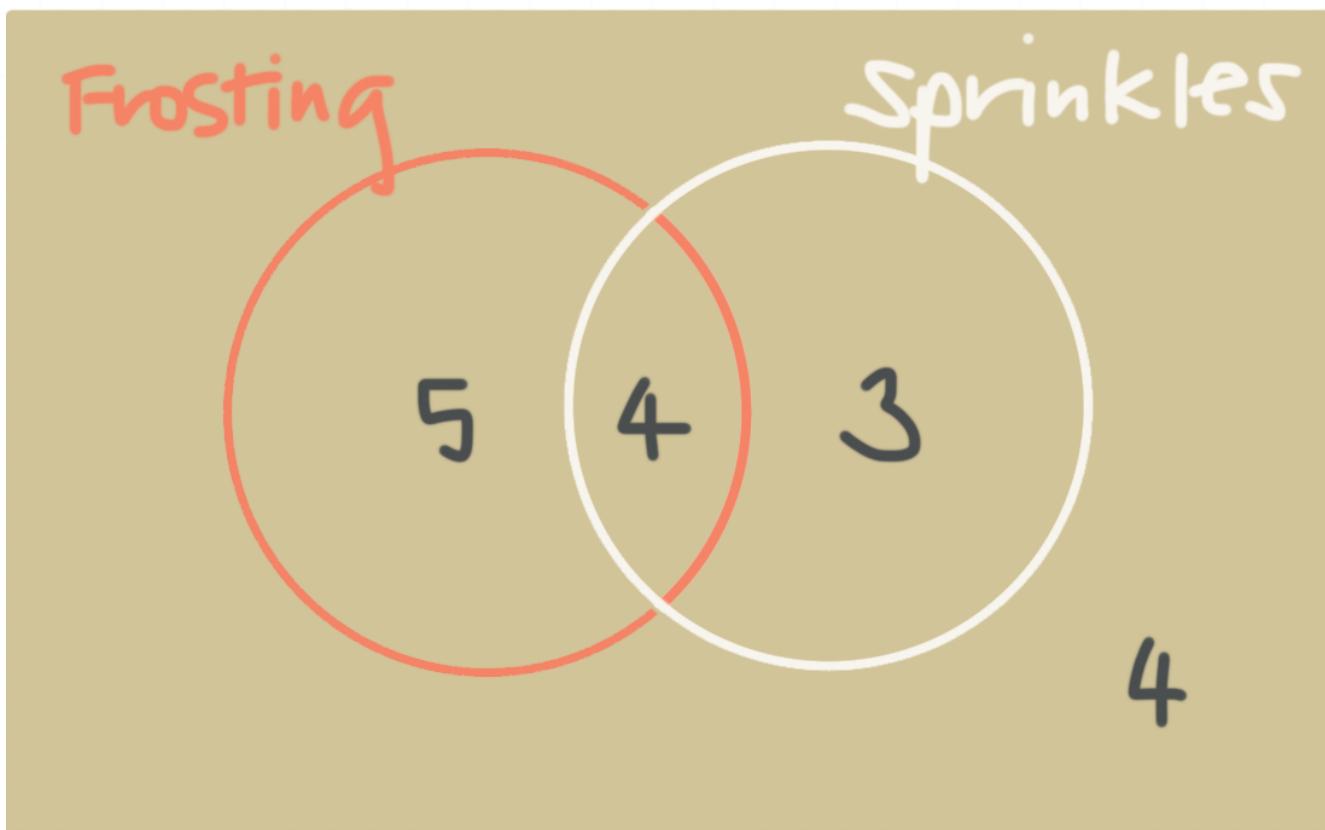
The ones with the red section are dipped in red frosting. The ones with the white dots are covered in white sprinkles. As we can see, some of our

cookies have both frosting and sprinkles, and some don't have either frosting or sprinkles.

Let's apply what we already know and represent this data in a **two-way table**. We'll go ahead and include totals as well.

	Sprinkles	No sprinkles	Total
Frosting	4	5	9
No frosting	3	4	7
Total	7	9	16

A Venn diagram representing this data might look like this:



Let's walk through this. The red circle represents the cookies with frosting and the gray circle represents cookies with sprinkles.

- Because we put a 5 inside the red circle but outside the gray circle, it means there are five cookies that have frosting but no sprinkles.
- Because we put a 3 inside the gray circle but outside the red circle, it means there are three cookies that have sprinkles but no frosting.
- Because we put a 4 inside both circles in the middle, it means there are four cookies that have both frosting and sprinkles.
- Because we put a 4 outside both circles in the corner, it means there are four cookies that don't have either frosting or sprinkles.

From the Venn diagram, we can also see that there are

- $5 + 4 = 9$  total cookies with frosting,
- $4 + 3 = 7$  total cookies with sprinkles,
- $5 + 4 + 3 = 12$  total cookies that have at least frosting or sprinkles or both, and
- $5 + 4 + 3 + 4 = 16$  cookies total

Notice that all of these numbers match the data in the two-way table.

So when we construct a Venn diagram, we put the number of items that only meet one condition in one circle, we put the number of items that only meet the other condition in the other circle, we put the number of items that meet both conditions in the middle where the circles overlap,

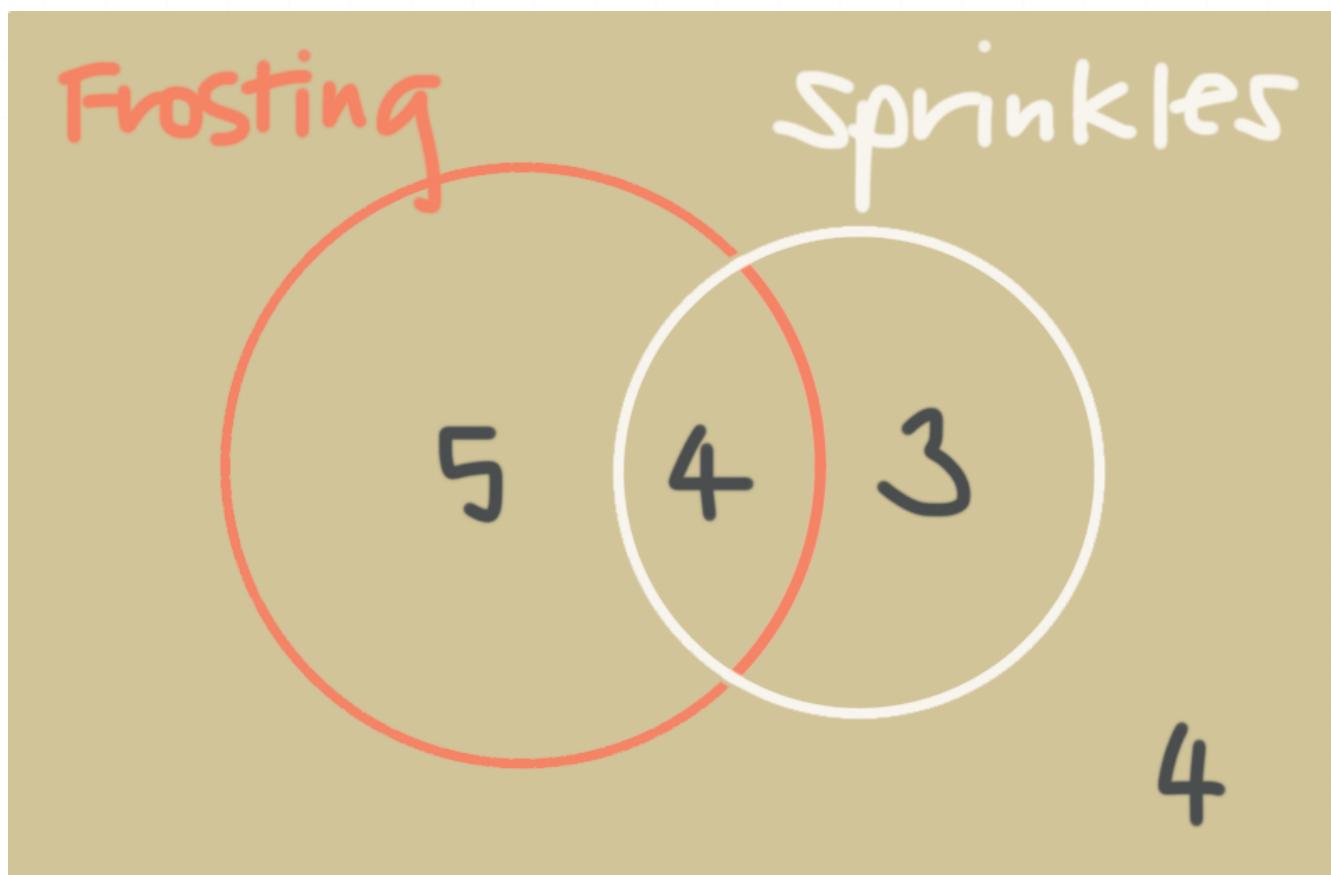


and we put the number of items that meet neither condition outside of all the circles but inside the box.

## Variations of Venn diagrams

We can also construct Venn diagrams with circles of different sizes, where the size of each circle is proportional to the number of items it represents.

So if we take the Venn diagram we made earlier for the cookies, we could adjust the circles so that the ratio of the frosting circle to the sprinkles circle is 9 : 7, since there are 9 cookies with frosting, and 7 circles with sprinkles.

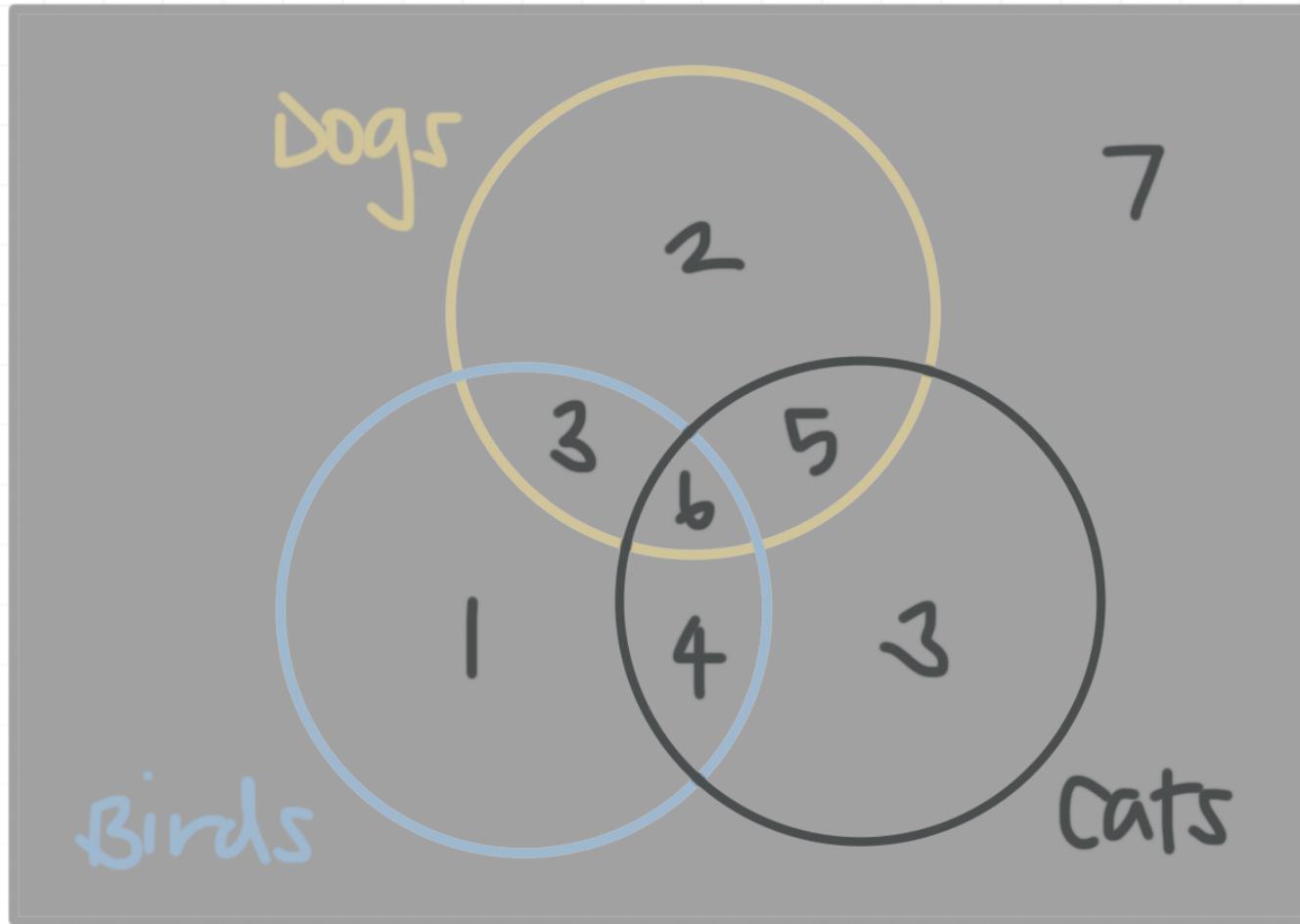


We can also create Venn diagrams with more than just two circles. For instance, suppose we asked students about their favorite kind of pet, and collected the following data.

- 2 students like dogs, but not cats or birds
- 3 students like cats, but not dogs or birds
- 1 student likes birds, but not dogs or cats
- 3 students like dogs and birds, but not cats
- 5 students like dogs and cats, but not birds
- 4 students like cats and birds, but not dogs
- 6 students like dogs, cats, and birds
- 7 students don't like any kind of pet

Even though there are three kinds of pets in this survey, dogs, cats, and birds, we can still construct a Venn diagram that models student preferences.

Let's use a brown circle for students who like dogs, a blue circle for students who like birds, and black circle for students who like cats.



From the diagram, we can see that,

- Because we put a 2 inside the brown circle but outside the black and blue circles, it means 2 students like dogs only.
- Because we put a 3 inside the black circle but outside the brown and blue circles, it means 3 students like cats only.
- Because we put a 1 inside the blue circle but outside the brown and black circles, it means 1 student likes birds only.
- Because we put a 3 inside the overlap of the blue and brown circles but outside the black circle, it means 3 students like dogs and birds, but not cats.

- Because we put a 5 inside the overlap of the brown and black circles but outside the blue circle, it means 5 students like dogs and cats, but not birds.
- Because we put a 4 inside the overlap of the blue and black circles but outside the brown circle, it means 4 students like birds and cats, but not dogs.
- Because we put a 6 inside the overlap of the blue, brown, and black circles, it means 6 students like birds, dogs, and cats.
- Because we put a 7 outside all three circle overlaps in the corner of the Venn diagram, it means 7 students don't like any kind of pet.



# Frequency tables and dot plots

Think of a **frequency table** as a table that displays how frequently or infrequently something occurs. A **dot plot** display can also be used to show the frequency of small data sets.

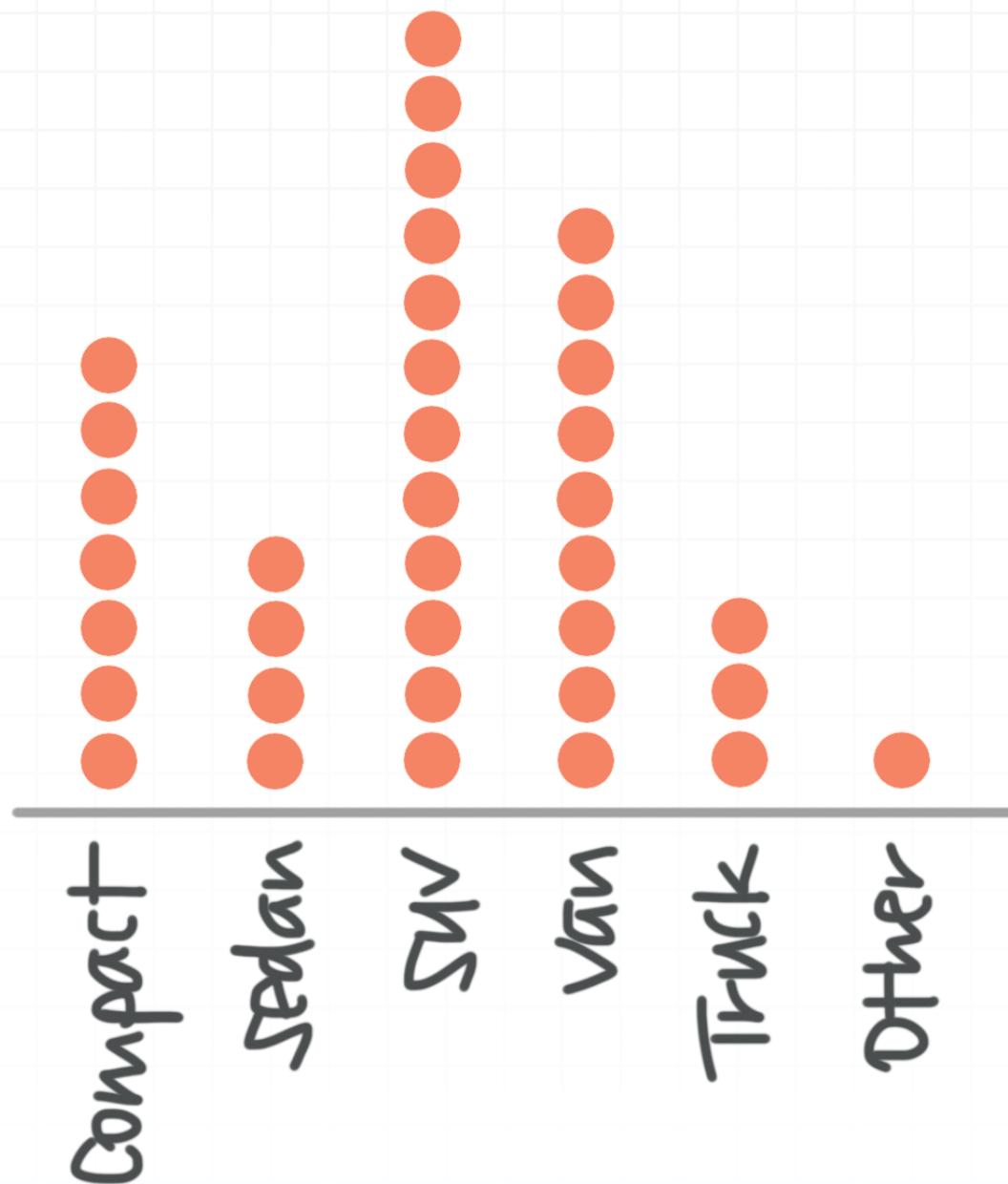
If I want to categorize the vehicles in a parking lot by type, I could do so in frequency table.

Type	Count
Compact	7
Sedan	4
SUV	12
Van	9
Truck	3
Other	1

To make this table, I counted the number of cars of each type, and recorded them in the table. Since I counted 7 compact cars, I wrote a 7 next to “Compact” in the table, and since I counted 9 vans, I wrote a 9 next to “Van.”

I could have also recorded this data in a dot plot.





To make the dot plot, we just took the data in the frequency table and changed the counts into dots. Since we counted 7 compact cars and put that in the frequency table, we put 7 dots above “Compact” in the dot plot.

Dot plots are a lot like bar graphs, in the sense that we can very easily see which items occur most frequently, based just on the height of the dots in the dot plot, like the height of the bar in a bar graph.

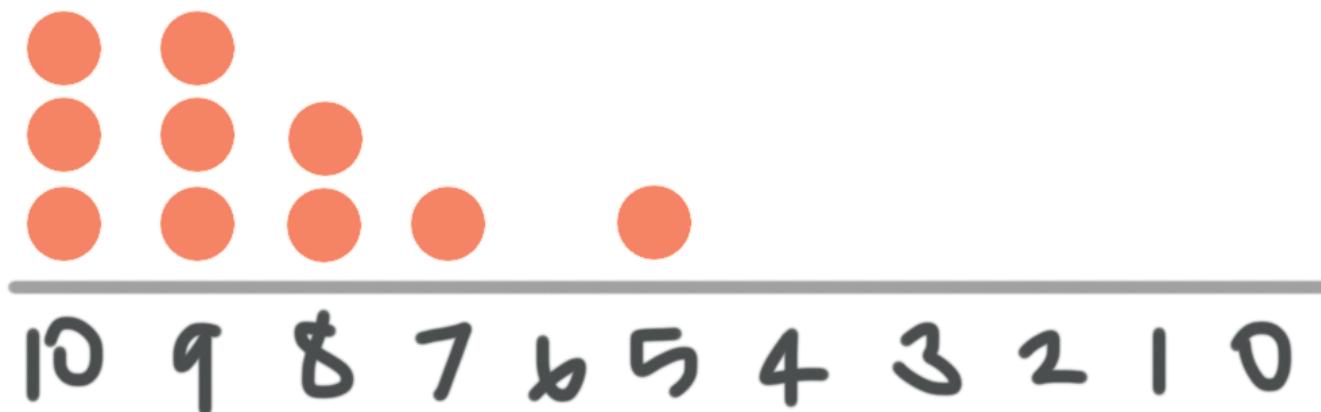
We can also convert between a list of data, a frequency table, and a dot plot. For example, 10 students took a quiz where they could earn scores between 0 and 10. Their scores were

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

A frequency table can be created to show the frequency of any given score between 0 and 10.

Quiz score	Count
10	3
9	3
8	2
7	1
6	0
5	1
4	0
3	0
2	0
1	0
0	0

Then the data from the frequency table can be turned into a dot plot.



# Relative frequency tables

We're familiar now with displaying data in a two-way table. For example, we might categorize the surfboards at Ricky's surf shop by length and by color.

	Blue	White
Longboard	20	15
Shortboard	5	3

But sometimes it's helpful to express the data in a two-way table as percentages. For example, of all the blue surfboards in Ricky's surf shop, what percentage of them are longboards, and what percentage of them are shortboards?

If we want to express these percentages, then we just need to change the two-way table into what's called a **relative frequency table**, which is a table that shows percentages instead of actual counts.

So to answer our question from earlier, we know from our table that there are  $20 + 5 = 25$  total blue surfboards, which means that  $20/25 = 0.8 = 80\%$  of the blue surfboards are longboards. And  $5/25 = 0.2 = 20\%$  of the blue surfboards are shortboards. We could do the same calculations for the white surfboards:

$$\text{Longboards: } 15/(15 + 3) = 15/18 = 0.\overline{8} \approx 83\%$$

$$\text{Shortboards: } 3/(15 + 3) = 3/18 = 0.\overline{16} \approx 17\%$$



So we could change the table from one that just shows a count of surfboards, into a relative frequency table that shows the percentage of longboards and shortboards based on color.

	Blue	White
Longboard	0.80	0.83
Shortboard	0.20	0.17
Column total	1.00	1.00

Now we can see that out of all the blue surfboards, 80% of them are longboards and 20% of them are shortboards. And out of all the white surfboards, about 83% are longboards and 17% are shortboards.

We could also create a relative frequency table from one-way table. Using the table from a previous section about the number of times each continent has played host to the summer Olympic games,

Continent	Count
Europe	16
North America	6
Australia	2
Asia	3
South America	1
Column total	28

we could turn this table into a relative frequency table.

Continent	Count
Europe	0.57
North America	0.21
Australia	0.07
Asia	0.11
South America	0.04
<b>Column total</b>	<b>1.00</b>

## Rows vs. columns

Notice that because we found the percent of the total within each column, that we were able to add a “Column total” to the bottom of the surfboard table. Each column sums to 1.00.

$$0.80 + 0.20 = 1.00$$

$$0.83 + 0.17 = 1.00$$

That's because in each cell, we found the percentage of the total for the column, and so each column sums up to 100 %. Therefore, the table we made is actually called a **column-relative frequency**, where the columns sum to 1.00 but the rows do not.

But we could have just as easily created a **row-relative frequency** by converting the rows into percentages instead of the columns into percentages.

We'll start again with the original table.



	Blue	White
Longboard	20	15
Shortboard	5	3

This time, we want to find the percentage of each color based on the kind of surfboard. In other words, what percentage of longboards are blue or white, and what percentage of shortboards are blue or white? Let's do the calculations for longboards:

$$\text{Blue: } 20/(20 + 15) = 20/35 \approx 0.571 \approx 57.1\%$$

$$\text{White: } 15/(20 + 15) = 15/35 \approx 0.429 \approx 42.9\%$$

Now we'll do the calculations for shortboards:

$$\text{Blue: } 5/(5 + 3) = 5/8 = 0.625 = 62.5\%$$

$$\text{White: } 3/(5 + 3) = 3/8 = 0.375 = 37.5\%$$

Now we can turn the table into a row-relative frequency table.

	Blue	White	Row total
Longboard	0.571	0.429	1.00
Shortboard	0.625	0.375	1.00

Out of all the longboard surfboards, 57.1% of them are blue and 42.9% of them are white. And out of all the shortboard surfboards, about 62.5% are blue and about 37.5% are white.

## Total relative frequency

We can also calculate the total-relative frequency for a two-way table by dividing everything by the grand total.

	Blue	White	Total
Longboard	20	15	35
Shortboard	5	3	8
Total	25	18	43

Since there are 43 total surfboards, we divide each value in the body of the table and each value in the total column and row by 43. Then we can express the values as decimals

	Blue	White	Total
Longboard	0.46	0.35	0.81
Shortboard	0.12	0.07	0.19
Total	0.58	0.42	1.00

or as percentages.

	Blue	White	Total
Longboard	46%	35%	81%
Shortboard	12%	7%	19%
Total	58%	42%	100%

# Joint distributions

A **joint distribution** is a table of percentages similar to a relative frequency table. The difference is that, in a joint distribution, we show the distribution of one set of data against the distribution of another set of data.

Let's say we study a group of 100 individuals, measuring the average number of hours each participant spent exercising each week over the course of the study, and we also gathered data about the total number of pounds of weight lost in total by each participant over that same period of time.

		Weight lost			
		0-2	2-4	4-6	6+
Hours exercising	0-3	4%	2%	2%	1%
	3-6	8%	6%	5%	0%
	6-9	1%	17%	10%	4%
	9-12	3%	3%	12%	9%
	12+	2%	3%	4%	4%

From this table, we can see that 4 % of the group, which would be 4 out of the 100 people studied, spent between 0 and 3 hours per week exercising, and lost between 0 and 2 pounds. More people in the study (17 % or 17 out of the 100) exercised between 6 and 9 hours per week on average and lost between 2 and 4 pounds in total.

This is an example of a joint distribution because we're comparing the distribution of two distributions: the distribution of average weekly hours



spent exercising, and the distribution of total weight lost over the course of the study.

When we read a joint distribution table, we'll oftentimes look at marginal and conditional distributions within the table.

## Marginal distribution

We could total up the data in each row and each column, and add those totals to the table:

		Weight lost				
		0-2	2-4	4-6	6+	Total
Hours exercising	0-3	4%	2%	2%	1%	9%
	3-6	8%	6%	5%	0%	19%
	6-9	1%	17%	10%	4%	32%
	9-12	3%	3%	12%	9%	27%
	12+	2%	3%	4%	4%	13%
	Total	18%	31%	33%	18%	100%

Think of a **marginal distribution** as the Total column or the Total row in this joint distribution. It's like only having one of the distributions, not both. So if we only had the distribution of weight lost, we'd have just the totals along the bottom of the table.

	Weight lost				
	0-2	2-4	4-6	6+	Total
Total	18%	31%	33%	18%	100%

This table gives us the marginal distribution for weight lost; it tells us the percentage of participants who lost 0 – 2 pounds, 2 – 4 pounds, etc. But it doesn't give us any information about how much exercise they did.

We could also look at only the distribution for hours spent exercising, taking the totals along the right side of the original joint distribution.

		Total
Hours exercising	0-3	9%
	3-6	19%
	6-9	32%
	9-12	27%
	12+	13%
Total		100%

This is the marginal distribution for hours of exercise, and we can see the percentage of participants who spent 0 – 3 hours exercising, 3 – 6 hours exercising, etc. But this table doesn't give us any information about how much weight was lost based on how many hours were spent exercising.

## Conditional distribution

Think of a **conditional distribution** as the distribution of one variable, given a particular value of the other variable.

For example, we could look at the conditional distribution of weight lost, given a particular amount of hours spent exercising:

		Weight lost				
		0-2	2-4	4-6	6+	Total
Hours exercising	0-3	44%	22%	22%	12%	100%
	3-6	42%	32%	26%	0%	100%
	6-9	3%	53%	31%	13%	100%
	9-12	10%	10%	40%	40%	100%
	12+	15%	23%	31%	31%	100%

Looking at the top row of this conditional distribution, we can say this:

Given that people spent 0 – 3 hours exercising,

44 % of them lost 0 – 2 pounds,

22 % of them lost 2 – 4 pounds,

22 % of them lost 4 – 6 pounds,

12 % of them lost 6+ pounds.

But this distribution is conditional on 0 – 3 hours spent exercising. Notice that the distributions in the table are conditional upon the value of hours spent exercising in each row, so we see the row totals at 100 % .

Or we could look at the conditional distribution of hours exercising, given a particular amount of weight lost:

		Weight lost			
		0-2	2-4	4-6	6+
Hours exercising	0-3	22%	6%	6%	6%
	3-6	44%	19%	15%	0%
	6-9	6%	55%	30%	22%
	9-12	17%	10%	36%	50%
	12+	9%	10%	13%	22%
	Total	100%	100%	100%	100%

Looking at the first column of this conditional distribution, we can say this:

Given that people lost 0 – 2 pounds of weight,

22 % spent 0 – 3 hours exercising

44 % spent 3 – 6 hours exercising

6 % spent 6 – 9 hours exercising

17 % spent 9 – 12 hours exercising

9 % spent 12+ hours exercising

But this distribution is conditional on 0 – 2 pounds of weight lost. Notice that the distributions in the table are conditional upon the value of pounds of weight lost in each column, so we see the column totals at 100 % .

# Histograms and stem-and-leaf plots

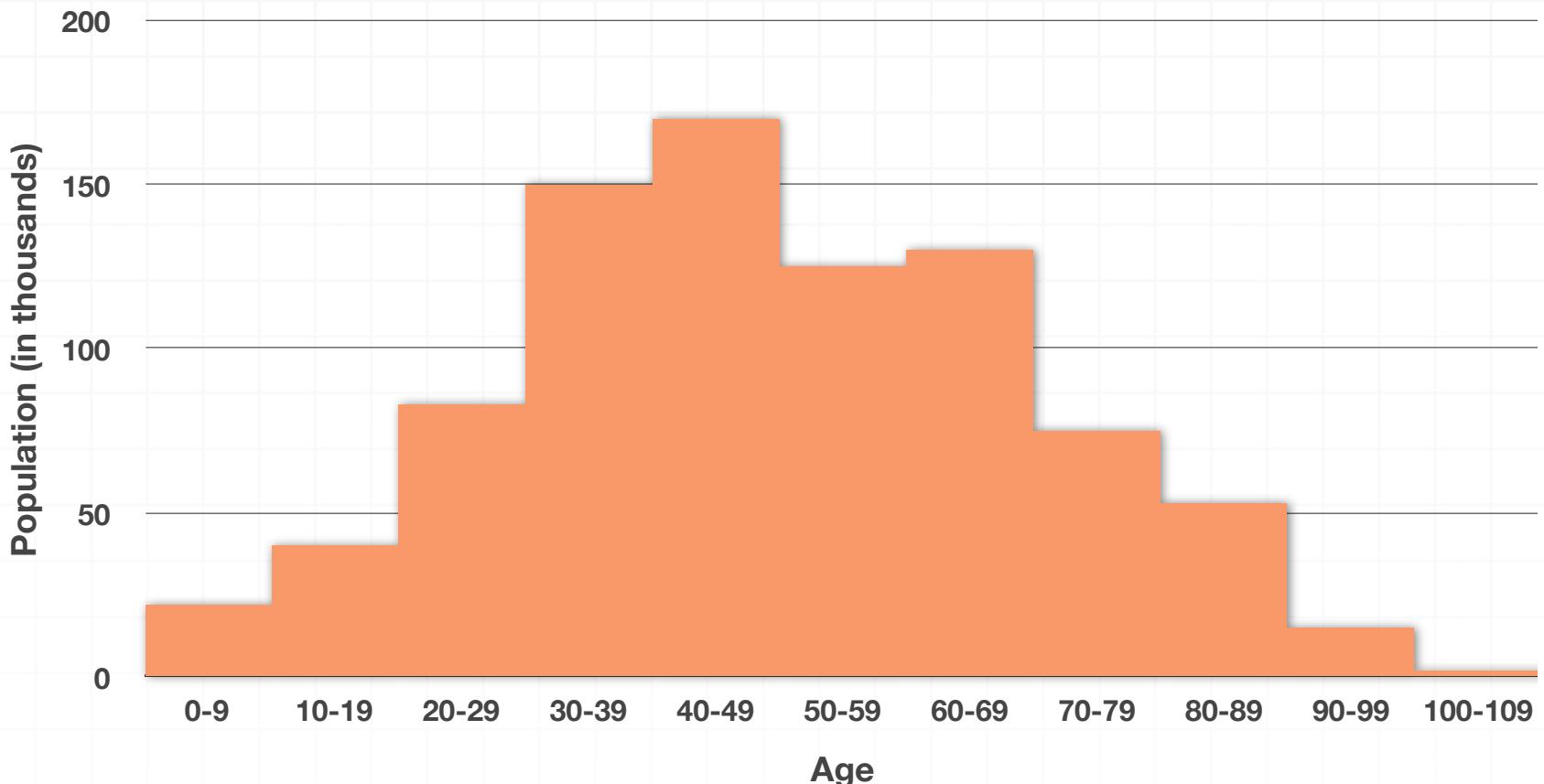
A **histogram**, also called a **frequency histogram**, is just like a bar graph, except that we collect the data into **buckets** or **bins**, and then sketch a bar for each bucket. Each bucket needs to be the same size, or width, so that they're capable of holding the same amount of data.

Unlike bar charts, histograms have no gaps between the bars (although some bars might be absent, which means there's no frequency in that "bucket"). A histogram represents a continuous data set, which is why there are no gaps between the buckets.

One reason we might want to use a histogram instead of a bar graph is because we have too many data points to plot individually. For example, maybe we want to use census data to make a graph of the number of people of each age in the entire city of San Francisco. In a typical bar graph, we have to show a bar for children younger than 1, another for 1-year-olds, for 2-year-olds, 3, 4, 5, all the way up to 100 or maybe even older. In other words, our bar graph might have 100 bars or more.

A histogram is the perfect solution to an overly-complicated bar graph. To create a histogram for the same information, we might group together 0 – 9 year-olds, 10 – 19 year-olds, 20 – 29 year-olds, etc. Notice that each of these buckets is the same size or length. That's important to remember when making a histogram. Putting people of similar age together in those groups would allow us to create a histogram with around 10 bars, instead of a bar graph with around 100 bars. The histogram might look like this:





The way the data is spread out in the histogram is called the **distribution**. As a very general rule, qualitative data is usually better in a bar graph, and quantitative data is usually better in a histogram.

## Stem-and-leaf plot

A **stem-and-leaf plot** (also called a **stem plot**), is just another way to summarize data. It's similar to a histogram, because both types of charts group together data points, and are good ways to show how many data points fall into a certain category or range.

For stem-and-leaf plots, we group data together by the first digit(s) in each number. In other words, let's say we have the finishing scores of golfers in a round of tournament golf:

66, 67, 67, 68, 68, 68, 68, 69, 69, 69, 69, 70, 70, 71, 71, 72, 73, 75

We could create a stem plot of the scores.

6	6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9
7	0, 0, 1, 1, 2, 3, 5

$$6|6 = 66$$

Let's use this particular plot to talk about what a stem plot shows. First, the “**stems**” are the numbers on the left, in this case the 6 and the 7. The “**leaves**” are all the other numbers on the right.

Each leaf represents one data point, in this case one golf score. So if we want to know how many data points are in the set, we could count the number of leaves on the right side. In this plot, there are 18 leaves, which means we collected 18 golf scores.

Notice that we also put “ $6|6 = 66$ ” below the stem-and-leaf plot. This is a key, or legend, that tells us that we intended for the stem to represent the tens place, and for the leaf to represent the units place. If the key had said  $6|6 = 606$ , that would have meant that each stem represented the hundreds place (6 would indicate 600), and the leaf would represent the units place (6 would indicate 6).

Each leaf needs to be attached to the stem from the same row in order to give us each data point. In other words, the stem isn't a data point on its own, and neither is the leaf. They only make a data point when we put them together. So if we take the first stem, 6 tens, and the first leaf, 6 ones, we put them together to get 66, and that's one golf score. If we put the second leaf, 7 ones, with the stem, we get 67, which is another golf



score in our data set. We could do this up to the last leaf on that line, to get golf scores from the first line of our plot of

66, 67, 67, 68, 68, 68, 68, 69, 69, 69, 69

We could do the same with the second line to see that we also have scores of

70, 70, 71, 71, 72, 73, 75

### No digit, or more than one digit in the stem?

In a stem-and-leaf plot, the leaf will always only have one digit, and the stem will take the rest of the digit. So for example, if another golfer had a terrible day and scored 103 and we wanted to add his score into our plot, we'd make the 3 the leaf, and the 10 would be the stem.

6	6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9
7	0, 0, 1, 1, 2, 3, 5
10	3

Even single-digit numbers can be included in a stem plot. If we wanted to add data points of 5, 7 and 9 to our plot, we can either leave the stem blank,



	5, 7, 9
6	6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9
7	0, 0, 1, 1, 2, 3, 5
10	3

or put a 0 in the stem spot.

0	5, 7, 9
6	6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9
7	0, 0, 1, 1, 2, 3, 5
10	3

# Building histograms from data sets

Now that we understand the basics of histograms, let's dive into the details about how we can actually build a histogram if we're only given the data set. Let's start by defining some important terms related to histograms.

The first one is **class interval**, or **class**, or **bin**, as we mentioned in the previous lesson. We always divide our data set into class intervals with equal **class width**. The class width is usually the difference between either the upper limits of two consecutive classes or between the lower limits of two consecutive classes.

For example, if the first class in our histogram is 5 – 9 and the second class is 10 – 14, then the class width is given by the difference between the upper limits,  $14 - 9 = 5$ , or by the difference between the lower limits,  $10 - 5 = 5$ .

A **class midpoint** is the value that's at the center of a particular class. The class midpoint is half the sum of the lower and upper limits of the class.

For example, for the class 100 – 104, the class midpoint is

$$\frac{100 + 104}{2} = 102$$

If at all possible, it's nice to choose a class width that's odd (like a width of 5, 13, 27, etc.) because it'll make the class midpoint an whole number, instead of a decimal number.



Let's look at the steps that we need to use to turn raw data into a histogram.

1. Put the data set in ascending order, then find the range as the difference between the largest and smallest values.
2. Determine the number of bins, or classes, that we want to have in our histogram. As a rule of thumb, it's best to use 5 – 6 classes for most of the data we'll work with during our statistical studies. However, we might want to use up to 20 classes when we deal with larger data sets. It all depends on how large our data set is and the number of classes that would best represent the data.
3. Divide the range by the number of classes, then round up the result to get the class width.
4. Build a table, putting each class in a separate row.
5. Find the frequency for each class by counting the data points that fall into each one. When adjacent bins are for intervals like 0 – 4 and 4 – 8, the data point 4 should be included in the second bin. In other words, the interval 0 – 4 actually contains data from 0 to 3.999..., and the interval 4 – 8 contains data from 4 to 7.999...
6. Graph the histogram by placing the classes along the horizontal axis and their frequencies along the vertical axis, such that the height of each bar is the frequency of each class.

Let's work through an example so that we can see these steps in detail.



## Example

A total of 20 students recorded the number of hours they spent doing homework last week. Construct a histogram for the number of homework hours spent by the group.

3, 2, 6, 13, 7, 5, 12, 1, 8, 4, 5, 9, 6, 15, 4, 3, 10, 14, 5, 11

First, put the values in ascending order. This makes it easier to see the maximum and minimum values of the data set.

1, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

The range is  $15 - 1 = 14$ . We'll divide the data set into 5 classes, which means the class width will be

$$\frac{14}{5} = 2.8 \approx 3$$

Remember that it's better to round the result up, which is why we rounded 2.8 to 3. Since 1 is the minimum value, we can use it as a starting point, or the lower limit of the first class. So our table for the classes and their frequencies will be

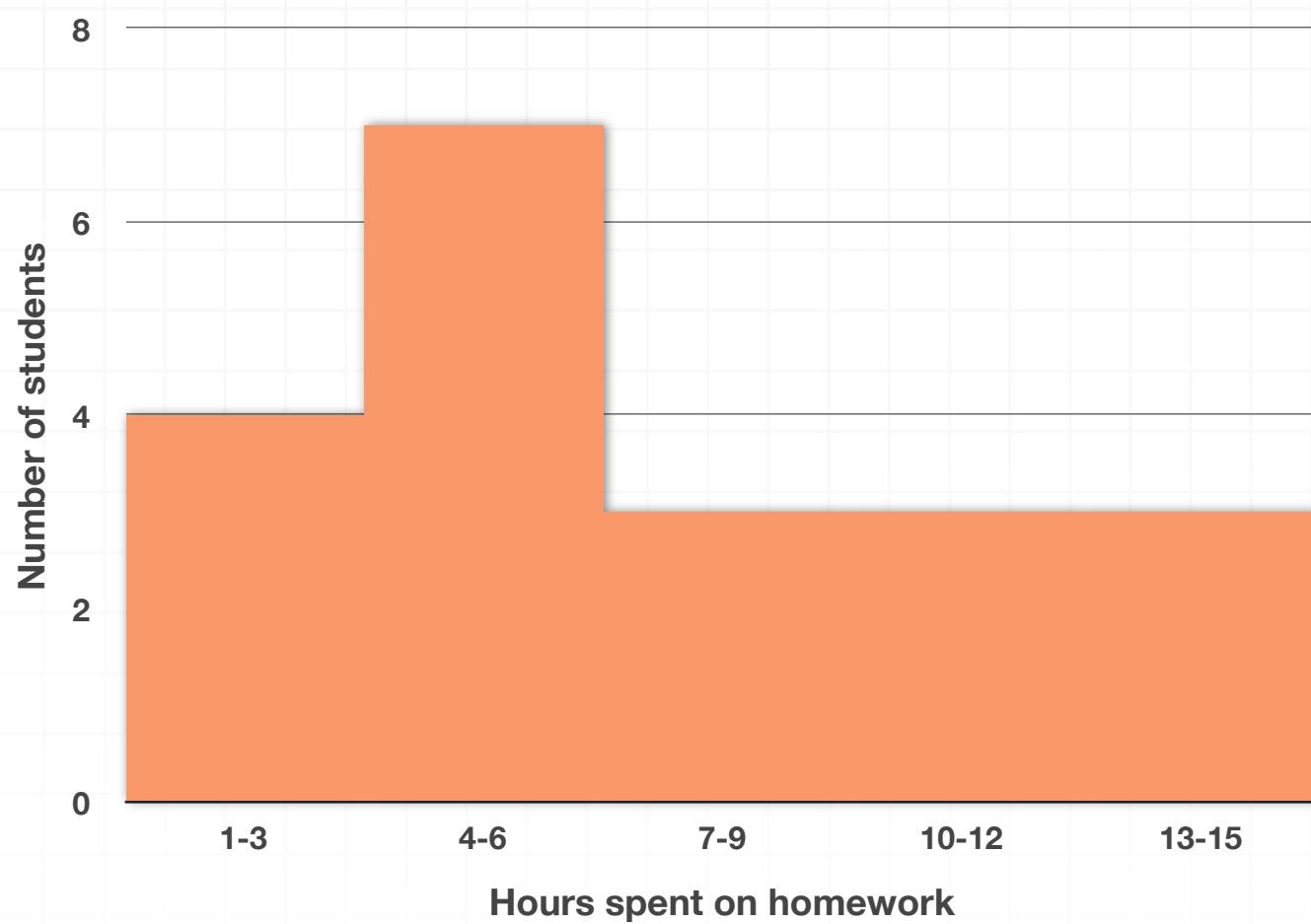


Classes	Frequency
1 - 4	4
4 - 7	7
7 - 10	3
10 - 13	3
13 - 16	3

We don't need it for this particular problem, but we could also determine the midpoint of each class.

Classes	Frequency	Class midpoint
1 - 4	4	2.5
4 - 7	7	5.5
7 - 10	3	8.5
10 - 13	3	11.5
13 - 16	3	14.5

Now we can use the table to draw the histogram.



# Measures of central tendency

We've talked a lot about data sets and the individual data points contained within them. And we've looked at ways to create visual representations of data.

Now we want to start analyzing the data set in a different way. In this section we're going to look at what we call **measures of central tendency**, which are different ways we've come up with to describe the "middle," "center," or most typical value of the data.

## Mean

We usually say "average," but technically we're thinking about **mean**, also called the arithmetic mean. We calculate the mean using a specific formula:

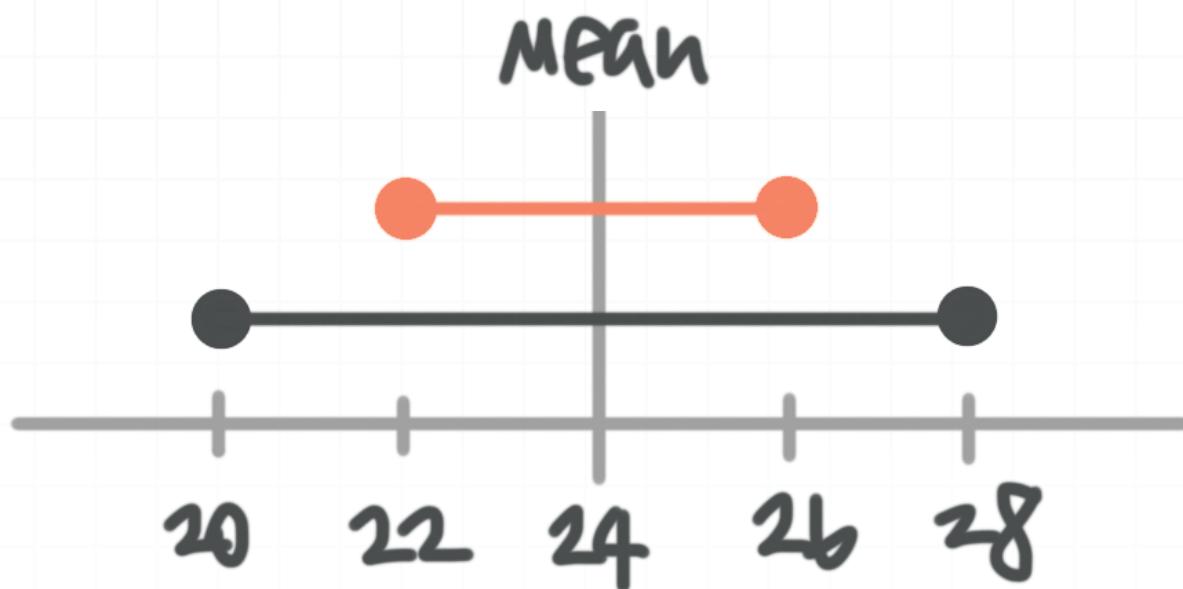
$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

In the formula,  $\mu$  (pronounced "mew") is the mean,  $n$  is the number of items in the data set, and the sum in the numerator tells us to start with the first data point, adding up all the data points together until we get to the last one. In other words, it would be just as correct to write the formula for mean as

$$\mu = \frac{\text{the sum of all the data points}}{\text{the number of data points}}$$



We can also think about the mean as the “balancing point” of the data. Let’s say that we have the data set 20, 22, 26, 28, where the data is evenly spread out. In this case, we can see what the mean is just by looking at the data set. We might predict that the mean is 24, and here’s why. We’ll illustrate the data on a number line.



We plotted 20 and 28 as gray dots, and 22 and 26 as orange dots. We don’t really need the color coding, because the point we’re trying to illustrate is that we have an equal amount of distance from the mean on the left side as we do on the right side. To be more specific, 22 and 26 are both 2 units away from the mean and 20 and 28 are both 4 units away from the mean.

What this tells us is that whenever we find the mean, what we’re really doing is creating a balance of distance between the points to the left and right of the mean. And in that way, the mean represents the balancing point of all the data. In other words, it’s the point that would balance all of the distances between the points in the data set. If the mean were moved a bit left or right then the balance would tip one way or the other.

Realize also that the formula for calculating the mean allows us to find more than just the mean. If we have the mean, but we're missing one data point in your set, we could figure out the missing data point.

### Example

Given the data set 20, 22,  $x$ , 28, and knowing that the mean is 24, find the missing value from the data set.

Let's plug everything we know, including the missing data point, into the formula for the mean.

$$\mu = \frac{\text{the sum of all the data points}}{\text{the number of data points}}$$

$$24 = \frac{20 + 22 + x + 28}{4}$$

Then we just use algebra to solve for the missing data point.

$$96 = 20 + 22 + x + 28$$

$$96 - 20 - 22 - 28 = x$$

$$26 = x$$

The missing data point is 26.

## Median

The **median** of a data set is the value at the middle of the data set when we line up all the data points in order from least to greatest. If the data set has an even number of points, there won't be one number in the middle. In this case, we take the mean of the pair of numbers in the middle. So the median is the value that divides the data set into halves, and the median does not necessarily need to be one of the actual data points in the set.

If we have an odd number of data points, the median will come from one number. For example, take the data set with 7 data points:

1, 2, 3, 4, 5, 6, 7

We need to cross out an equal number of data points on each end of the data set until we get to the number in the center. In this case, we can cross out three data points on each side.

~~1, 2, 3, 4, 5, 6, 7~~

The median is 4.

When there are an even number of data points in the set, the process for finding the median is slightly different. In that case, we cross out everything but the middle two terms.

1, 2, 3, 4, 5, 6, 7, 8

~~1, 2, 3, 4, 5, 6, 7, 8~~

Then to find the median of the data set, we find the mean of the two data points in the middle.



$$\mu = \text{median} = \frac{4+5}{2} = \frac{9}{2} = 4.5$$

The median is 4.5.

When we have a large number of data points in the set, we can use

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

where  $n$  is the number of values in the set, in order to find the location of the mean. For example, if the data set has 121 values, then the median is located at the 61st value.

$$\frac{n+1}{2} = \frac{121+1}{2} = 61$$

Alternatively, if the data set has 78 values, then the median is located between the 39th and 40th values.

$$\frac{n+1}{2} = \frac{78+1}{2} = 39.5$$

Just remember that the location of the median isn't the same as the value of the median.

## Mode

The **mode** of a data set is the value that occurs most often, more than any other value. We could think about it as the most typical value in the set. In the set 1, 2, 3, 4, 4, 6, 7, the mode is 4 because 4 occurs twice and every



other value only occurs once, which means that 4 occurs more often than any other value.

Sometimes we'll have a set like 1, 2, 3, 4, 4, 6, 6. In this set, 4 and 6 occur twice, and every other value occurs once. Which means 4 and 6 occur most often. Because there's not a "clear winner" between 4 and 6, sometimes people will say the data set has no mode, others will say that the set has two modes and the data set is called **bi-modal**.

# Measures of spread

We looked at measures of central tendency, which we saw were various ways of representing the “middle” of a data set. But central tendency isn’t the only thing we’re interested in when it comes to data.

We also want to know about **spread**, which is how, and by how much, our data set is spread out around its center. We also call measures of spread **measures of dispersion**, or **scatter**.

Range, interquartile range (IQR), variance, and standard deviation are all measures of spread.

## Range

The **range** of a data set is the difference between the largest value and smallest value. For example, in this stem-and-leaf plot of golf scores,

6	6, 7, 7, 8, 8, 8, 8, 9, 9, 9, 9
7	0, 0, 1, 1, 2, 3, 5

the lowest score is 66 and the highest score is 75. Therefore the range of this data set is

$$75 - 66 = 9$$

## Interquartile range (IQR)

You know how when you divide something into four parts, you cut it in half and then cut each half in half again? Each of those pieces is then a quarter of the original whole.

In a similar way, we can divide a data set into quarters by using the medians in the data. We cut the data in half at the median, and then find the median of each half, splitting the data at those points. Each quarter of the data that we've created is bounded by the data's **quartiles**. The median of the lower half of the data is called the first quartile  $Q_1$ , the median of the upper half of the data set is called the third quartile  $Q_3$ , and the median of the entire data set is the second quartile  $Q_2$ .

The **interquartile range** is the difference between the median of the upper half and the median of the lower half, or  $Q_3 - Q_1$ . Let's see how it's done.

If we list out all of our golf scores from the stem chart, the data set is

66, 67, 67, 68, 68, 68, 68, 69, 69, 69, 69, 70, 70, 71, 71, 72, 73, 75

The median is  $(69 + 69)/2 = 69$ .

~~66, 67, 67, 68, 68, 68, 68, 69, 69, 69, 70, 70, 71, 71, 72, 73, 75~~

To find the IQR, we'll now split the data in half. Since this data set has 18 data points, we'll have 9 data points in the lower half, and 9 data points in the upper half. We then need to find the median of each half.

The median of the lower half is 68.

~~66, 67, 67, 68, 68, 68, 68, 69, 69~~



The median of the upper half is 71.

~~69, 69, 70, 70, 71, 71, 72, 73, 75~~

To see this visually, let's look at the numbers we picked out from the original data set.

66, 67, 67, 68, **68**, 68, 68, 69, 69, 69, 69, 70, 70, **71**, 71, 72, 73, 75

Median of the  
lower half

Median

Median of the  
upper half

Now that we have the median of both halves of the data, we can find the interquartile range by taking the difference of those medians. For this data set, the IQR is  $71 - 68 = 3$ .

Notice that in this data set of golf scores we had an even number of data points, and we therefore just divided the number of data points in two to get 9 data points in the lower half and 9 data points in the upper half.

We also need to know how to calculate the IQR when we have an odd number of data points. This data set has 11 data points:

66, 67, 68, 69, 69, 69, 70, 71, 71, 72, 75

If we find the median, we can see that it's 69.

~~66, 67, 68, 69, 69, 69, 70, 71, 71, 72, 75~~

To separate the data into two halves with an equal number of data points in each, we'll take the lower half as everything below the median (not including the median), and the upper half as everything above the median



(not including the median). So the lower half will be 66, 67, 68, 69, 69, and its median is 68. The upper half will be 70, 71, 71, 72, 75 and its median is 71. The IQR of this data set with an odd number of data points is therefore  $71 - 68 = 3$ .

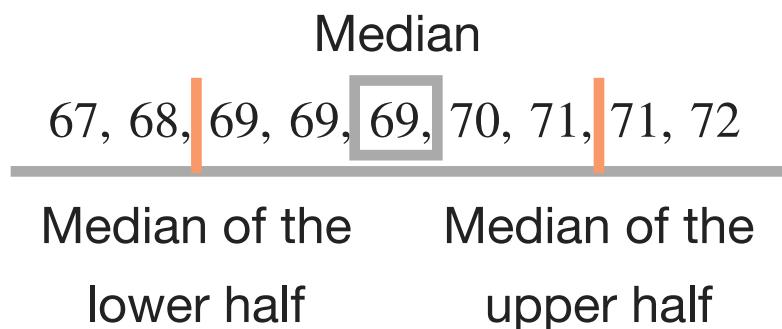
We can also have a data set with an odd number of data points where the median is the middle number, but then the upper and lower half of the data includes an even number of data points, so the median of each half will be the mean of two numbers. If we take away the first and last data point from the data set above, then the new data set is

67, 68, 69, 69, 69, 70, 71, 71, 72

The median of this set is 69.

~~67, 68, 69, 69, 69, 70, 71, 71, 72~~

But then the lower half of the data set is 67, 68, 69, 69 and the upper half of the data set is 70, 71, 71, 72. To find the median of these halves, we use the same process we always use to find the median of a data set with an even number of data points. We find the middle two numbers, and then take their mean. The median of the lower half, 67, 68, 69, 69, is  $(68 + 69)/2 = 68.5$ . The median of the upper half, 70, 71, 71, 72, is  $(71 + 71)/2 = 71$ .



Therefore, the IQR of this data set is  $71 - 68.5 = 2.5$ .

# Changing the data, and outliers

In this section, we want to see what happens to our measures of central tendency and spread when we make changes to our data set. Specifically the changes made either by changing all the values in the set at once, or by adding a single data point to, or removing a single data point from, the data set.

## Changing the entire data set

### Shifting (addition and subtraction)

What happens to measures of central tendency and spread when we add a constant value to every value in the data set? To answer this question, let's pretend we have the data set 3, 3, 7, 9, 13, and let's calculate our measures for the set.

$$\text{Mean: } (3 + 3 + 7 + 9 + 13)/5 = 7$$

$$\text{Median: } 7$$

$$\text{Mode: } 3$$

$$\text{Range: } 13 - 3 = 10$$

$$\text{IQR: } 11 - 3 = 8$$

If we add 6 to each data point in the set, the new set is 9, 9, 13, 15, 19. And our new measures of central tendency and spread are



Mean:  $(9 + 9 + 13 + 15 + 19)/5 = 13$

Median: 13

Mode: 9

Range:  $19 - 9 = 10$

IQR:  $17 - 9 = 8$

What we see is that adding 6 to the entire data set also adds 6 to the mean, median, and mode, but that the range and IQR stay the same.

And this will always be true. No matter what value we add to the set, the mean, median, and mode will shift by that amount but the range and the IQR will remain the same. The same will be true if we subtract an amount from every data point in the set: the mean, median, and mode will shift to the left but the range and IQR will stay the same.

So to summarize, whether we add a constant to each data point or subtract a constant from each data point, the mean, median, and mode will change by the same amount, but the range and IQR will stay the same.

### Scaling (multiplication and division)

Let's look at what happens when we multiply our data set by a constant value. Again starting with the set 3, 3, 7, 9, 13, the measures are

Mean:  $(3 + 3 + 7 + 9 + 13)/5 = 7$

Median: 7

Mode: 3



Range:  $13 - 3 = 10$

IQR:  $11 - 3 = 8$

Let's multiply the set by 2, making the new set 6, 6, 14, 18, 26. The new measures of central tendency and spread are

Mean:  $(6 + 6 + 14 + 18 + 26)/5 = 14$

Median: 14

Mode: 6

Range:  $26 - 6 = 20$

IQR:  $22 - 6 = 16$

What we see is that multiplying the entire data set by 2 multiplies all five measures by 2 as well. The mean, median, mode, range, and IQR are all doubled when we double the values in the data set.

And this will always be true. No matter what value we multiply by the data set, the mean, median, mode, range, and IQR will all be multiplied by the same value. The same will be true if we divide every data point in the set by a constant value: the mean, median, mode, range, and IQR will all be divided by the same value.

So to summarize, if we multiply our data set by a constant value or divide our data set by a constant value, then the mean, median, mode, range, and IQR will all be scaled by the same amount.



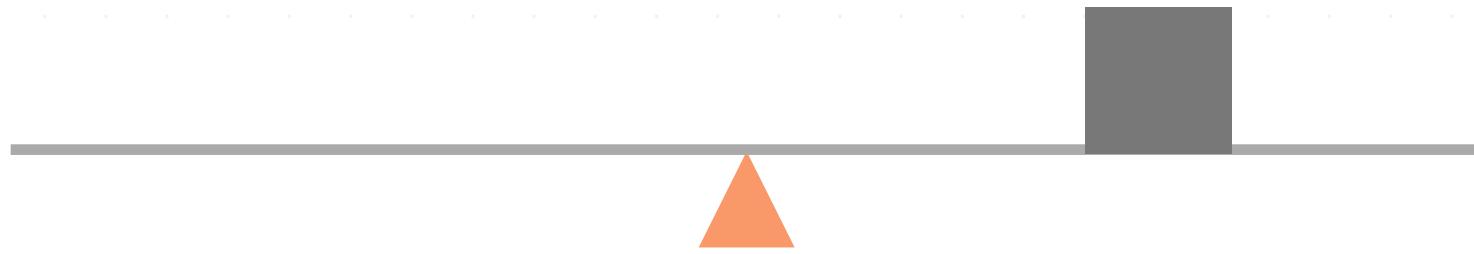
## Adding or removing a data point from the set

### Mean

Thinking back to our discussion about the mean as a balancing point, we want to realize that adding another data point to the data set will naturally effect that balancing point. In fact, adding a data point to the set, or taking one away, can effect the mean, median, and mode.

If we add a data point that's above the mean, or take away a data point that's below the mean, then the mean will increase. If take away a data point that's above the mean, or add a data point that's below the mean, the mean will decrease.

Adding or removing a number on one side of the mean will force us to move the mean if we want to stay balanced



The mean is at the balancing point

### Median

If we add or remove a data point from the set, it can effect the median, but it may not. In the set 1, 2, 3, 4, 4, 6, 6, the median is 4. If we take out 3, the median of 1, 2, 4, 4, 6, 6 is still 4; it's unchanged. But if we take out a 6, the

median of 1, 2, 3, 4, 4, 6 is now 3.5; it changes. The same will be true for adding in a new value to the data set. Depending on the value, the median might change, or it might not.

### Effect on the mean vs. median

It's also important that we realize that adding or removing an extreme value from the data set will affect the mean more than the median.

Let's take an easy example, and use the data set 1, 2, 3. The mean is 2 and the median is 2. Let's add a huge value to the data set, like 1,000, so that the new data set is 1, 2, 3, 1,000. The mean of this new data set is about 252, and the median of the new data set is 2.5.

What we see is that adding an extreme value to the data set barely had any effect on the median at all: it went up from 2 to 2.5. But adding the new value had an enormous effect on the mean: it shifted the mean from 2 up to 252.

### Example

Let's say we play a round of golf with three friends, and our scores are the set 70, 71, 71, 103. What effect does removing the 103 have on the mean and median of the set?

In a set like this one, we have a few data points clustered tightly together, and then a data point that is much different than the others. Removing the data point that's far from the cluster effects the mean and median in



interesting ways. We can see that the median of the set is 71, and we can calculate that the mean is

$$\mu = \frac{70 + 71 + 71 + 103}{4} = \frac{315}{4} \approx 79$$

If we remove the 103 from the data set, the median doesn't change at all because the median of the set 70, 71, 71 is still 71. But the mean will change significantly. The new mean is

$$\mu = \frac{70 + 71 + 71}{3} = \frac{212}{3} \approx 71$$

Which makes sense, because the single data point of 103 would tend to skew the data more by bringing up the average. So when it's removed, the mean drops back down to a value that more accurately reflects most of the scores. On the other hand, the 103 barely changes the median, which is why the median didn't change when we removed the 103.

A number that has the power to change a data set in this way is called an **outlier**; it's a number on the extreme upper end or extreme lower end of a data set.

## Mode

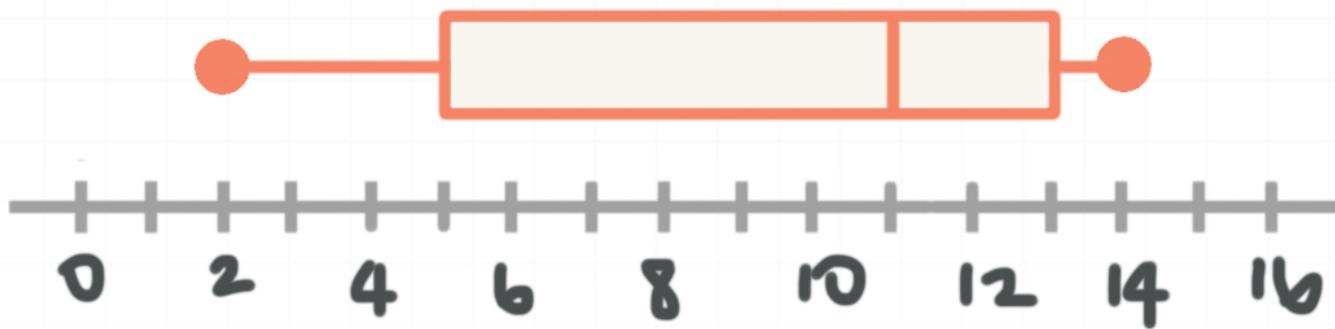
The mode could also be effected by adding a data point or taking one away. For example, in the set 1, 2, 3, 4, 4, 6, 7, we could add a 4 and it wouldn't change the mode. We could also take away a 2, and it wouldn't change the mode. But, if we were to take away a 4, the mode of the set would change from 4 to the set having no mode at all.



# Box-and-whisker plots

**Box-and-whisker plots** (also called box plots) are a great way to represent a data set when we want to show the median and spread of the data at the same time.

In general, a box-and-whisker plot might look like this:



The big rectangle in the center is the **box**, and the little lines extending out from the sides are the **whiskers**.

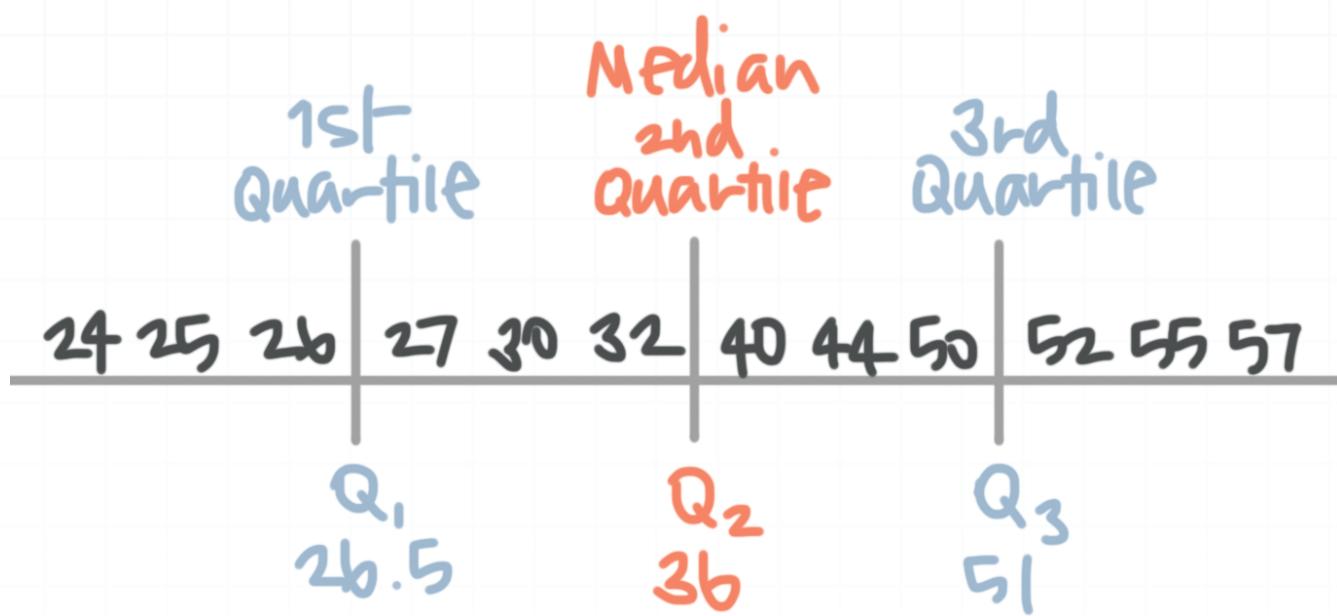
The great thing about a box plot is that we know the median, range, upper and lower bounds just by looking at it. The interquartile range is also just a simple calculation.

The vertical line inside the box is the median of the data set, so the median of the data set represented in the plot above is 11.

The dot at the end of the left whisker is the minimum of the data set, and the dot at the end of the right whisker is the maximum of the data set. So in this plot, we can say that the minimum is 2, that the maximum is 14, and so we know right away that the range of the data is  $14 - 2 = 12$ .

The IQR is given by the ends of the box. Since the box above extends from 5 to 13, the IQR is  $13 - 5 = 8$ .

The box-and-whisker plot also shows us where each quartile of the data is located. A **quartile** is a number that divides the data set into quarters. The first quartile,  $Q_1$ , separates the lowest 25% of data points from the second 25%. The second quartile,  $Q_2$ , is the median, and it separates the data set into halves. The third quartile,  $Q_3$ , separates the third 25% of data points from the upper 25% of data points.



In a box-and-whisker plot, the left end of the box represents  $Q_1$ , the median represents  $Q_2$ , and the right end of the box represents  $Q_3$ . Based on the box plot above,

- 25% of the data points lie between 2 and 5
- 25% of the data points lie between 5 and 11
- 25% of the data points lie between 11 and 13
- 25% of the data points lie between 13 and 14

Therefore,

- 5 is the first quartile

- 11 is the second quartile
- 13 is the third quartile

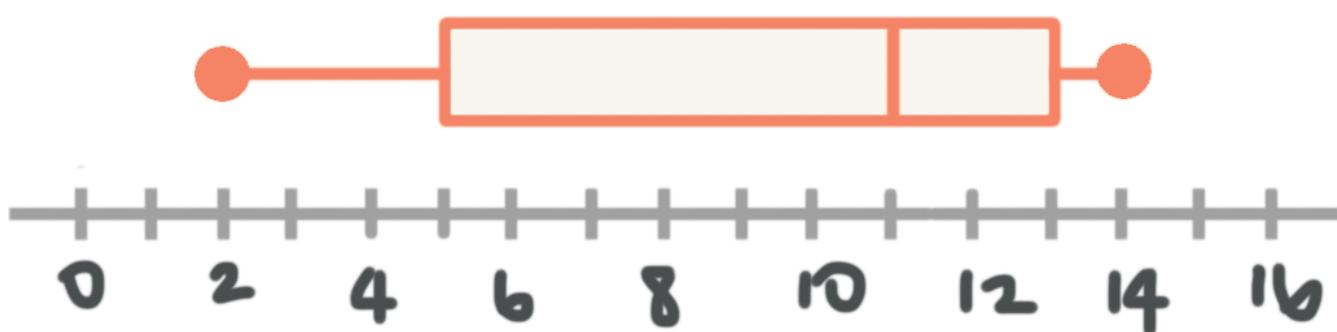
In a box-and-whisker plot, the middle 50 % of the data is represented inside the box, the lowest 25 % in the whisker on the left, and the highest 25 % in the whisker on the right.

As far as our quarters, this means that the first quarter is represented by the whisker on the left, the second quarter is represented by the part of the box to the left of the median, the third quarter is represented by the part of the box to the right of the median, and the fourth quarter is represented by the whisker on the right.

## Five-number summary

The **five-number summary**, also called the five-figure summary, for any set of data will include the minimum and maximum values, the median, and  $Q_1$  and  $Q_3$  for the data set. We usually give the five-number summary in a table, and we can easily gather all of this information from a box-plot.

The five-number summary for the box plot



is

Min	$Q_1$	Median	$Q_3$	Max
2	5	11	13	14

# Mean, variance, and standard deviation

Before we dive into standard deviation and variance, it's important for us to talk about populations and population samples. A **population** is the entire group of subjects that we're interested in. A **sample** is just a subsection of the population.

So, as an example, if we're interested in data about polar bears in the arctic, the population would be every single polar bear in that region. It would be very difficult, if not impossible, for us to ensure we'd looked at every polar bear. So we might choose instead to take a sample of the population, maybe only 25 bears, and use the data we collect about that smaller group in order to draw conclusions about the population as a whole.

If, on the other hand, we were interested in data about all the students in our math class, there might only be 30 other students, so it might be very reasonable for us to collect data about the entire population.

It's important to know whether we're talking about a population or a sample, because in this section we'll be talking about variance and standard deviation, and we'll use different formulas for variance and standard deviation depending on whether we're using data from a population or data from a sample.

In all the formulas we use that involve a count of the number of subjects or participants, we'll denote the number of subjects in a population as capital  $N$ , and the number of subjects in a sample as lowercase  $n$ .



## Mean

We learned previously that the formula for the mean was

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

Now that we're a little more advanced and we want to start distinguishing between populations and samples, let's update the mean formula and say that the **mean of a population** is

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

The mean of a population is still defined as  $\mu$ , but we'll define the **mean of a sample** with  $\bar{x}$ , pronounced "x-bar":

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Notice the capital  $N$  in the population formula and the lowercase  $n$  in the sample formula. Remember the capital  $N$  means we've included everyone (the population), and the lowercase  $n$  means we've selected just a few individuals (the sample).

## Variance

**Variance** is the measure of how far the data is spread from the mean. Population variance is given by  $\sigma^2$  (pronounced "sigma squared"). The



reason we define the population variance formula in terms of  $\sigma^2$  is because doing so will help us with some concepts we'll learn later on. The formula for population variance is:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Notice that  $\mu$  is the population mean, which means that  $x_i - \mu$  gives the distance of each point from the mean, which is the deviation of each point. Then  $(x_i - \mu)^2$  is the squared deviation, we're summing together all those squared deviations in the numerator, and then we're dividing that result by the number of objects in the population,  $N$ , in order to get population variance,  $\sigma^2$ .

Finding **sample variance** is a very similar process to finding population variance, but we use a slightly different formula:

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

Notice that the formula for sample variance,  $s^2$ , is identical to the formula for population variance, except that we've swapped out  $\mu$  for  $\bar{x}$  (since  $\bar{x}$  is sample mean, whereas  $\mu$  is population mean), and we've changed  $N$  to  $n$  (since  $n$  refers to sample size, whereas  $N$  refers to population size).

But we need to be really careful here. While this sample variance formula is correct, it's not usually the one we use, because it's actually not that accurate. We won't go into detail about why it's not super accurate, but we'll say that, because it's not that accurate, we usually say that the formula above gives **biased sample variance**.



Interestingly, the easy way to make the sample variance formula a lot more accurate is to divide by  $n - 1$  instead of  $n$ . Dividing by  $n$  will underestimate sample variance, and dividing by  $n - 2$  will overestimate sample variance. In other words, the better formula for sample variance, and therefore the one we want to use is

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

We say that this formula gives us the **unbiased sample variance**. Sometimes, in order to distinguish these formulas from one another, we'll see them written as

$$s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \text{ for biased sample variance}$$

and

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \text{ for unbiased sample variance}$$

However, because the formula for unbiased sample variance always gives us a more accurate figure for the variance of a sample, very often we won't worry about indicating the left-hand side of the formula as  $s_n^2$  or  $s_{n-1}^2$ , because we just assume that we always want unbiased sample variance. And therefore, we agree that the formula we always want to use for sample variance is this one:

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



Let's do an example where we have to calculate the mean and variance for a data set.

### Example

A school principal wants to know the average age of teachers in her school. She took a sample of 10 teachers and recorded their ages. Find the sample mean and sample variance.

43, 35, 38, 56, 29, 33, 46, 63, 49, 40

Let's find the sample mean first.

$$\bar{x} = \frac{43 + 35 + 38 + 56 + 29 + 33 + 46 + 63 + 49 + 40}{10} = 43.2$$

Now we'll use a table to calculate sample variance. We'll first calculate how far each data point is from the sample mean to get the deviation of each data point, and then we can calculate squared deviations.

Data	Deviations	Squared deviations
43	43-43.2=-0.2	0.04
35	35-43.2=-8.2	67.24
38	38-43.2=-5.2	27.04
56	56-43.2=12.8	163.84
29	29-43.2=-14.2	201.64
33	33-43.2=-10.2	104.04
46	46-43.2=2.8	7.84
63	63-43.2=19.8	392.04
49	49-43.2=5.8	33.64
40	40-43.2=-3.2	10.24

Find the sum of the squared deviations, and divide the result by  $n - 1$ .

$$s^2 = \frac{0.04 + 67.24 + 27.04 + 163.84 + 201.64 + 104.04 + 7.84 + 392.04 + 33.64 + 10.24}{10 - 1}$$

$$s^2 \approx 111.96$$

Therefore, the mean age of teachers, based on the sample, is 43.2. The sample variance is 111.96.

## Standard deviation

**Standard deviation** is a measure of how much the data in a set varies from the mean. The larger the value of standard deviation, the more the data in

the set varies from the mean. The smaller the value of standard deviation, the less the data in the set varies from the mean.

**Population standard deviation** is the positive square root of population variance. Since population variance is given by  $\sigma^2$ , population standard deviation is given by  $\sigma$ .

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

So when we want to calculate the standard deviation for a population, just find population variance, and then take the square root of the variance, and we'll have population standard deviation.

Similarly, we'll find **sample standard deviation** by taking the square root of unbiased sample variance (the one we found by dividing by  $n - 1$ ). Since sample variance is given by  $s^2$ , sample standard deviation is given by  $s$ .

$$s = \sqrt{s_{n-1}^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Keep in mind that, even though we start with unbiased sample variance, when we take the square root to find sample standard deviation, we reintroduce some bias into the value. The amount of bias in the sample standard deviation just depends on the kind of data in the data set.

Here's a table that summarizes the formulas from this section.



	<b>Population</b>	<b>Sample</b>
# of subjects	$N$	$n$
Mean	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$

Note:  $s^2$  is the formula for unbiased sample variance, since we're dividing by  $n - 1$ .

<b>Standard deviation</b>	$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$
---------------------------	--	---

Note: Finding  $s$  by taking  $\sqrt{s^2}$  reintroduces bias.

Let's extend our previous example.

### Example

A school principal wants to know the average age of teachers in her school. She takes a sample of 10 teachers and finds  $\bar{x} = 43.2$  and  $s^2 \approx 111.96$ . What is the standard deviation of her sample?

We need to take the square root of the sample variance in order to find sample standard deviation.

$$s = \sqrt{s^2} = \sqrt{111.96} \approx 10.58$$



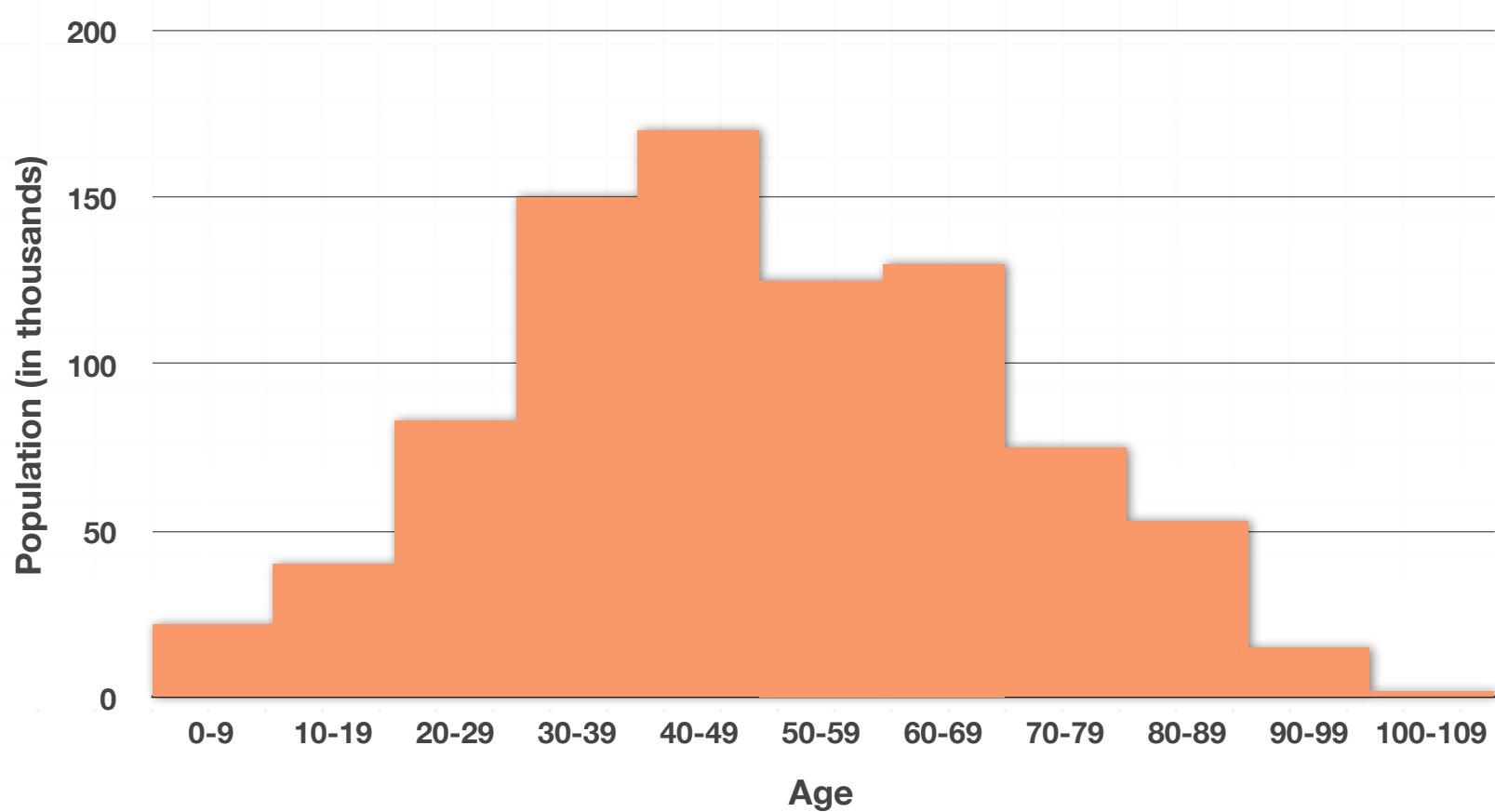
The sample standard deviation is approximately 10.58.

---



# Frequency histograms and polygons, and density curves

Earlier we learned about creating histograms by collecting the data in our set into small groups, and then graphing each group together. The grouping of data points is what makes it a histogram instead of just a bar graph. Each bar essentially shows the frequency of that group. In other words, in the histogram below,



if we want to know the frequency at which 30 – 39 year-olds occur in the data, we just look at the bar to see that there are about 150,000 30 – 39 year-olds in the data set.

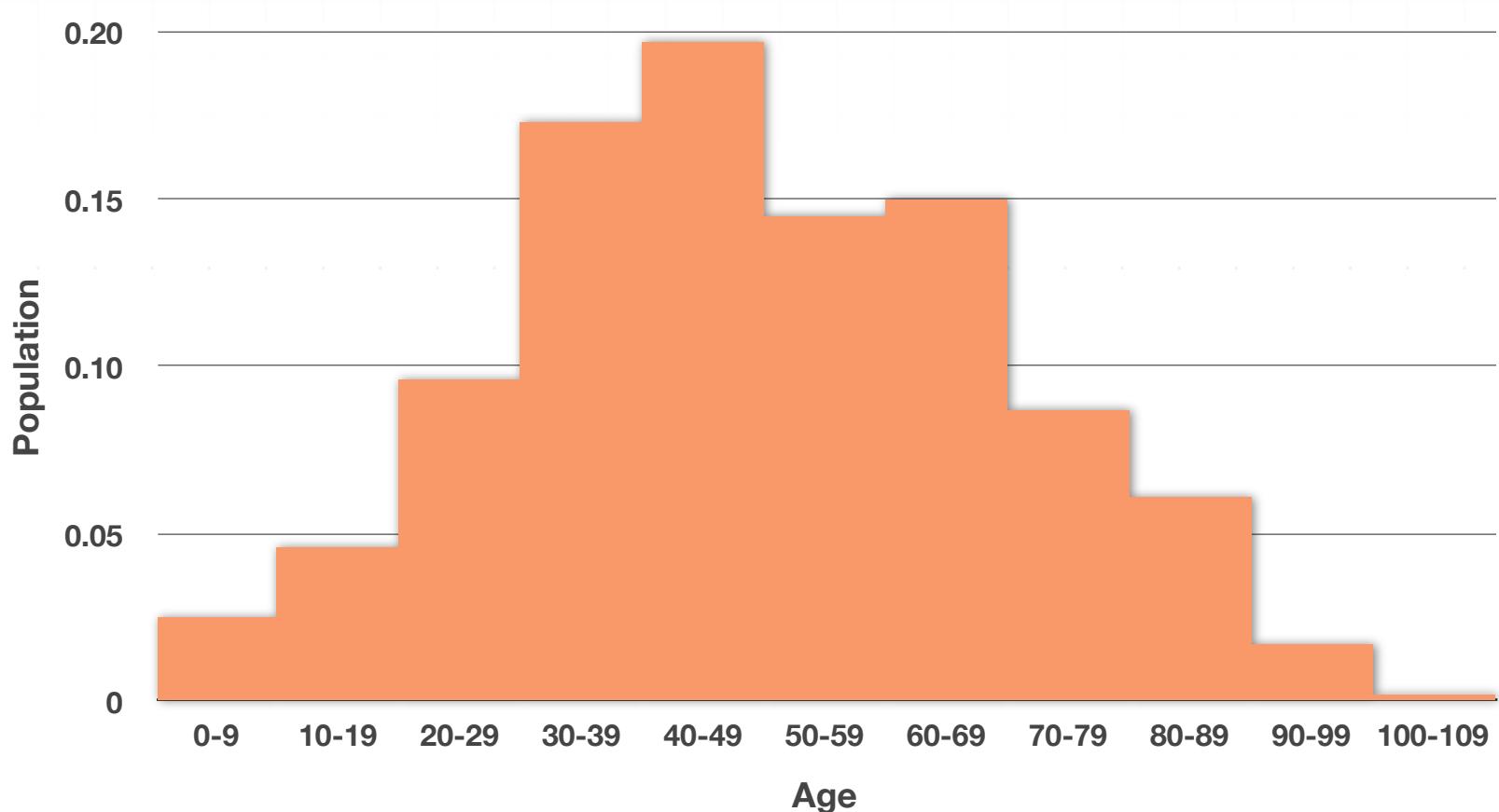
For this reason, a histogram is often also called a **frequency histogram**, since it shows the frequency at which each category occurs.

## Relative frequency histogram

We can also make a **relative frequency histogram**, which is the same as a regular histogram, except that we display the frequency of each category as a percentage of the total of the data.

In the histogram above, we're showing the number of people in each interval, and there are 865,000 total people in the data set. To find the percentage of people represented in the 30 – 39 age group, we'd take the number of people in that group, 150,000, and divide by the total number of people in the data set, 865,000. We'd see that  $150,000/865,000 \approx 0.173$ .

Which means that group represents about 17.3 % of the data. If we repeat that process for the rest of the groups, then we can put the new data into a relative frequency histogram:

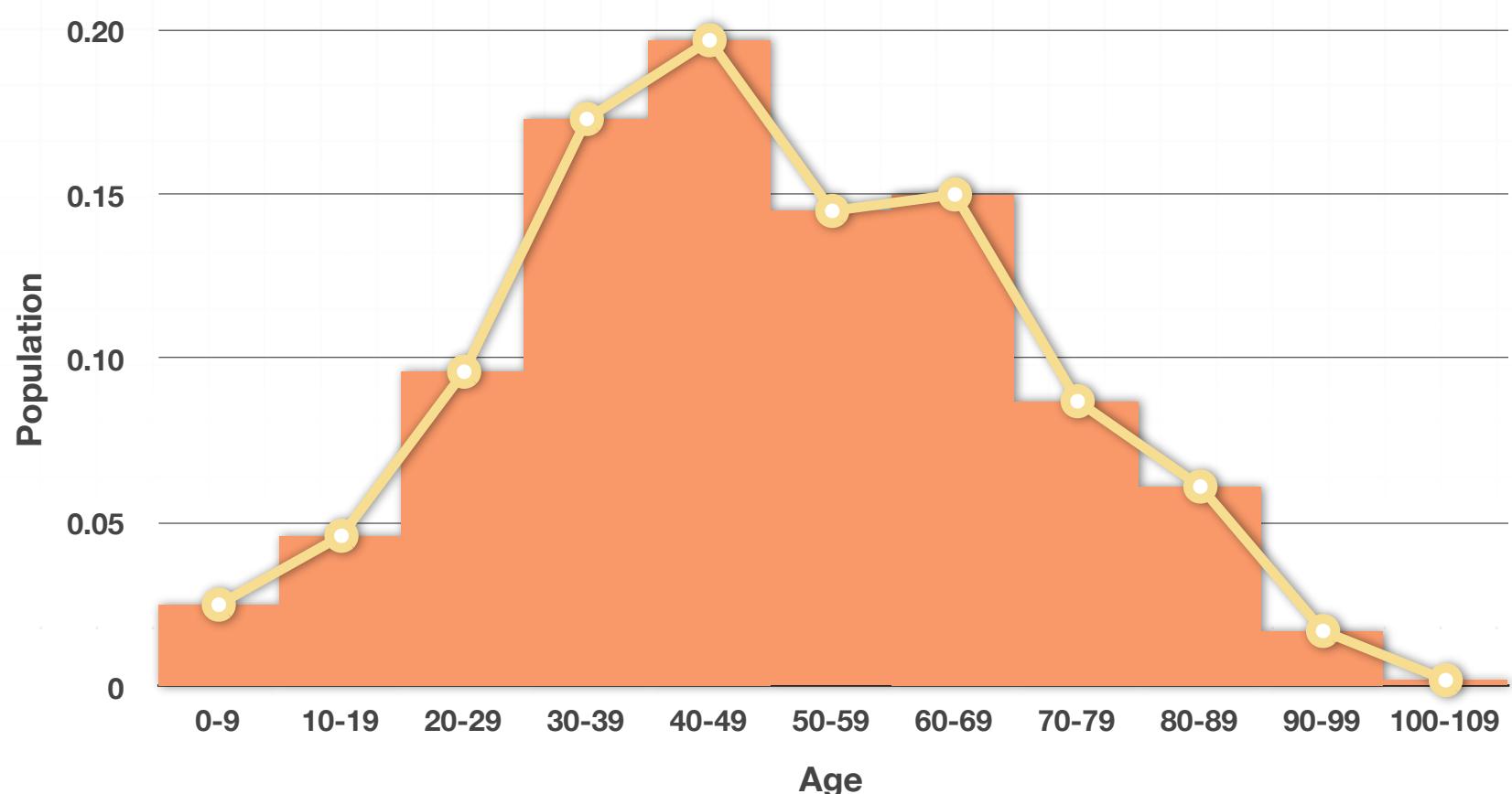


Notice that we marked off the  $y$ -axis differently. We can see from this relative frequency histogram that the largest group in the data set is

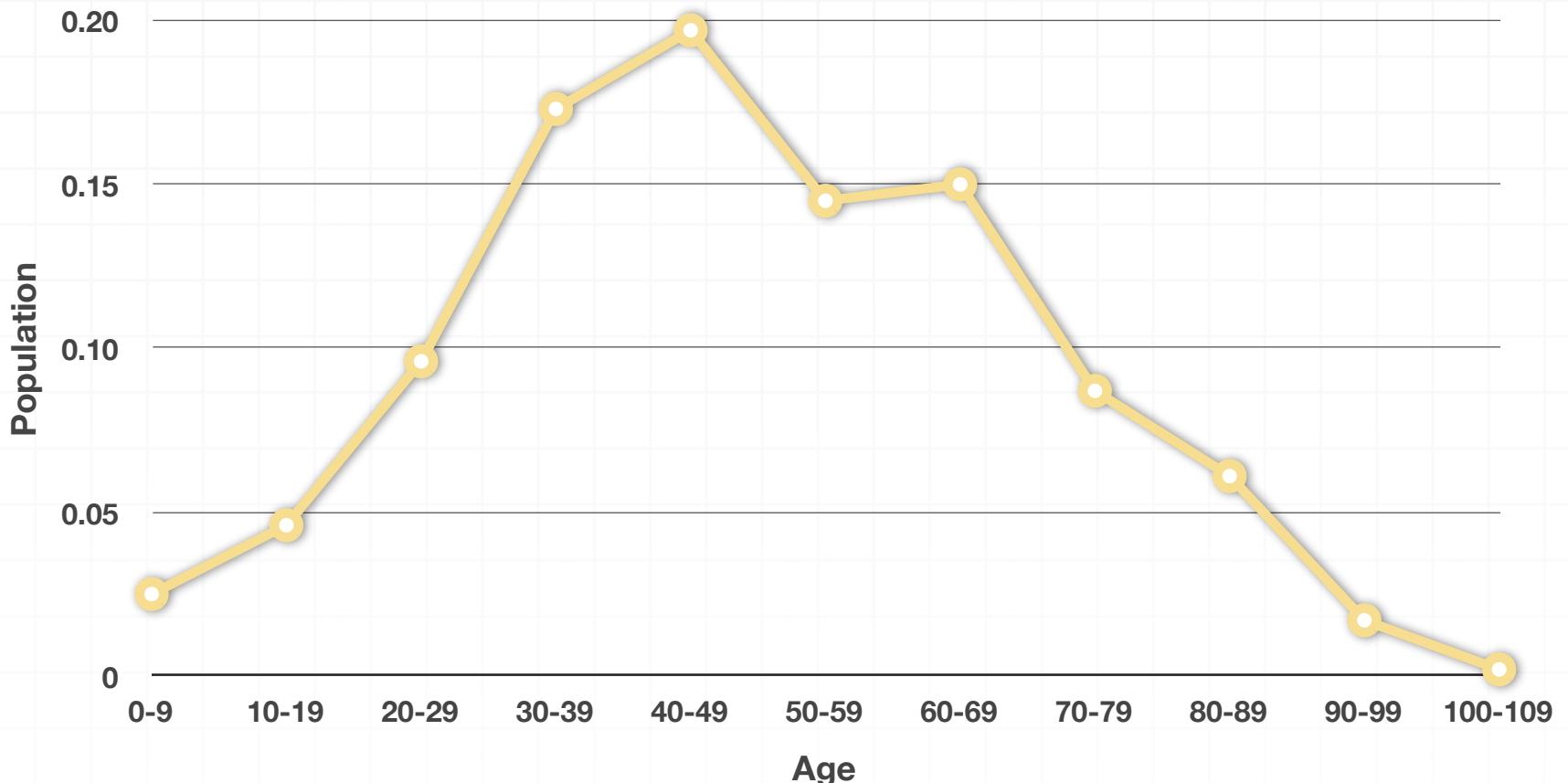
40 – 49 year-olds, and that they represent just under 20 % of the total population.

## Frequency polygon

We can turn any histogram or relative frequency histogram into a **frequency polygon** by connecting the top of each bar with a line. Our relative frequency histogram becomes



Then we remove the bars, leaving only the line graph.



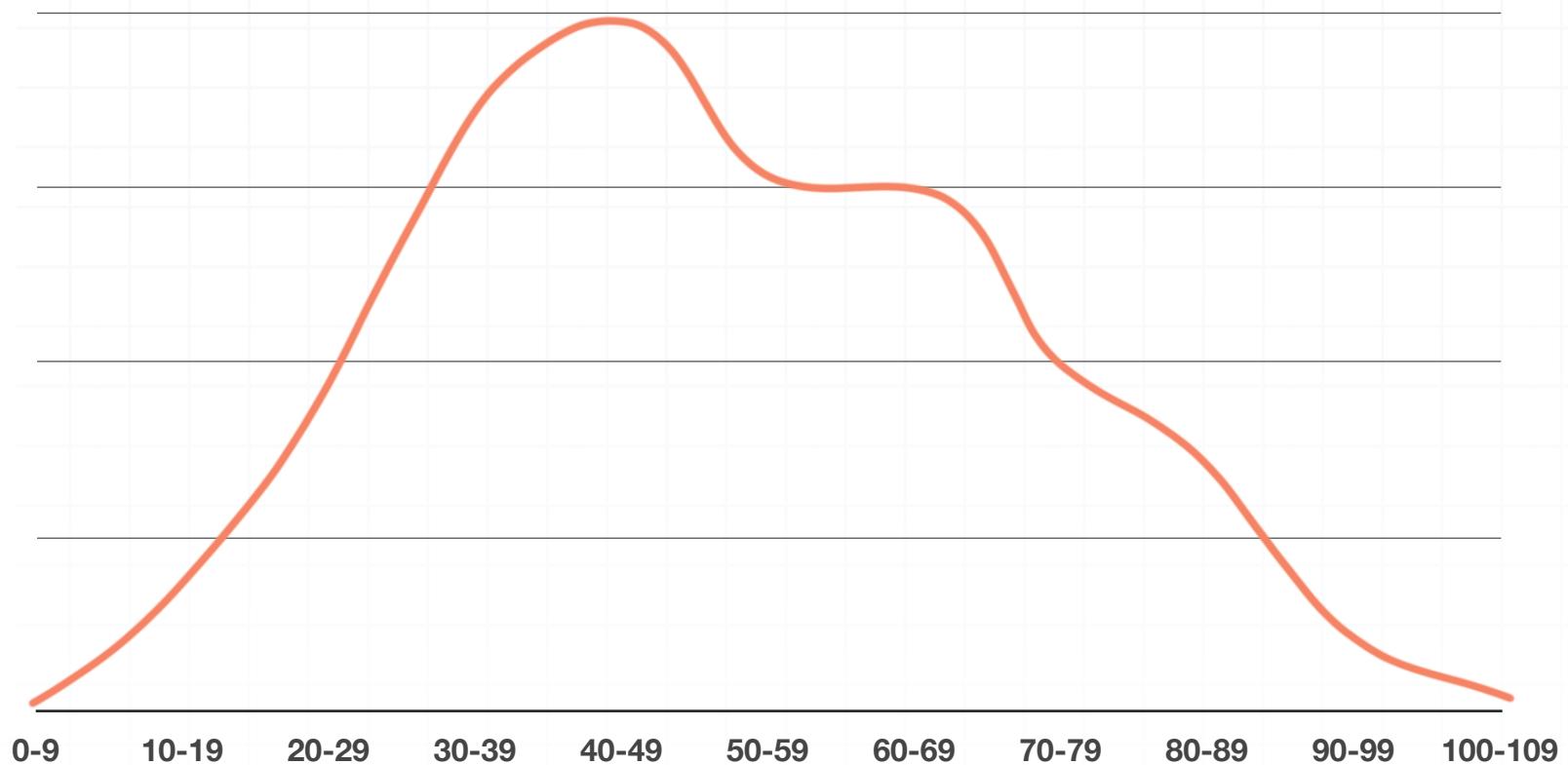
This is now a frequency polygon, so-called because it's a polygon-shaped figure that shows the frequency at which each age range occurs in the data set. Frequency polygons are nice because they can give us a visual glance at the distribution of the data set.

## Density curve

In the histogram we've been using, we had our data grouped into 11 categories based on age. From that, we were able to see the “density” of where most of our data was occurring. In this particular histogram, for example, most of the data occurs between age 30 and age 69.

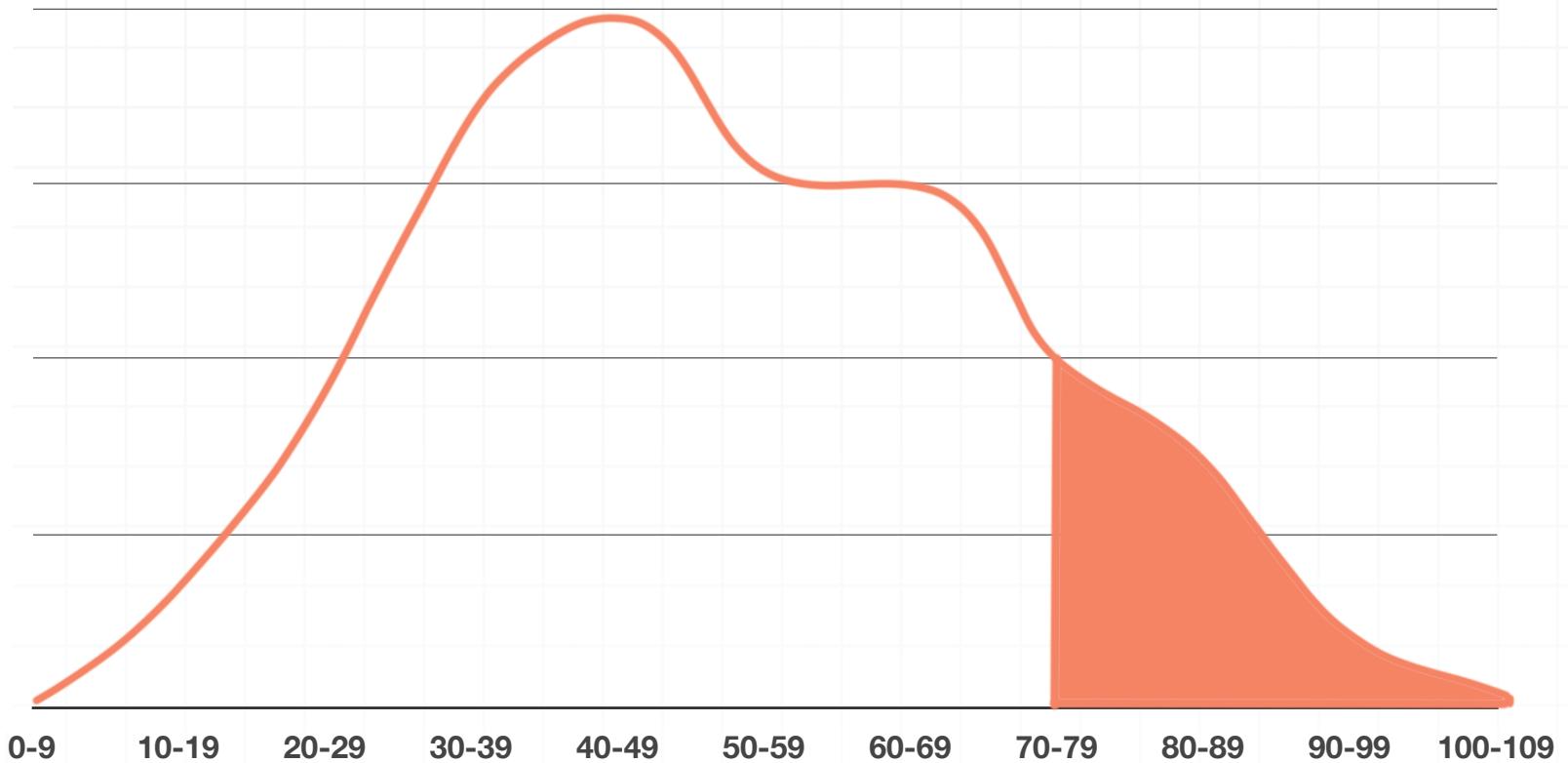
If we were to use more and more categories, instead of just 11, until finally we were using infinitely many categories, and the bars in the histogram were infinitely thin, we'd be able to create a perfectly smooth curve by

connecting the tops of each bar with a smooth line. This smooth curve is called a **density curve**.



There are a couple of important things we want to remember about density curves. First, the area under a density curve will always represent 100% of the data, or 1.0. The curve will never dip below the  $x$ -axis.

If we want to know how much of our data falls within a certain interval, then we want to look at the amount of total area that falls under the curve within that interval. As an example, if we want to know how much of the population is roughly age 70 and older, we're looking at the area under the curve on that interval:

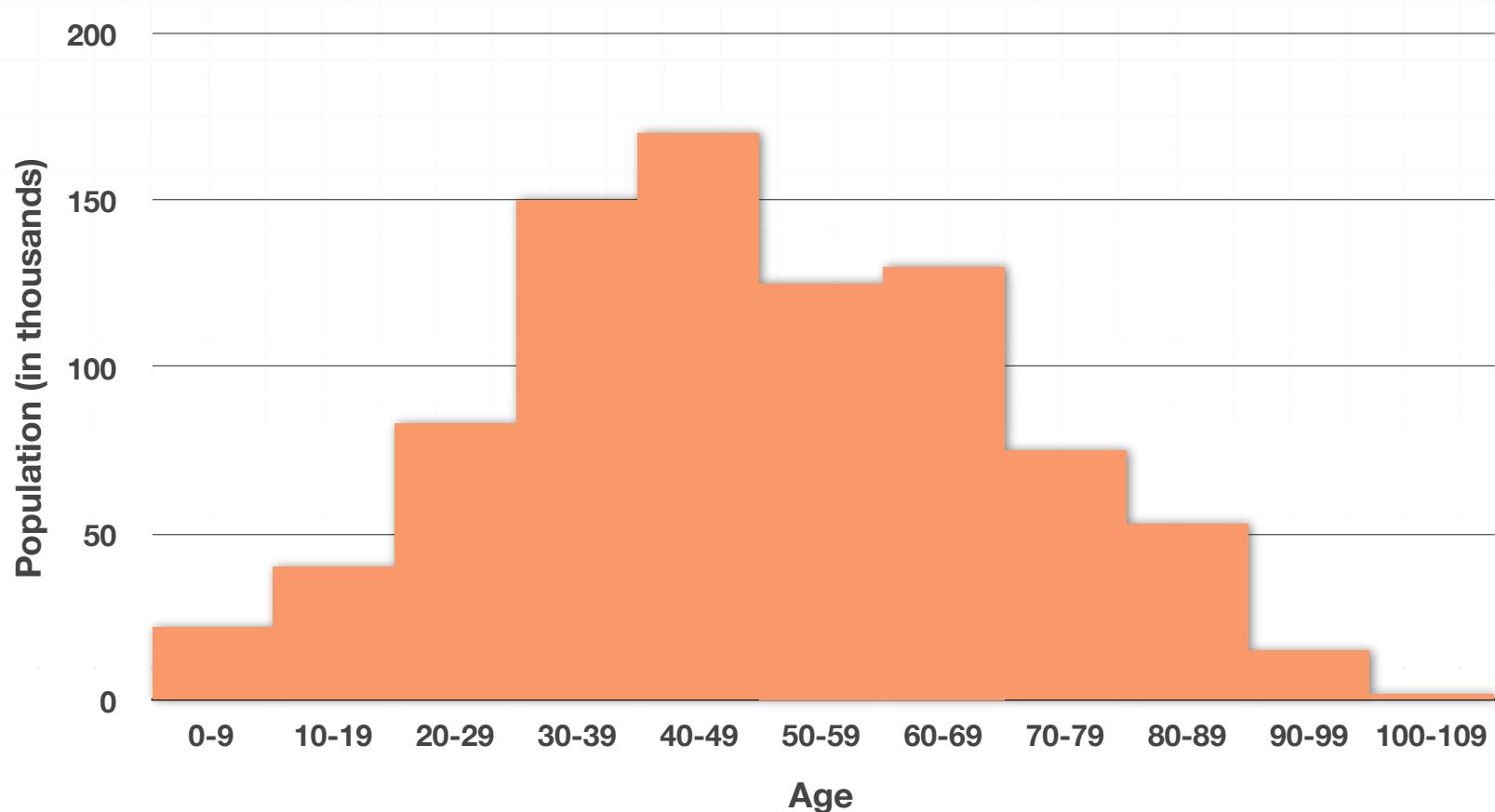


We're estimating here, but that looks like it might be roughly 25 % of the total area under the curve, so we might say that about 25 % of our population is age 70 or older.

# Symmetric and skewed distributions and outliers

A density curve is technically the smooth line that encloses a **distribution**. We call it a distribution because the area under the curve shows us the distribution of our data.

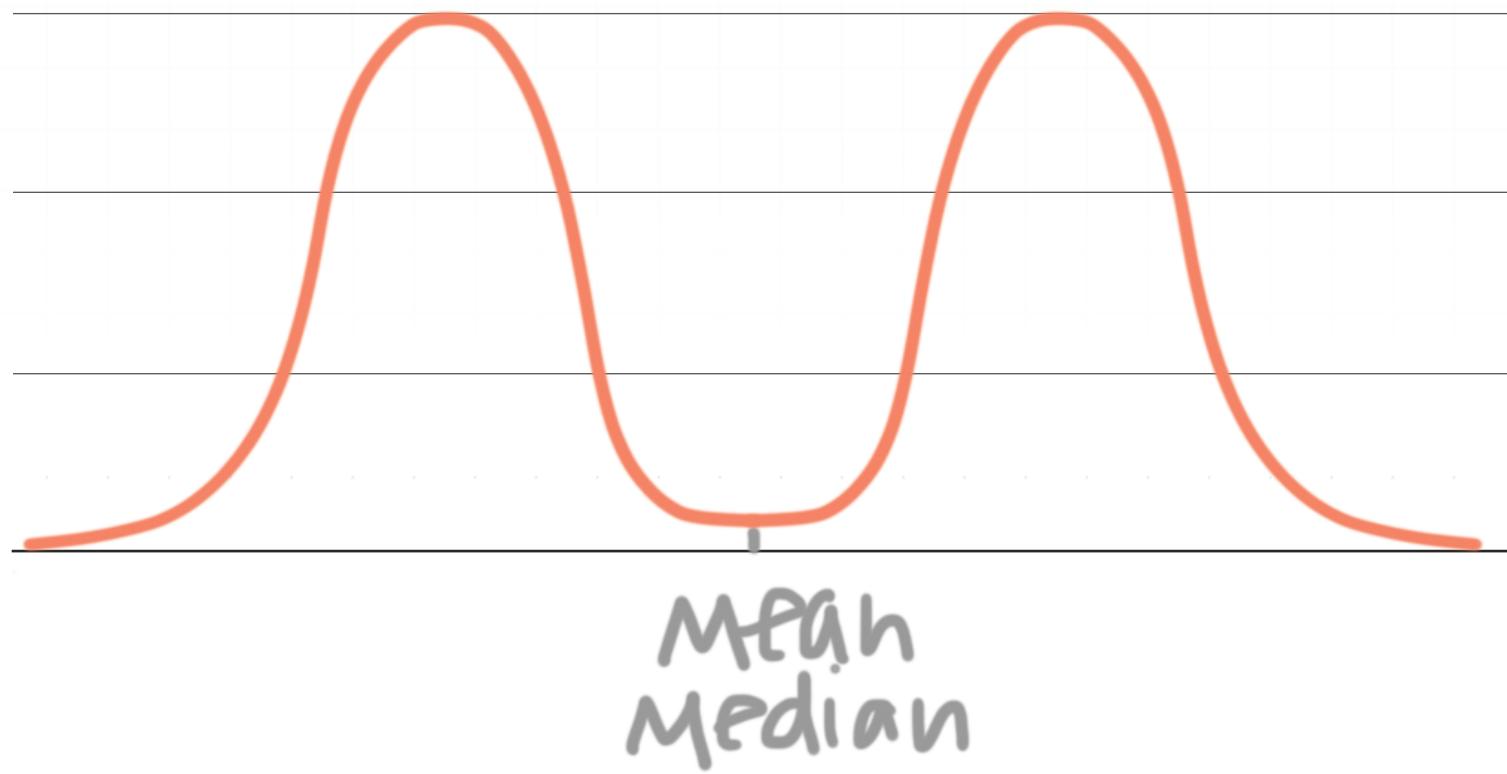
For example, in the distribution we drew for the ages of San Francisco residents,



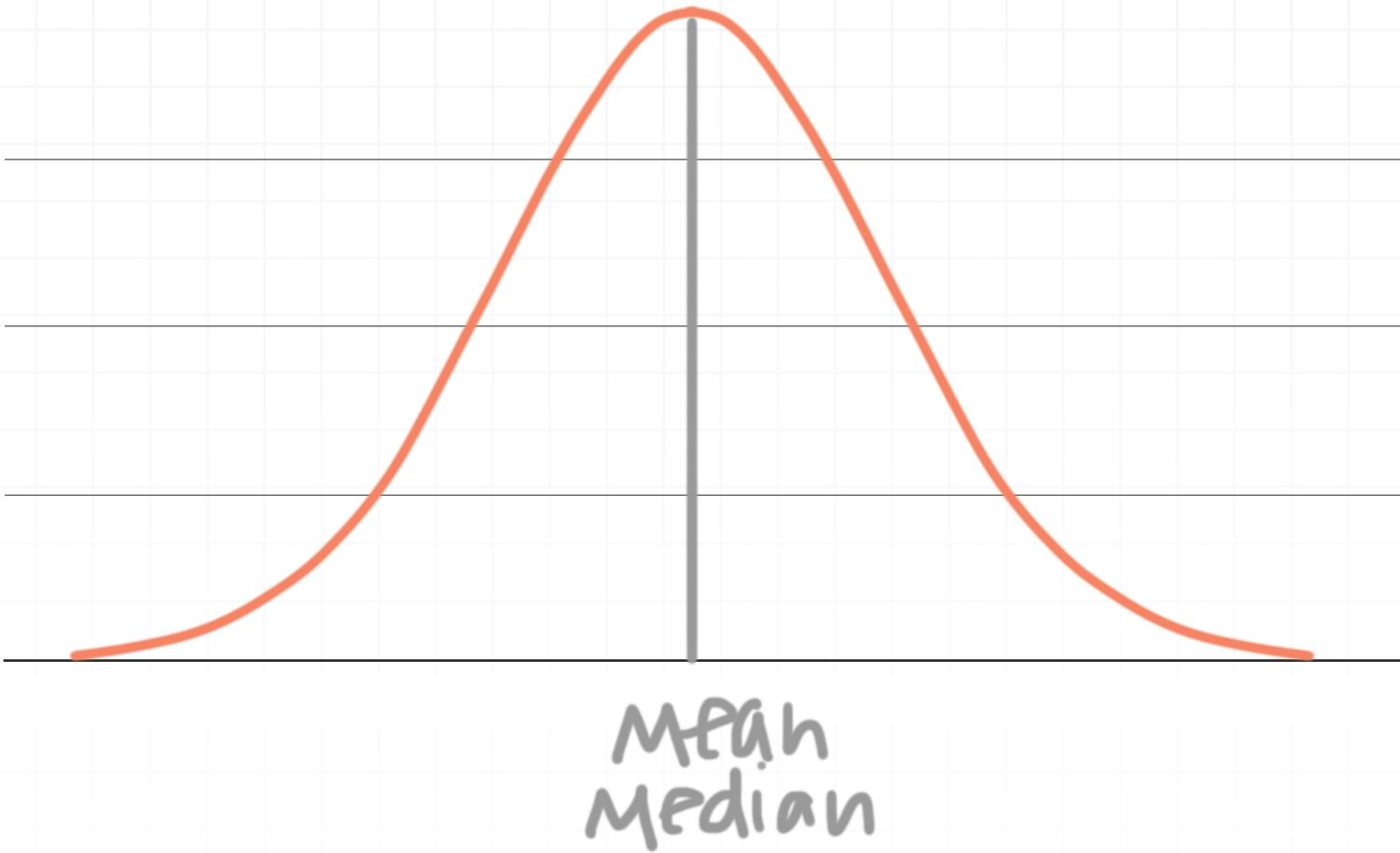
we were able to see roughly how much of the population was age 70 or older, and we could similarly estimate how much of the population was between age 30 and age 59. The ages of all of the people in the population are “distributed” between age 0 and 109.

## Symmetric distributions

When a density curve is perfectly symmetric, then the mean and the median are both at the very center of the distribution. The mean and median for a symmetric distribution will always be wherever there's an equal amount of area on the left and right. This is one example of a symmetric, non-normal distribution:



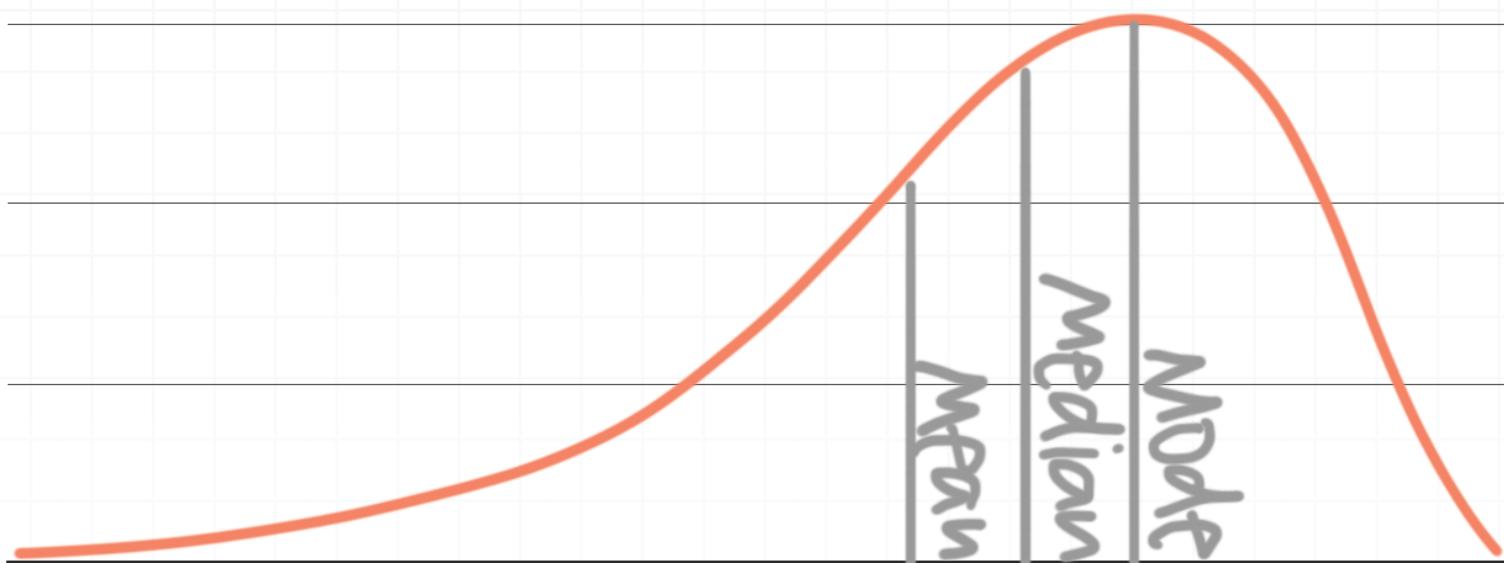
Symmetric distributions can be any shape (as long as they're symmetric, of course), but we'll deal a lot with what we call a **normal distribution**, which is a symmetric, bell-shaped distribution:



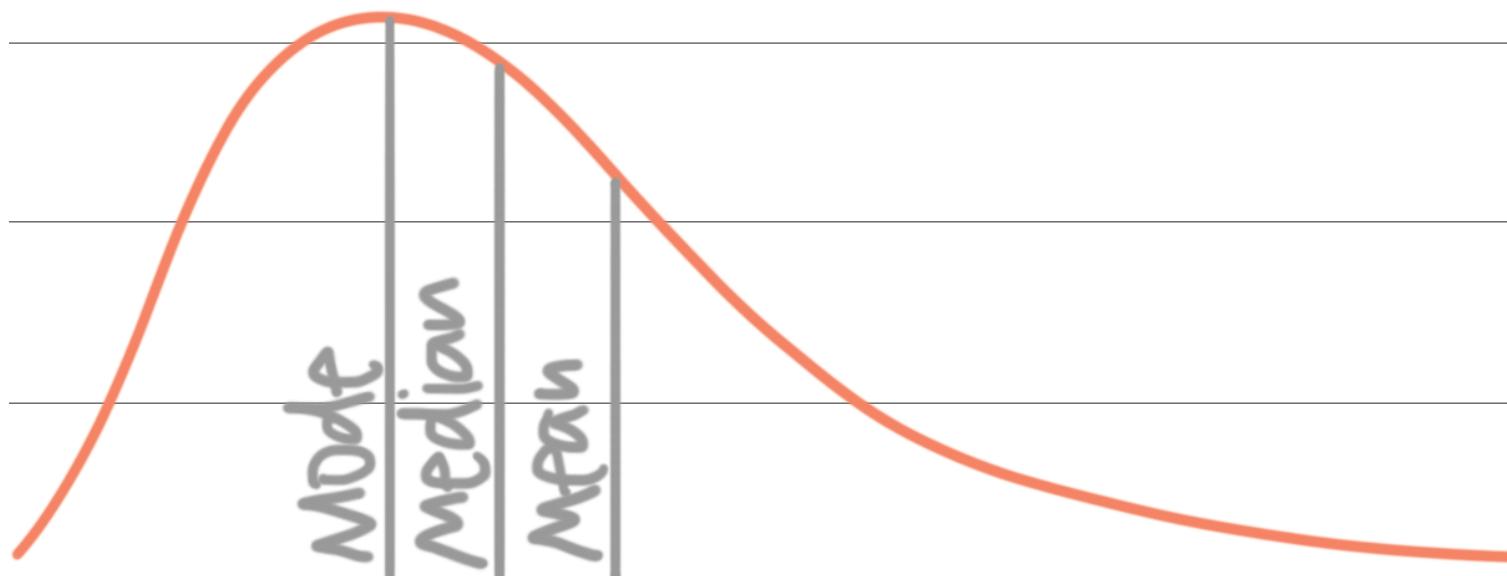
## Skewed distributions

**Skewed distributions** are non-symmetric distributions that lean right or left. We'll look at **negatively skewed distributions** (also called left-skewed distributions or left-tailed distributions), and **positively skewed distributions** (also called right-skewed distributions or right-tailed distributions).

In a left-skewed distribution, the “tail” is on the left. The median of a left-skewed distribution is still at the point that divides the area into two equal parts. The mean is further to the left than the median, more towards the tail on the left side, and the mode is where the data peaks:



In a right-skewed distribution, the “tail” is on the right. The median of a right-skewed distribution is still at the point that divides the area into two equal parts. The mean is further to the right than the median, more towards the tail on the right side, and the mode is still where the data peaks:



## Outliers

The reason we get skewed distributions is because data is disproportionately distributed. Specifically, the majority of the data is clustered in one area, and there are one or more outliers away from the majority of the data. **Outliers** are data points that are unlike most of the rest of the data.

Oftentimes we can't just "eyeball" an outlier. If there's a data point that's really far from most of the data, then we can probably call it an outlier. But there's also a technical way to calculate outliers.

We use what's called the **1.5-IQR rule**, and it will identify both **high outliers** (outliers above the majority of the data) and **low outliers** (outliers below the majority of the data). The rule says that a low outlier is anything less than  $Q_1$  (the first quartile) minus 1.5(IQR), and that a high outlier is anything greater than  $Q_3$  (the third quartile) plus 1.5(IQR).

Low outliers:  $Q_1 - 1.5(\text{IQR})$

High outliers:  $Q_3 + 1.5(\text{IQR})$

For example, if  $Q_1 = 25$ ,  $Q_3 = 35$ , and therefore  $\text{IQR} = 10$ , then the low outliers would be the data points below  $25 - 1.5(10) = 10$  and the high outliers would be the data points above  $35 + 1.5(10) = 50$ .

When we have a data set with outliers that skew the data, the median will be a better measure of central tendency than the mean, and the interquartile range will be a better measure of spread than standard



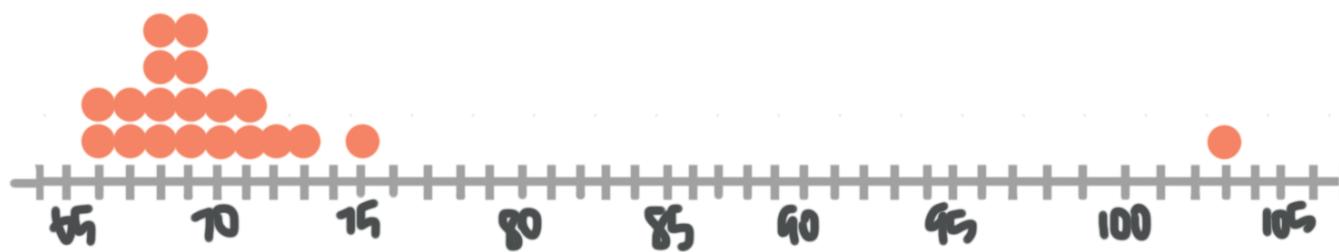
deviation. That's because mean and standard deviation will take into account all points in the data set, including the outliers. But median and IQR can ignore these outliers, giving us more accurate measurements of the data.

So if our data is skewed or if there are outliers, use median for central tendency and IQR for spread. But if our data is fairly symmetrical or there aren't outliers, then consider using mean and standard deviation for central tendency and spread, respectively.

## Describing distributions

When we want to describe the general shape of a distribution, we should mention what we know about its shape, center, spread, and outliers.

Let's take a look at this dot plot of golf scores.



If we imagine drawing a smooth curve over this data, including the point all the way out at 103, we would have a skewed distribution where the long thin tail is on the right side, which means this is a right-skewed, or right-tailed distribution.

The range of the data is  $103 - 66 = 37$ , and there is one outlier: 103. We can tell just by looking at the dot plot that the median is probably close to about 69, but since we have all of the actual data points, we could also

calculate it precisely to see that it is in fact 69. We could also find the IQR to be  $71 - 68 = 3$ .

Remember that, because we have a skewed distribution, the median will be a better measure of center than the mean, and IQR will be a better measure of spread than standard deviation. If we calculate the mean, we find that it's just about 71.

Let's summarize what we found about this distribution of golf scores:

Shape: Positively (right) skewed

Center: Median of 69 (more accurate); Mean of 71 (less accurate)

Spread: IQR of 3

Outliers: 103

# Normal distributions and z-scores

In the last section, we talked about a normal distribution, which is a bell-shaped, symmetric curve for **normally distributed data**, that looks something like this:



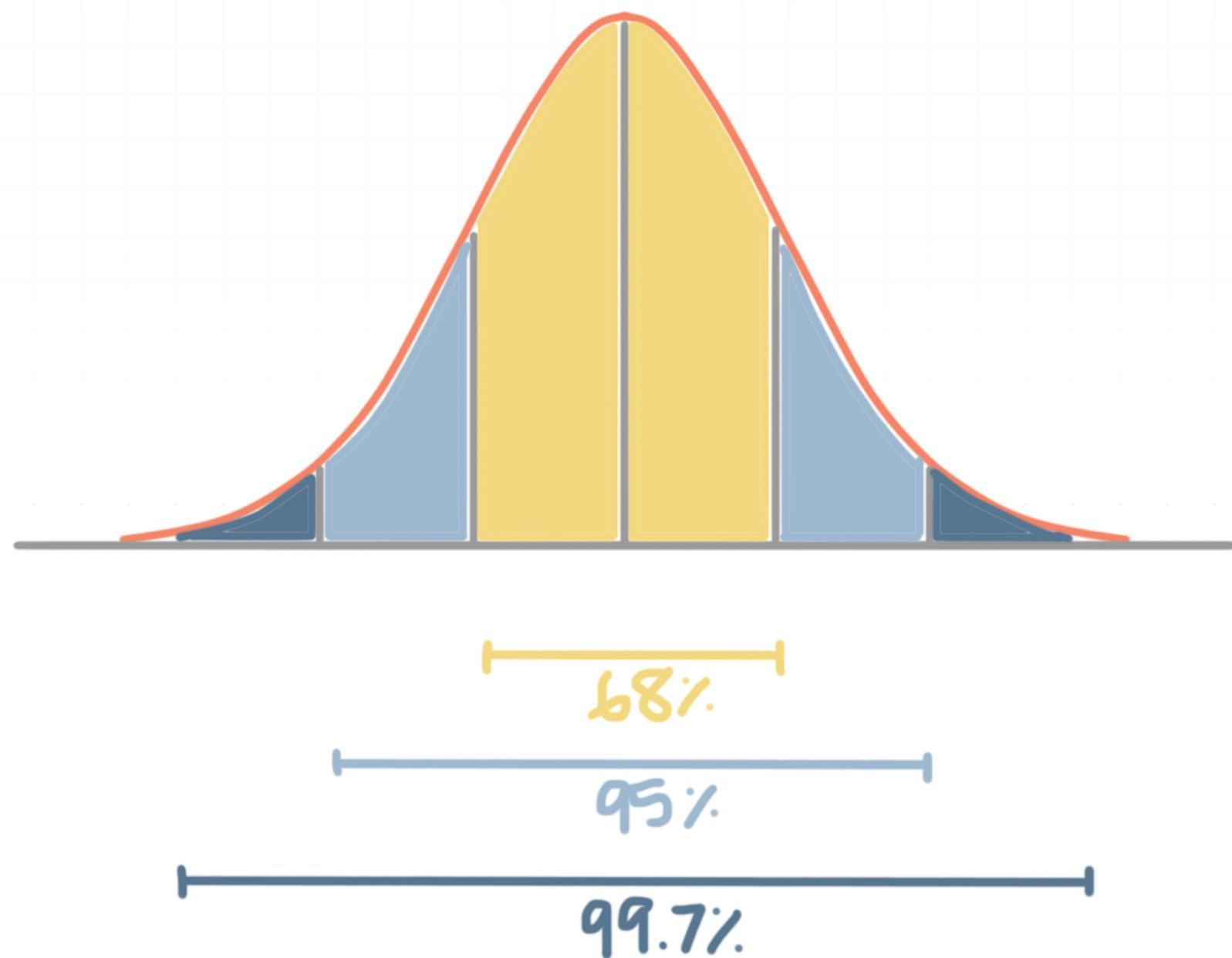
We'll spend a lot of time working with distributions like this, so let's talk about some of the most important properties of a normal distribution.

## The empirical rule

Normal distributions follow the **empirical rule**, also called the **68-95-99.7 rule**. The rule tells us that, for a normal distribution, there's a

- 68% chance a data point falls within 1 standard deviation of the mean
- 95% chance a data point falls within 2 standard deviations of the mean
- 99.7% chance a data point falls within 3 standard deviations of the mean

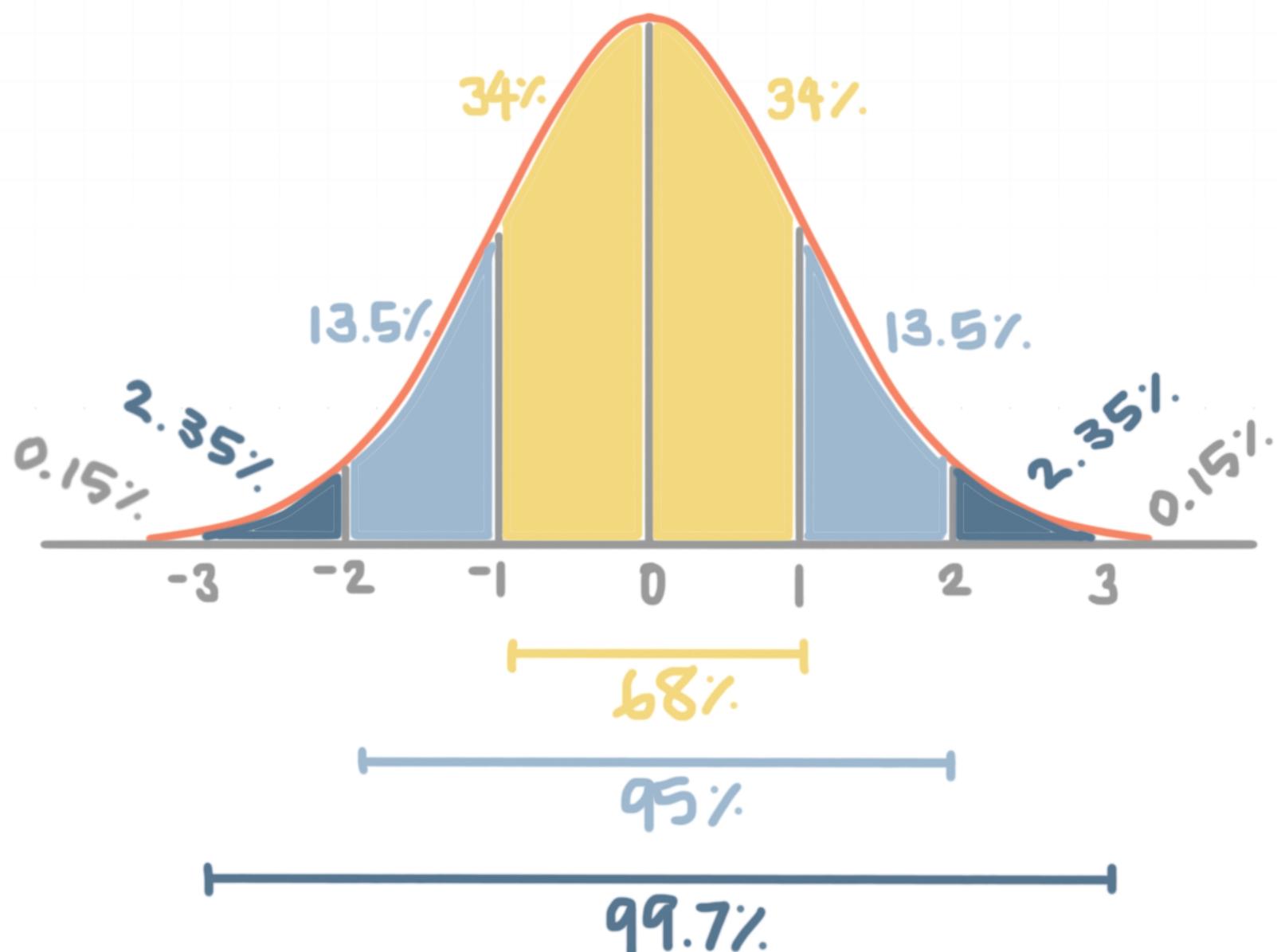
In other words, if we want to show this graphically,



we can show that 68% of the data will fall within 1 standard deviation of the mean, that within 2 full standard deviations of the mean we'll have 95%

of the data, and that within 3 full standard deviations from the mean we'll have 99.7 % of the data.

And we can draw all kinds of conclusions based on this information, and the fact that all the area under the graph represents 100 % of the data. For example, since total area is 100 % , and the data within three standard deviations is 99.7 % , that means that we'll always have 0.3 % of the data in a normal distribution that lies outside three standard deviations from the mean. Or if we wanted to know how much of our data will lie between one and two standard deviations from the mean, we can say that it's  $95\% - 68\% = 27\%$  .



## Percentile

We look a lot at percentiles within a normal distribution. The ***n*th percentile** is the value such that *n* percent of the values lie below it. In other words, a value in the 95th percentile is greater than 95 % of the data. The 50th percentile in a normal distribution always gives the median, or  $Q_2$ , and the IQR is always found using the 75th percentile,  $Q_3$ , minus the 25th percentile,  $Q_1$ .

## Z-scores

A ***z*-score** tells us the number of standard deviations a point is from the mean. To calculate a *z*-score for normally distributed data (normal distributions) we use the formula

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

where  $x$  is the data point,  $\mu$  is the mean, and  $\sigma$  is the standard deviation.

The *z*-score for a data point is how far it is from the mean, and we always want to give the *z*-score in terms of standard deviations. Therefore, to find the *z*-score at a certain point in the distribution, we use the formula above, taking the data point, subtracting the mean, and then dividing that result by the standard deviation. That gives us a value for  $z$ .

We'll look up the *z*-score in a *z*-table, which is a table that takes the number of standard deviations and tells us the percentage of the area under the curve up to that point.



Data points that are less than the mean will be to the left of the mean and will have a negative  $z$ -score. They should be looked up in the table of negative  $z$ -scores:



<b>z</b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>-3.4</b>	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
<b>-3.3</b>	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
<b>-3.2</b>	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
<b>-3.1</b>	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
<b>-3.0</b>	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
<b>-2.9</b>	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
<b>-2.8</b>	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
<b>-2.7</b>	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
<b>-2.6</b>	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
<b>-2.5</b>	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
<b>-2.4</b>	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
<b>-2.3</b>	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
<b>-2.2</b>	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
<b>-2.1</b>	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
<b>-2.0</b>	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
<b>-1.9</b>	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
<b>-1.8</b>	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
<b>-1.7</b>	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
<b>-1.6</b>	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
<b>-1.5</b>	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
<b>-1.4</b>	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
<b>-1.3</b>	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
<b>-1.2</b>	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
<b>-1.1</b>	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
<b>-1.0</b>	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
<b>-0.9</b>	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
<b>-0.8</b>	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
<b>-0.7</b>	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
<b>-0.6</b>	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
<b>-0.5</b>	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
<b>-0.4</b>	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
<b>-0.3</b>	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
<b>-0.2</b>	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
<b>-0.1</b>	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
<b>0.0</b>	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641



Data points that are greater than the mean will be to the right of the mean and will have a positive  $z$ -score. They should be looked up in the table of positive  $z$ -scores:



<b>z</b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>0.0</b>	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
<b>0.1</b>	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
<b>0.2</b>	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
<b>0.3</b>	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
<b>0.4</b>	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
<b>0.5</b>	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
<b>0.6</b>	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
<b>0.7</b>	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
<b>0.8</b>	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
<b>0.9</b>	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
<b>1.0</b>	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
<b>1.1</b>	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
<b>1.2</b>	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
<b>1.3</b>	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
<b>1.4</b>	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
<b>1.5</b>	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
<b>1.6</b>	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
<b>1.7</b>	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
<b>1.8</b>	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
<b>1.9</b>	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
<b>2.0</b>	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
<b>2.1</b>	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
<b>2.2</b>	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
<b>2.3</b>	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
<b>2.4</b>	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
<b>2.5</b>	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
<b>2.6</b>	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
<b>2.7</b>	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
<b>2.8</b>	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
<b>2.9</b>	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
<b>3.0</b>	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
<b>3.1</b>	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
<b>3.2</b>	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
<b>3.3</b>	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
<b>3.4</b>	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



A  $z$ -score is unusual if it's further than three standard deviations from the mean. Essentially the  $z$ -score tells us the percentile rank of the data point that we started with. If the  $z$ -score for our data point is 0.7123, it means that the data point is greater than 71.23 % of the data, meaning that our data point is in the 71.23 percentile. Remember, the  $z$ -table always gives us the percentage of data that's below our data point. Therefore, to find the percentage of data above our data point, we have to take 1 minus the value from the table.

## Thresholds

Sometimes we want to know the **threshold**, or cutoff, in our data set. In other words, we might want to know “What’s the minimum value needed in order to be in the “top 30 %” of the data?

In order to figure this out, we need to work backwards starting from the  $z$ -table. For example, if we want to find the top 30 % of the data, we’d use the  $z$ -table to find the first  $z$ -score that’s just barely above 70 %, or 0.7000. Then we’ll look at the row and column headers that correspond with a  $z$ -table value of 0.7000. The decimal number given by the row and column headers tells us how many standard deviations above the mean we need to be in order to be above 70 %, or, in the top 30 %.

If we multiply that decimal number by the standard deviation, and then add the result to the mean, that will tell us the value that’s at the bottom of the top 30 %. If instead we were looking up the “bottom 40 %” in the  $z$ -table, we’d need to look for the  $z$ -table value that’s just under 0.4000.



## Example

Let's say the mean finishing time for male speed skaters at the winter Olympics on the 500 meter track is 70.42 seconds, with a standard deviation of 0.34 seconds (the data is normally distributed). What is the maximum time a skater can post if he wants to skate faster than 95 % of his competitors?

We know that  $\mu = 70.42$  and  $\sigma = 0.34$ . This athlete wants to be faster than 95 % of the event's participants, which means he wants his time to be in the fastest 5 %. Keep in mind here that, if he's finishing in the fastest 5 %, that means his finishing time is in the lowest 5 % of times.

In other words, he wants his time to be in the top 5 % of finishers, which is equivalent to having a finishing time in the fastest 5 % of all finishing times, which is equivalent to having a value in the bottom 5 % of the normal distribution, so we need to look in a  $z$ -table for the negative  $z$ -score that will keep us under 0.05.

We look for the largest value in the body of the negative  $z$ -table that's still below 0.05. That value is 0.0495, since the next smallest value of 0.0505 surpasses our 0.05 threshold.



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	<b>.0495</b>	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

The  $z$ -score that keeps us within the top 5% of participants is  $-1.65$ . A  $z$ -score of  $-1.64$  would push us into the bottom 95% of competitors, over the 5% threshold.

If our standard deviation is  $\sigma = 0.34$ , and our  $z$ -score is  $-1.65$ , then we can calculate the maximum skater's time to be in top 5%.

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

$$-1.65 = \frac{x - 70.42}{0.34}$$

$$-1.65(0.34) = x - 70.42$$

$$x = 70.42 - 1.65(0.34)$$

$$x = 69.859 \text{ seconds}$$

Therefore, if the skater's time is 69.859 seconds or faster, he'll be within the fastest 5% of finishers in the event.

# Chebyshev's Theorem

The empirical rule tells us that the percentage of data that falls within 1, 2, and 3 standard deviations of the mean in a normal distribution is 68 %, 95 %, and 99.7 %, respectively.

Of course, the constraint of the empirical rule is that it only applies to normally distributed data. We can't use it when our distribution is left-skewed or right-skewed. But that's where Chebyshev's Theorem comes in.

**Chebyshev's Theorem** tells us that at least  $(1 - 1/k^2)\%$  of our data must be within  $k$  standard deviations of the mean, for  $k > 1$ , and regardless of the shape of the data's distribution. For instance, here's what the theorem can conclude for any distribution when  $k = 2, 3$ , and  $4$ :

- At least 75 % of the data must be within  $k = 2$  standard deviations of the mean.
- At least 89 % of the data must be within  $k = 3$  standard deviations of the mean.
- At least 94 % of the data must be within  $k = 4$  standard deviations of the mean.

Keep in mind that  $k$  doesn't have to be an integer, but it *does* have to be greater than 1. So we could use Chebyshev's Theorem for  $k = 1.32$ ,  $k = 2$ , or  $k = 2.14$ , but not for  $k = 1$  or for  $k = 0.46$ .

Notice how these percentages are less than the corresponding percentages for the normal distribution. For instance, in a normal



distribution, the Empirical Rule tells us that 95 % of the data falls within 2 standard deviations of the mean, but Chebyshev's Theorem only lets us conclude that 75 % of the data will fall within 2 standard deviations of the mean. Similarly, the Empirical Rule tells us that 99.7 % of the data falls within 3 standard deviations of the mean, but Chebyshev's Theorem only lets us conclude that 89 % of the data will fall within 3 standard deviations of the mean.

Because Chebyshev's Theorem has to work for distributions of all shapes, unlike the Empirical Rule which applies only to the normal distribution, Chebyshev's Theorem is required to be more conservative. And that's why we see smaller percentages for Chebyshev's Theorem than we do for the Empirical Rule.

Let's do an example so we can see how to apply Chebyshev's Theorem.

### Example

A statistics class of 40 students has a mean final exam score of 86, with a standard deviation of 3. How many students scored between 81 and 91 on the final exam?

We need to determine the distance from the mean of 81 and 91, in terms of standard deviations.

$$k = \frac{81 - 86}{3} = -\frac{5}{3} \approx -1.67$$



$$k = \frac{91 - 86}{3} = \frac{5}{3} \approx 1.67$$

Then we can apply Chebyshev's Theorem, using the value of  $k$  that's greater than 1, which is  $k = 1.67$ .

$$1 - \frac{1}{k^2} = 1 - \frac{1}{1.67^2} \approx 0.64$$

Because we found 0.64, we know at least 64% of the students scored between 81 and 91 on the final exam. Because there are 40 students in the class, 64% of the class is  $0.64(40) = 25.6$ .

So Chebyshev's Theorem tells us that at least 25.6 students scored between 81 and 91 on the exam. It doesn't make sense to say 25.6 students, but "at least 25.6 students" doesn't meet the threshold of "at least 26 students," so we round down to get our conclusion:

*"According to Chebyshev's Theorem, at least 25 students scored between 81 and 91 on the final exam."*

We can also use Chebyshev's Theorem to work backwards through probability problems.

### Example

A statistics class of 40 students has a mean final exam score of 86, with a standard deviation of 3. Find the score range for the central 80% of test scores.



Using Chebyshev's Theorem,

$$0.8 = 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} = 1 - 0.8$$

$$1 = 0.2k^2$$

$$k^2 = 5$$

$$k \approx 2.24$$

Approximately 2.24 standard deviations above the mean gives us a test score of

$$86 + 2.24(3)$$

$$86 + 6.72$$

$$92.72$$

And 2.24 standard deviations below the mean gives us a test score of

$$86 - 2.24(3)$$

$$86 - 6.72$$

$$79.28$$

So at least 80 % of the final exam scores fell between 79.28 and 92.72. If test scores can only be integers, then at least 80 % of the final exam scores must have fallen between 80 and 92.

---



# Simple probability

Up to now we've been talking about statistics, which is all about data, and how to display, summarize, and analyze data. Now we'll transition into probability, which is all about the likelihood of whether or not some event will occur.

The reason we study statistics and probability together is because when we collect data as part of a statistical study, we want to be able to use what we know about probability to say how likely it is that our results are reliable. So in that way, statistics and probability go hand-in-hand.

## Probability

So what is probability in its most basic form? **Probability** is how likely it is that something will occur. We can write a probability as a fraction, decimal or percent, but all probabilities are numbers equal to or between 0 and 1.

A probability of 0 means there's a 0 % chance that something will occur (it's impossible), whereas a probability of 1 means there's a 100 % chance that something will occur (it's guaranteed). Probabilities between 0.5 and 1 mean something is more likely to occur than not, whereas probabilities between 0 and 0.5 mean something is more likely not to occur. A probability of 0.5 means there's an equal chance of the event occurring vs. not occurring.



Typically, we talk about something like the probability that we'll get heads when we flip a coin, or the probability that we get a queen when we pull a card from a deck of playing cards.

But what we're basically asking is, "How likely is it that we'll get heads when we flip a coin?" or "How likely is it that we'll get a queen when we pull a card from a deck?" How likely these things are depends on the full set of all possible outcomes, and how many of those possible outcomes meet our particular criteria.

We usually denote the probability of an event as  $P(\text{event})$ . So the probability that we'll get heads when we flip a coin is  $P(H)$ , and the probability that we get a queen we draw from a deck of cards might be  $P(Q)$ . Let's find these two probabilities.

The formula for simple probability is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

The collection of "all possible outcomes" from the denominator is called the **sample space**. For example, when we flip a coin, the sample space  $S$  is  $S = \{H, T\}$ , because there are two possible outcomes: heads or tails.

Now think about flipping a coin. If we flip a coin one time, there are two possible outcomes: we'll either get heads or tails. We want to know the probability of getting heads, and there's only one outcome that satisfies this: heads. Therefore, the probability of getting heads is

$$P(H) = \frac{1}{2}$$



In other words, there's a 50% chance that we'll get heads when we flip a coin one time.

What about the probability of drawing a queen from a deck of playing cards. Well, there are 52 cards in a deck, which means there are 52 total possible outcomes. There are only 4 queens in the deck though, which means there are only 4 outcomes that meet our criteria. So the probability of drawing a queen is

$$P(Q) = \frac{4}{52} = \frac{1}{13}$$

Keep in mind that in order to use this simple probability formula, all of the possible outcomes need to be **equally likely** to occur. In other words, it needs to be equally likely that we'll get heads or tails when we flip the coin. And it needs to be equally likely that we'll pull any of the 52 cards in the deck. Which means that, when we're dividing by "all possible outcomes" in the simple probability formula, we're actually dividing by "all possible **equally likely outcomes**."

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible equally likely outcomes}}$$

## Experimental and theoretical probability

Flipping a coin one time is one **experiment**. Drawing a card from a deck is one experiment. So when we talk about the probability of getting heads or tails when we flip a coin, we could do several experiments in a row.



For example, let's say we flip a coin four times in a row. As we know, it's completely possible that, just by chance, we end up with four heads in a row. Based on that result, we might say that the probability of getting heads is

$$P(H) = \frac{4}{4} = 1 = 100\%$$

But how can this be true? Before we saw that the probability of getting heads on one flip was 50%, but now we're calculating the probability of getting heads four times in a row at 100%. What's going on?

We're looking at the difference between experimental and theoretical probability. **Experimental probability** (also called empirical probability) is the probability we find when we run experiments. Flipping four heads in a row tells us that we've found the experimental probability of getting heads as 100%. But if we flip the coin a fifth time and get tails this time, then the experimental probability of getting heads after 5 experiments is

$$P(H) = \frac{4}{5} = 0.8 = 80\%$$

In other words, the experimental probability of an event will be constantly changing as we run more and more experiments over time. If the experiment is a good one, the idea is that over time the experimental probability will get very close to the theoretical probability.

**Theoretical probability** (also called classical probability) is the probability that an event will occur if we could run an infinite number of experiments. Or, we can think about the theoretical probability as the one we get from the simple probability formula:



$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

We know from using this formula that the probability of getting heads when we flip a coin is 50 %. Therefore, the theoretical probability is 50 %, which means that the more experiments we run, the closer our experimental probability should get to 50 %.

This is also called the **law of large numbers**. It says that, if we could run an infinite number of experiments, that our experimental probability would eventually equal our theoretical probability.

### Example

Find the theoretical probability of rolling an even number when we roll a 6-sided die one time.

Since we've been asked for theoretical probability, we'll use the formula

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

We know that if we roll the 6-sided die, that there are 6 possible, equally likely outcomes: 1, 2, 3, 4, 5, 6. Of those possible outcomes, 2, 4, and 6 are even numbers, so there are 3 outcomes that meet our criteria. Therefore, the probability of rolling an even number is

$$P(\text{even}) = \frac{3}{6} = \frac{1}{2} = 50\%$$





# The addition rule, and union vs. intersection

Sometimes we'll need to find the probability that two events occur together within one experiment. Remember that an **event** is a specific collection of outcomes from the **sample space**. For example, what's the probability that we roll a pair of 6-sided dice and either get at least one 1, or an even sum when we add the dice together?

When we roll two dice together, there are 36 possible outcomes. There are 11 rolls out of the 36 where we get at least one 1:

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

And there are 18 possible outcomes where the sum of the dice is even.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

So we might be tempted to say that the probability of getting at least one 1 or an even sum is  $P(1 \text{ or even}) = (11 + 18)/36$ , or  $29/36$ . But we've neglected to consider that there's some overlap between these two sets. We have the rolls 1 – 1, 1 – 3, 1 – 5, 3 – 1, and 5 – 1 in both sets, so we're double-counting those in our probability calculation.

Which means we have to subtract out the values that are overlapping. Since there are 5 overlapping values, the probability calculation is actually

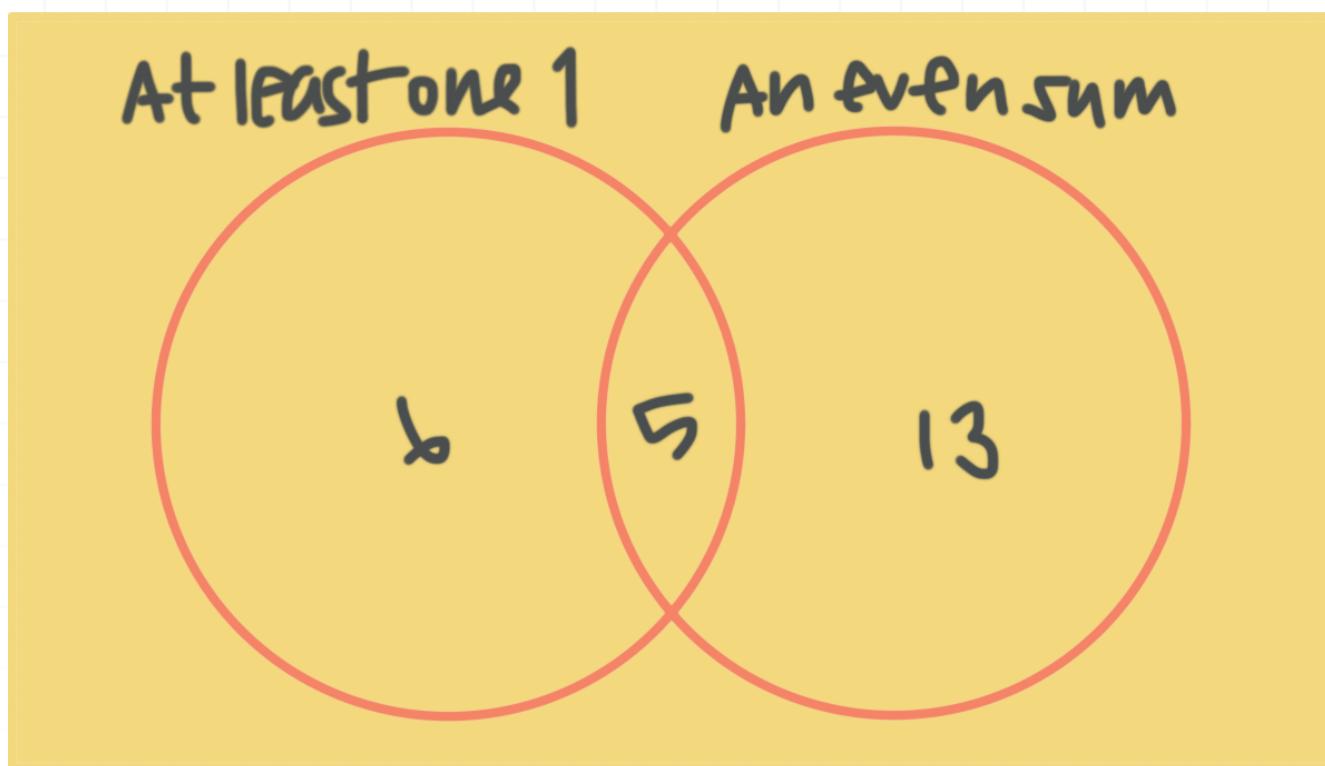
$$P(1 \text{ or even}) = \frac{11 + 18 - 5}{36}$$

$$P(1 \text{ or even}) = \frac{24}{36}$$

$$P(1 \text{ or even}) = \frac{2}{3}$$

A great way to illustrate this kind of overlapping probability is with a Venn diagram. We would build a Venn diagram to show that there are 11 rolls where we get at least one 1, that there are 18 rolls where the sum is even,

and that there are 5 rolls where we get at least one 1 and the sum is also even.



Then, from the Venn diagram, we just add the  $6 + 5 = 11$  and the  $5 + 13 = 18$ , and then subtract the overlapping 5, in order to get all of the outcomes that meet our criteria, but without double-counting any of the outcomes. And then our probability again is

$$P(1 \text{ or even}) = \frac{11 + 18 - 5}{36} = \frac{24}{36} = \frac{2}{3}$$

## Addition rule

This idea of making sure that we don't double-count the overlap is called the **addition rule** (or sum rule) for probability, and it's given as:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Now what happens if there's no overlap between  $A$  and  $B$ ? In that case,  $A$  and  $B$  are called **mutually exclusive** (or disjoint), and  $P(A \text{ and } B)$  will be 0. Which means the addition rule will simplify this way:

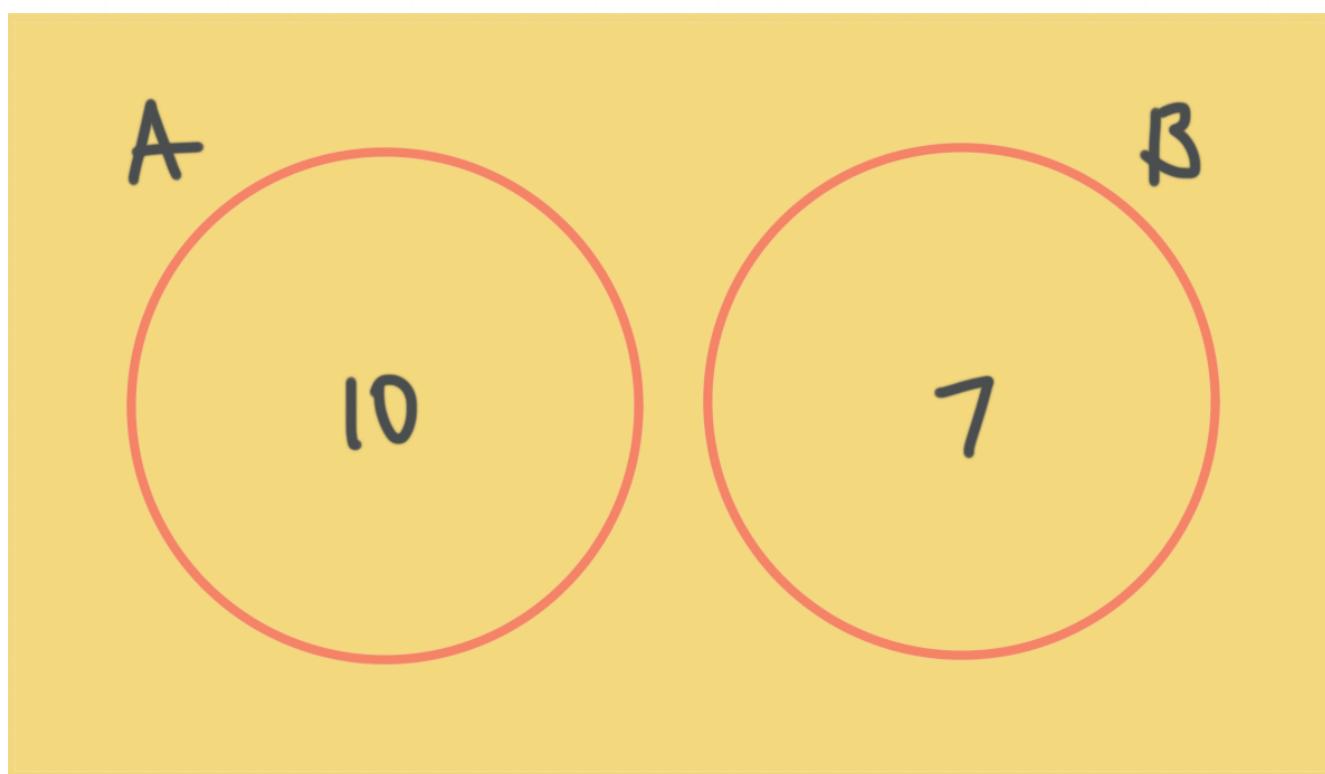
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ or } B) = P(A) + P(B) - 0$$

$$P(A \text{ or } B) = P(A) + P(B)$$

Which tells us that when events are mutually exclusive/disjoint, we can calculate the probability of either event  $A$  happening or event  $B$  happening simply by adding together the probability of each one happening individually.

For instance, the events in this Venn diagram are disjoint, since the circles don't overlap:



Because there are  $10 + 7 = 17$  total events, the probability of event  $A$  is  $P(A) = 10/17$ . And the probability of event  $B$  is  $P(B) = 7/17$ . So the probability that both events occur is

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ or } B) = \frac{10}{17} + \frac{7}{17}$$

$$P(A \text{ or } B) = \frac{17}{17}$$

$$P(A \text{ or } B) = 1$$

## Union and intersection

In the first version of the addition rule formula, we use the words “or” and “and.” But we can also write the formula as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This second formula is the same addition rule calculation, but we use the  $\cup$  and  $\cap$  symbols instead of the words “or” and “and,” respectively.

$P(A \cup B)$  is called the **union** of  $A$  and  $B$ , and it means the probability of either  $A$  or  $B$  or both occurring.  $P(A \cap B)$  is called the **intersection** of  $A$  and  $B$ , and it means the probability of both  $A$  and  $B$  both occurring.

---

## Example



We surveyed 100 people about their favorite sport, and recorded their gender and favorite sport in a table.

Sport	Male	Female	Total
Football	22	16	38
Basketball	13	8	21
Other	25	16	41
<b>Total</b>	<b>60</b>	<b>40</b>	<b>100</b>

1. What is the probability that a participant is male?
2. What is the probability that a participant's favorite sport is football?
3. What is the probability that a participant is female or prefers a sport other than football or basketball?

We know from the table that 60 of the 100 participants are male, so the probability that a participant is male is

$$P(\text{male}) = \frac{60}{100} = \frac{3}{5}$$

And from the table we can see that 38 of the 100 participants like football best, so the probability that a participant's favorite sport is football is

$$P(\text{football}) = \frac{38}{100} = \frac{19}{50}$$



These were both simple probability questions, but the third question requires us to use the addition rule. There are 40 female participants, and 41 participants who prefer a sport other than football or basketball.

But there are 16 participants in the “overlap” group: the group of females who also prefer a sport other than football or basketball. Therefore, we’ll apply the addition rule and say that the probability that a participant is female or likes a sport other than football and basketball is

$$P(\text{female or other}) = \frac{40 + 41 - 16}{100} = \frac{65}{100} = \frac{13}{20}$$

---



# Independent and dependent events and conditional probability

## Independent probability

Up to this point, we've been focusing on **independent events**, which are events that don't affect one another. For example, if I flip a coin two times in a row, the result of the first flip doesn't affect the second flip, so those flips are independent events.

In other words, if I get heads on the first flip, the second flip still has an equally likely chance of producing heads or tails. If instead I get tails on the first flip, the second flip still has an equally likely chance of producing heads or tails.

## The multiplication rule

When we want to find the probability of multiple independent events (also called a **joint occurrence**), we'll multiply their probabilities. This is called the **multiplication rule**. So for example, the probability that we get heads twice when we flip a coin two times in a row is

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

$$P(HH) = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \frac{1}{4}$$

In other words, the probability of getting heads on the first flip is  $1/2$ . The probability of getting heads on the second flip is  $1/2$ . These are independent events, so the result of one flip doesn't affect the other flip.



Therefore, the probability of getting heads twice in a row is just a product of the individual probabilities.

We could have also calculated the probability of getting heads twice in a row by realizing that there are four possible outcomes when we flip a coin twice:  $HH$ ,  $HT$ ,  $TH$ , and  $TT$ . The outcome we're interested in is  $HH$ , which means that the probability of getting heads twice in a row is

$$P(\text{event}) = \frac{\text{outcomes that meet our criteria}}{\text{all possible outcomes}}$$

$$P(HH) = \frac{1}{4}$$

When we flip a coin, there's a 50% chance of getting heads, and a 50% chance of getting tails. But that's not what makes the events independent. It's the fact that one flip doesn't affect any other flip. But let's say that a soccer player makes 70% of all penalty kicks that he attempts. In other words, for every 100 penalty kicks he attempts, he's going to make about 70 of them.

Given his success rate, we can calculate the likelihood that he makes 5 penalty kicks in a row. Each penalty kick is an independent event, since the result of the first kick theoretically has no effect on the second kick, and so on. Therefore, the probability that he makes 5 penalty kicks in a row is

$$P(KKKKK) = (0.7)(0.7)(0.7)(0.7)(0.7)$$

$$P(KKKKK) = (0.7)^5$$

$$P(KKKKK) = 0.16807 \approx 0.17$$

Even though the probability that he makes any single penalty kick is 70 %, and each kick is an independent event, the probability that he makes 5 in a row is about 17 %.

### Example

We want to find the probability of drawing a jack from a deck of 52 cards, and then drawing another jack, assuming we put the first jack back in the deck before drawing the second.

Because we're replacing the first jack before we draw the second, these events are independent. The second draw won't be affected by the outcome of the first draw.

Because there are four jacks in a regular deck, the probability that we get a jack on the first draw is

$$P(J_1) = \frac{4}{52} = \frac{1}{13}$$

After the first draw, we replace the jack, which completely resets the deck. So the probability that we get a jack on the second draw is

$$P(J_2) = \frac{4}{52} = \frac{1}{13}$$

Therefore, the probability that we get two jacks in a row when we draw cards with replacement is



$$P(J_1) \cdot P(J_2) = \left(\frac{1}{13}\right) \left(\frac{1}{13}\right) = \frac{1}{169}$$


---

## Dependent probability

Contrast this with **dependent events**, which are events that have an effect on one another. Pulling cards from a deck without replacing the ones we've pulled would be an example of dependent events.

If we pull one card from a 52-card deck, the probability of getting that exact card is  $1/52$ . If we set that card to the side without replacing it in the deck, and then pull another card, the probability of getting a specific card is no longer  $1/52$ , it's  $1/51$ . Since the probability changed, these are dependent events.

When we talk about dependent events like drawing two cards from a deck without replacing the first, we express that probability as  $P(A | B)$  (read as “the probability of  $A$  given  $B$ ”), which gives the probability of event  $A$  happening given that event  $B$  has already happened. Dependent probability is also called **conditional probability**.

When events are dependent, we have to think about the probability of each event.

### Example



If we draw a card from a deck of playing cards, and then without replacing it, draw a second card, what is the probability that we'll get two jacks in a row?

Since we're dealing with dependent probability, we need to look at the probability of each event. The probability that we get a jack on the first draw is

$$P(J_1) = \frac{4}{52} = \frac{1}{13}$$

The probability that we get a jack on the second draw is

$$P(J_2 | J_1) = \frac{3}{51} = \frac{1}{17}$$

Therefore, the probability that we get two jacks in a row is

$$P(J_1) \cdot P(J_2 | J_1) = \left(\frac{1}{13}\right) \left(\frac{1}{17}\right) = \frac{1}{221}$$

If we put this result with the jacks into a formula, we can say that

$$P(A \text{ and } B) = P(A) \cdot P(B | A)$$

The vertical bar in  $P(B | A)$  means “given”, so  $P(B | A)$  is the probability that  $B$  occurs given that  $A$  has already occurred.

This formula tells us that the probability that  $A$  and  $B$  both happen (that we pull a jack on the first draw and a jack on the second draw) is the product of the probability of the first event happening and the probability that the second event happens, given that the first event already happened.

From this formula, we can prove that events  $A$  and  $B$  are independent if we can show that  $P(A | B) = P(A)$ , because this means that the probability of event  $A$  happening is the same as the probability of event  $A$  happening, even if event  $B$  already happened, which means event  $B$  must have had no effect on event  $A$ , and therefore that  $A$  and  $B$  are independent. The events are also independent if  $P(B | A) = P(B)$ . If this is not true of two events, then they're not independent events and we call them **dependent events**.

# Bayes' Theorem

**Bayes' Theorem**, also known as Bayes' Law or Bayes' Rule, tells us the probability of an event, given prior knowledge of related events that occurred earlier. Bayes' Theorem is

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

To simplify Bayes' Theorem problems, it can be really helpful to create a tree diagram. If we're ever having trouble figuring out a conditional probability problem, a tree diagram is a great tool to fall back on, because it shows all of the sample space of the problem.

## Example

We have two dice. One is fair, and the other one is weighted to land on 6, 50% of the time. There's an equal probability for the other five faces on the biased die. Without knowing which one we're choosing, we pick one of the dice, roll it, and get a 6. What is the probability that we rolled the biased die?

Since we're looking for the probability that a die is biased given that we already rolled a 6, we can say that we're looking for  $P(\text{Biased} | 6)$ . In this case,  $P(A) = P(\text{Biased})$  and  $P(B) = P(6)$ . So we need to find the following values and put them into a formula:

- $P(6 | \text{Biased})$



- $P(\text{Biased})$
- $P(6)$

We know from the problem that  $P(6 | \text{Biased}) = 1/2$ . And since there are two die and each choice is equally likely  $P(\text{Biased}) = 1/2$ .

The probability of rolling a 6 is the probability of choosing the biased die and rolling a 6 or the probability of choosing the fair die and rolling a 6.

Let's first find the probability that the die is biased and we roll a 6.

$P(\text{Biased}) = 1/2$  and the probability of a 6 on the biased die is 50%. So the probability of biased and 6 is

$$P(\text{Biased and } 6) = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \frac{1}{4}$$

Now let's find the probability that the die is fair and we roll a 6.

$$P(\text{Fair and } 6) = \left(\frac{1}{2}\right) \left(\frac{1}{6}\right) = \frac{1}{12}$$

Therefore, the probability of rolling a 6 is

$$P(6) = \frac{1}{4} + \frac{1}{12}$$

$$P(6) = \frac{3}{12} + \frac{1}{12}$$

$$P(6) = \frac{4}{12}$$



$$P(6) = \frac{1}{3}$$

Putting these values into Bayes' Theorem, we get

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

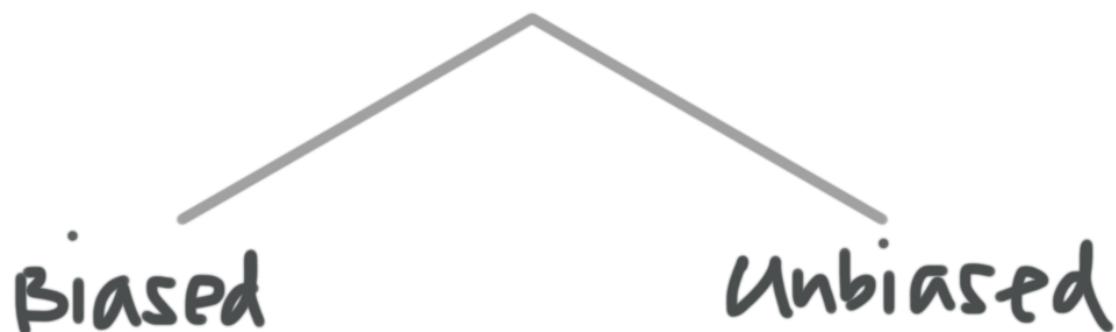
$$P(A | B) = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{3}}$$

$$P(A | B) = \frac{\frac{1}{4}}{\frac{1}{3}}$$

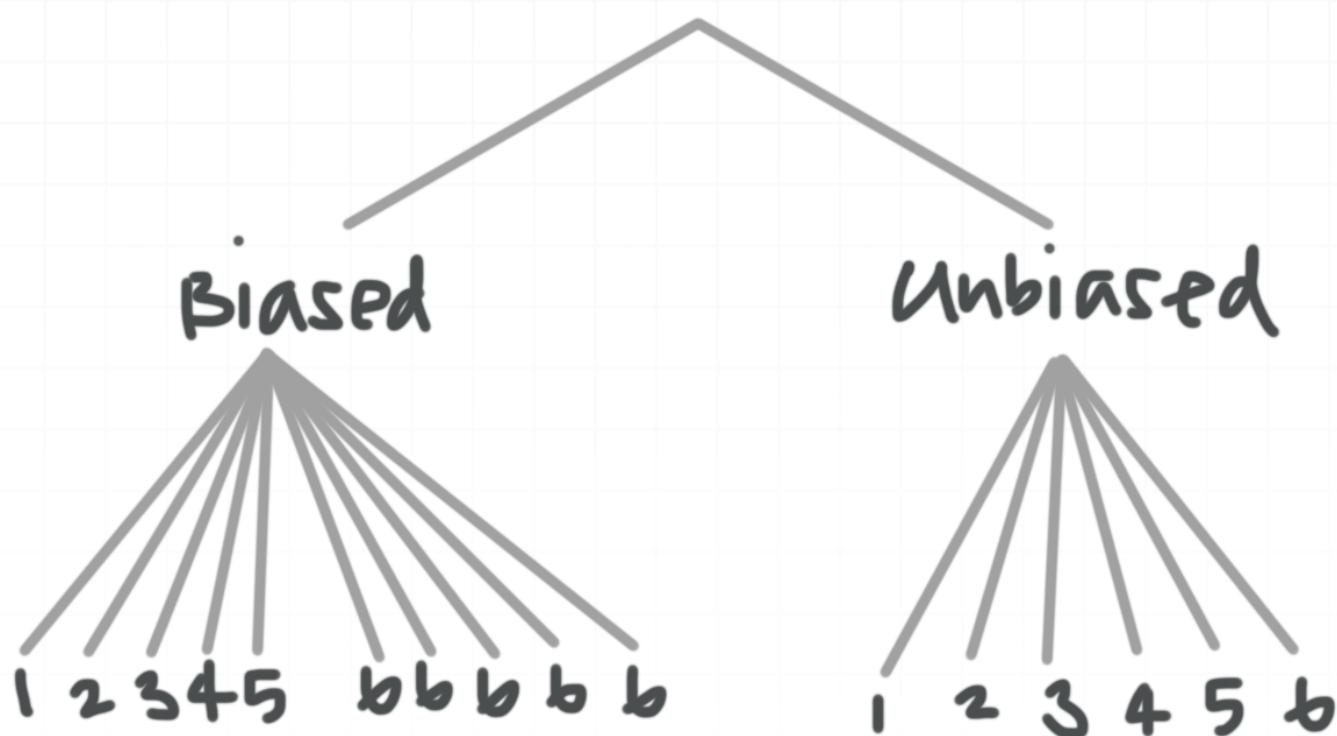
$$P(A | B) = \frac{1}{4} \cdot \frac{3}{1}$$

$$P(A | B) = \frac{3}{4}$$

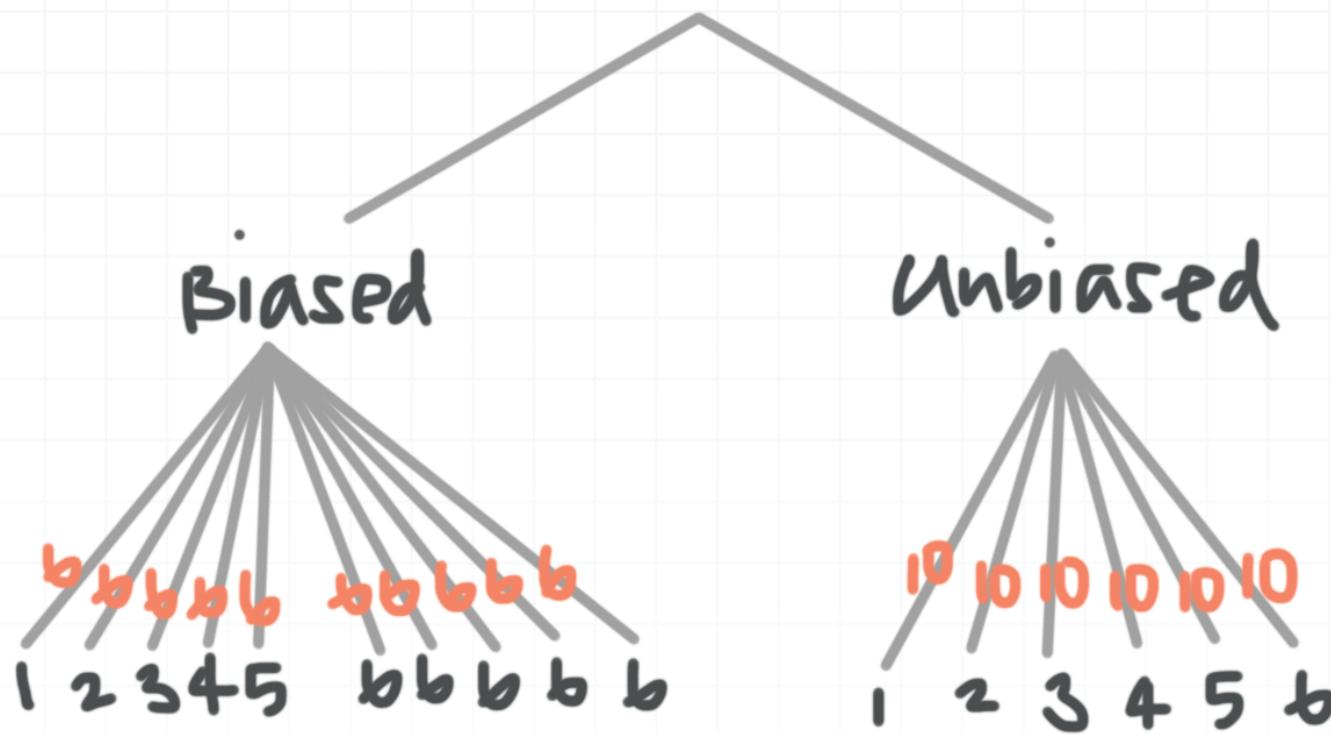
Let's look at how we would do this with a tree diagram. We need to show branches in our tree for every event in our problem. First, we picked a die. It can either be the biased die or the unbiased die, and we have an equal chance of picking each one.



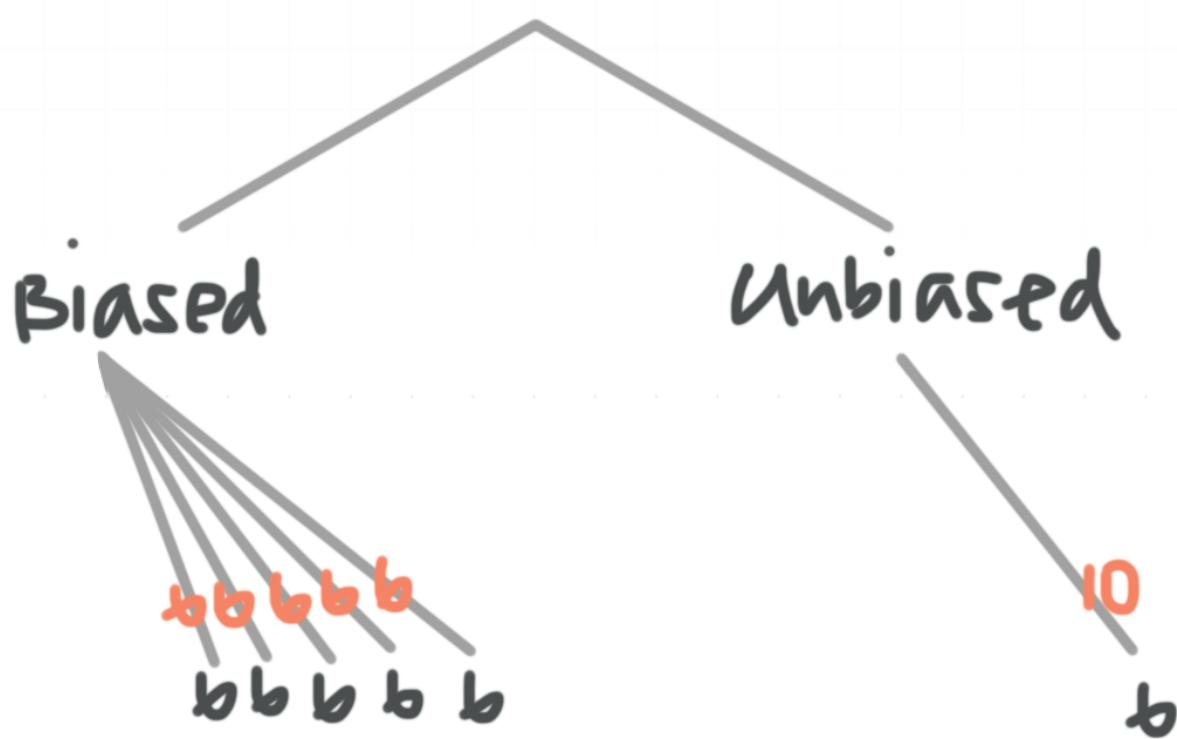
Each die can lead to six possible outcomes, so we'll show these in our tree. We also know that the biased die will land on 6, 50% of the time, and that all five other outcomes are equally likely, which means we could think of this as 5 possible outcomes leading to 6, and one possible outcome leading to each of the other numbers.



But we always need to make sure our tree is balanced. We can't have 10 branches coming down from the biased side, and 6 branches coming down from the unbiased side. We have to have an equal number of branches on both sides, so we'll scale up both sides to the least common multiple, which is 60. That means we need to scale up all the biased branches by 6 and all the unbiased branches by 10.



The next part of the problem tells us that we rolled one die and got a 6. So we'll trim all the branches of our tree that lead to results that didn't occur.



Now we want to know the probability that we rolled the biased die. Remember that each of the branches on the biased side represents 6 possibilities, so there are 30 branches coming from the biased side. The branch on the unbiased side represents 10 possibilities, so there are 10 branches coming from the unbiased side. Therefore, the probability that

we rolled the biased die is all the outcomes coming from the biased side, 30, divided by all of the branches in total, 40:

$$P(\text{Biased}) = \frac{30}{30 + 10} = \frac{30}{40} = 75\%$$

---



# Discrete probability

## Discrete random variables and probability distributions

A **discrete random variable** is a variable that can only take on discrete values. For example, if we flip a coin twice, we can only get heads zero times, one time, or two times. We can't get heads 1.5 times, or 0.31 times. The number of heads we can get takes on a discrete set of values: 0, 1, and 2. A **continuous random variable**, on the other hand, can take on any value in a certain interval.

In probability distributions for all random variables, the probabilities of each of the possibilities has to sum to 1, or 100%.

For example, if I flip a coin twice, I can get any of the following outcomes:

*HH*

*HT*

*TH*

*TT*

There are four possible outcomes, and one of them where I get 0 heads, so the probability of getting 0 heads is  $1/4$ . In *HT* and *TH* I get 1 heads, so the probability of getting 1 heads is  $2/4$ . In *HH* I get 2 heads, so the probability of getting 2 heads is  $1/4$ .

Now we can tell that this is a valid discrete probability distribution, because



$$\frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1 = 100\%$$

The fact that a valid probability distribution always sums to 100% allows us to find missing values in our data. For example, if instead we'd been told that the table below tells us the probability of getting a certain number of heads when we flip a coin twice,

Heads	Probability
0	0.25
1	0.50
2	

we could calculate the missing value by subtracting the known probabilities from 1.00. So we could say that the probability of getting exactly 2 heads is

$$P(2 \text{ heads}) = 1.00 - 0.25 - 0.50 = 0.25$$

And then we could complete the table.

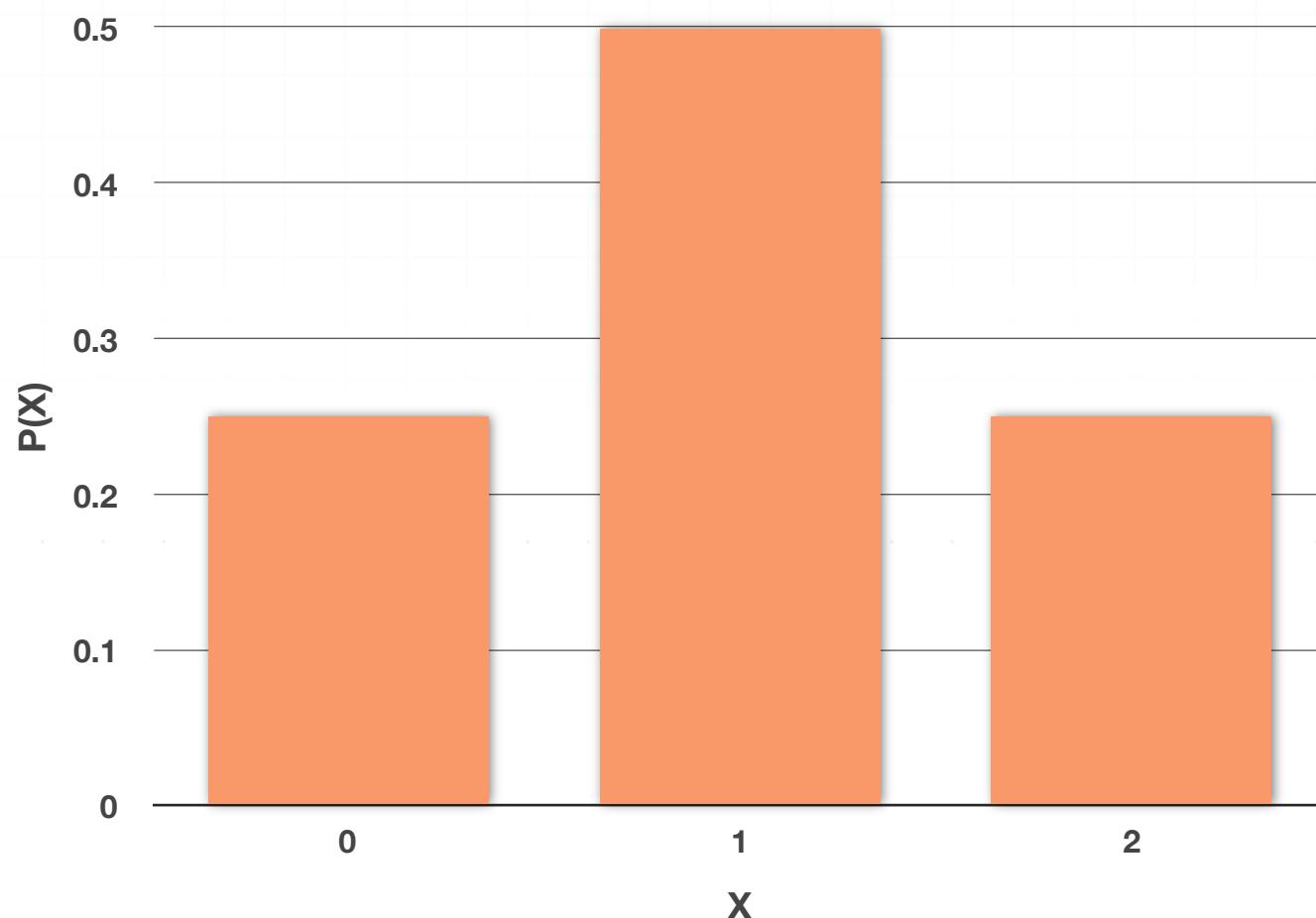
Heads	Probability
0	0.25
1	0.50
2	0.25

Keep in mind that we often use capital  $X$  to represent a discrete random variable. Which means that, for the example of flipping the coin twice, we could call the number of heads  $X$ , and the probability of getting a certain

number of heads  $P(X)$ . And we could give the probability distribution table as

X	P(X)
0	0.25
1	0.50
2	0.25

Or we could take the same information and graph the distribution this way:



## Expected value

Once we have a probability distribution for a discrete random variable,  $X$ , we can calculate the **expected value**  $E(X)$ , which is the mean of  $X$ . The

expected value is often referred to as the “long term average.” When we run an experiment over and over and over again, this is the mean we’d expect to find. To find this value for a discrete random variable, we have to “weight” each value.

For example, if we want to find the expected value for the number of heads when we flip a coin two times, we’ll multiply each value of  $X$  by the corresponding value of  $P(X)$ , and then add them all together.

X	P(X)
0	0.25
1	0.50
2	0.25

So the expected value is

$$E(X) = \mu_X = 0(0.25) + 1(0.50) + 2(0.25)$$

$$\mu_X = 0 + 0.50 + 0.50$$

$$\mu_X = 1$$

Therefore, on average, we’ll expect to get 1 heads when we flip a coin two times. We can then extrapolate this to guess that, for example, we should get 50 heads when we flip a coin 100 times.

## Variance and standard deviation

We can also find the variance and standard deviation for discrete random variables. To find the variance, we'll take the difference between  $X$  and the mean,  $\mu_X$ , square that difference, and then multiply the result by the probability of  $X$ , called  $P(X)$ . We'll do that for each value of  $X$ , and then add all those results together to get the **variance**,  $\sigma_X^2$ .

$$\sigma_X^2 = \sum_{i=1}^n (X_i - \mu)^2 P(X_i)$$

Let's find the variance when  $X$  is the number of heads we get when we flip a coin two times, remembering that we already found  $E(X) = 1$  for this probability distribution.

$$\sigma_X^2 = (0 - 1)^2(0.25) + (1 - 1)^2(0.50) + (2 - 1)^2(0.25)$$

$$\sigma_X^2 = (-1)^2(0.25) + (0)^2(0.50) + (1)^2(0.25)$$

$$\sigma_X^2 = 1(0.25) + 0(0.50) + 1(0.25)$$

$$\sigma_X^2 = 0.25 + 0.25$$

$$\sigma_X^2 = 0.50$$

We can also find the standard deviation of  $X$ ,  $\sigma_X$ , which is just the square root of the variance.

$$\sigma_X = \sqrt{0.50}$$

$$\sigma_X \approx 0.71$$

Let's do an example where we find mean, variance, and standard deviation of a discrete random variable.



## Example

We're playing a game of chance in which a computer randomly chooses four numbers from 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9, with replacement. We pay \$3 to play the game. If we pick the same four numbers as the computer, we win \$10,000 and get our \$3 back, so we profit \$10,000. If we fail to match all four of the computer's numbers, we lose our \$3, and our profit is –\$3. What is our expected profit in the long run if we play this game over and over again?

Let  $X$  be the profit on each play.  $X$  is a discrete random variable whose possible values are only –3 and 10,000.

Since the computer picks from 10 numbers, the probability that we choose one of the computer's numbers is 1/10, and the probability of choosing all four numbers correctly is

$$\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right)\left(\frac{1}{10}\right) = \frac{1}{10,000} = 0.0001$$

Therefore, the probability we win is 0.0001 and the probability we lose is  $1 - 0.0001 = 0.9999$ .

$X$	$P(X)$	$XP(X)$
-3	0.9999	-2.9997
10,000	0.0001	1

Now we can find the expected value of playing the game long term.



$$E(X) = \mu_X = -2.9997 + 1 = -1.9997$$

So if we play this game over and over again, we expect to lose a little less than \$2 each time we play. To think about it a different way, here's our expected profit based on number of games played.

If we play 100 games, we expect approximately

$$100(-\$1.9997) = -\$199.97 \text{ in loss}$$

If we play 1,000 games, we expect approximately

$$1,000(-\$1.9997) = -\$1,999.70 \text{ in loss}$$

If we play 10,000 games, we expect approximately

$$10,000(-\$1.9997) = -\$19,997.00 \text{ in loss}$$

The variance is

$$\sigma_X^2 = (-3 - (-1.9997))^2(0.9999) + (10,000 - (-1.9997))^2(0.0001)$$

$$\sigma_X^2 = (-1.0003)^2(0.9999) + (10,001.9997)^2(0.0001)$$

$$\sigma_X^2 = 1.0005 + 10,003.9998$$

$$\sigma_X^2 \approx 10,005$$

so the standard deviation is

$$\sigma_X \approx 100.02$$

# Transforming random variables

Remember previously that we talked about how our measures of central tendency and spread would change if we shifted or scaled our data set.

Shifting the data set by a constant  $k$  means adding  $k$  to every value in the data set, or subtracting  $k$  from every value in the data set. On the other hand, scaling the data set by a constant  $k$  means multiplying or dividing every value in the data set by  $k$ .

We learned that shifting the data set would shift the mean, median and mode by the same amount as the constant, but that the range and IQR would stay the same. For example, shifting a data set up by  $k$  might look like this:

## Original data set

Mean: 6

Median: 7

Mode: 3

Range: 10

IQR: 8

## Shifted data set

Mean:  $6 + k$

Median:  $7 + k$

Mode:  $3 + k$

Range: 10

IQR: 8

To this list, let's add standard deviation. When we shift the data set up or down by  $k$  units, the standard deviation will stay the same. So

## Original data set

## Shifted data set

Mean: 6

Mean:  $6 + k$ 

Median: 7

Median:  $7 + k$ 

Mode: 3

Mode:  $3 + k$ 

Range: 10

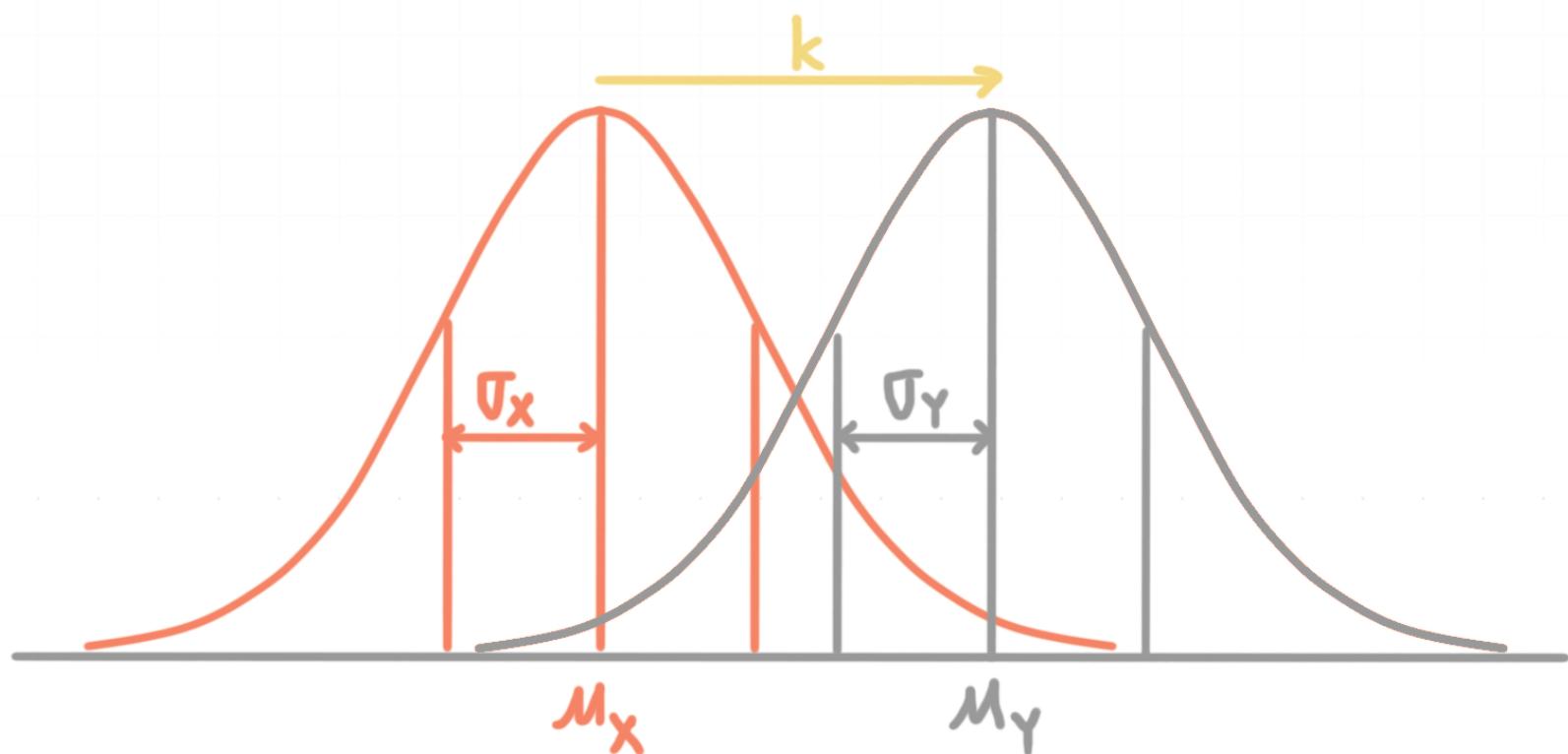
Range: 10

IQR: 8

IQR: 8

**Standard deviation:**  $\sigma$ **Standard deviation:**  $\sigma$ 

Here's how we'd visually represent shifting in the data.



We also learned that scaling the data set would equally scale the mean, median, mode, range and IQR. In other words, they all scale by the same factor. For example, scaling a data set by multiplying by  $k$  might look like this:

Original data setScaled data set

Mean: 6

Median: 7

Mode: 3

Range: 10

IQR: 8

Mean:  $6k$

Median:  $7k$

Mode:  $3k$

Range:  $10k$

IQR:  $8k$

But when we scale the data set by  $k$  units, the standard deviation will scale by the same value. So

### Original data set

Mean: 6

Median: 7

Mode: 3

Range: 10

IQR: 8

**Standard deviation:**  $\sigma$

### Scaled data set

Mean:  $6k$

Median:  $7k$

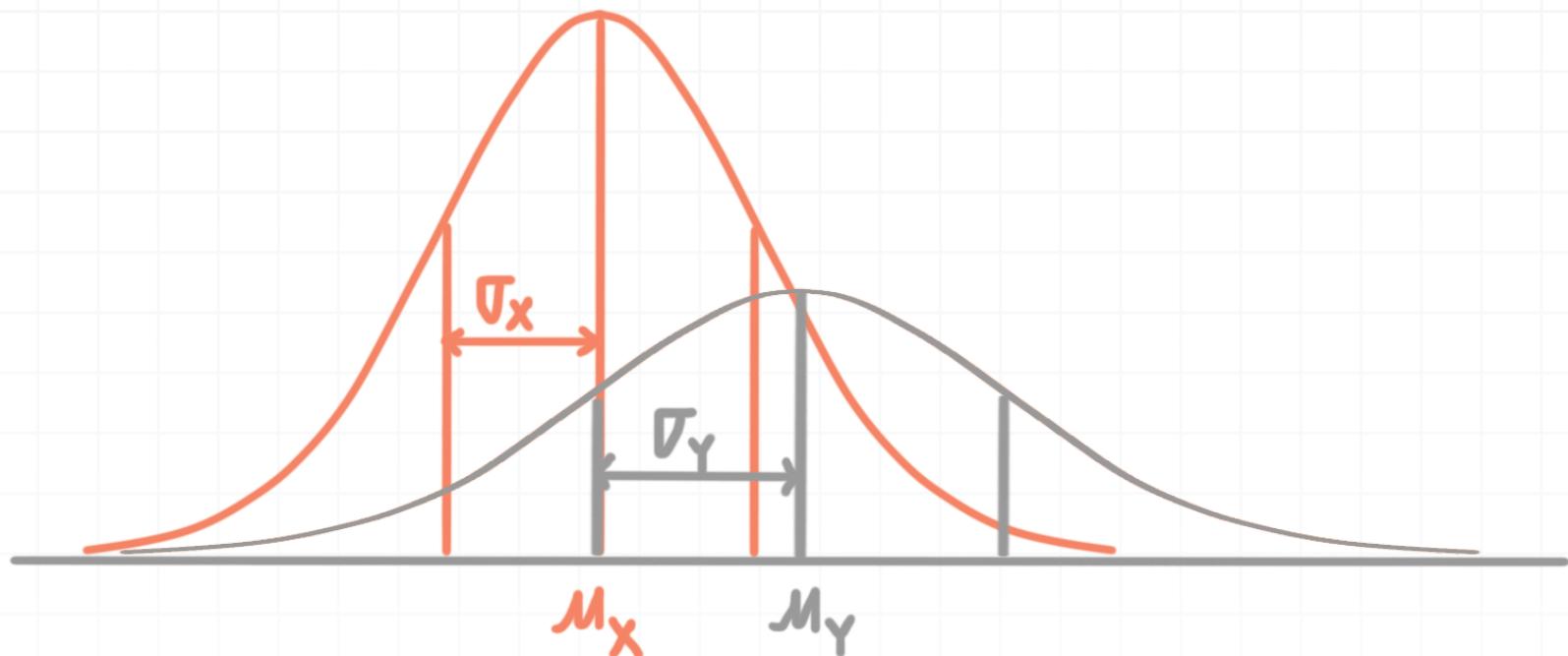
Mode:  $3k$

Range:  $10k$

IQR:  $8k$

**Standard deviation:**  $\sigma k$

Here's how we'd visually represent scaling in the data. The original distribution would become the new scaled version:



Let's do an example.

### Example

We're playing a game in which we pay \$5 for the chance to shoot a basketball two times. We win \$3 for every shot we make. Find the mean and standard deviation of the profit (or loss) we can expect if we play this game, assuming the table represents the probability distribution of  $X$ , the number of shots we make.

X	0	1	2
P(X)	0.25	0.49	0.26

The mean, or expected value, is

$$E(X) = \mu_X = 0(0.25) + 1(0.49) + 2(0.26)$$

$$E(X) = \mu_X = 0.49 + 0.52$$

$$E(X) = \mu_X = 1.01$$

Then the variance is

$$\sigma_X^2 = (0 - 1.01)^2(0.25) + (1 - 1.01)^2(0.49) + (2 - 1.01)^2(0.26)$$

$$\sigma_X^2 = (1.0201)(0.25) + (0.0001)(0.49) + (0.9801)(0.26)$$

$$\sigma_X^2 = 0.255025 + 0.000049 + 0.254826$$

$$\sigma_X^2 = 0.5099$$

so the standard deviation is

$$\sigma_X \approx 0.7141$$

From here, we can set up a net gain equation. Every time we play the game, if we make  $X$  shots, our expected net gain (or loss) can be given by  $N(X) = 3X - 5$ , because we get \$3 for every shot we make, but we have to pay \$5 to play. So the net gain (or loss) for every possible value of  $X$  is

$$N(0) = 3(0) - 5 = -5$$

$$N(1) = 3(1) - 5 = -2$$

$$N(2) = 3(2) - 5 = 1$$

and the probability distribution is therefore

$N$	-5	-2	1
$P(N)$	0.25	0.49	0.26

Although we could calculate the mean and standard deviation using this new table, let's do it instead by using  $N(X) = 3X - 5$  as a transformation of the variable  $X$ . In  $N(X) = 3X - 5$ ,  $X$  is scaled first,  $3X$ , and then shifted by  $-5$ .

The mean is affected by both scaling and shifting, so the mean of the net gain is

$$\mu_N = 3\mu_X - 5$$

$$\mu_N = 3(1.01) - 5$$

$$\mu_N = 3.03 - 5$$

$$\mu_N = -1.97$$

The standard deviation is affected only by scaling, not shifting, so the standard deviation of the net gain is

$$\sigma_N = 3\sigma_X$$

$$\sigma_N = 3(0.7141)$$

$$\sigma_N = 2.1423$$

Therefore, every time we play the game, we expect to lose \$1.97, with a standard deviation of 2.1423.

# Combinations of random variables

## Linear combinations of random variables

We just reviewed what happens when we shift or scale a data set by a constant value. But now we want to look at what happens when we combine two data sets, either by adding them or subtracting them.

For example, let's say I have two variables: how much time I spend each day walking and biking,  $W$  and  $B$  respectively. And let's say that I have data on my walking and biking habits for a full year, and I've already found the mean and standard deviation for both variables.

$$\mu_W = 1.1$$

$$\sigma_W = 0.2$$

$$\mu_B = 0.6$$

$$\sigma_B = 0.1$$

But now I want to know the mean for the sum of my walking and biking time together. In other words, I spend some time walking and biking each day, so I'd like to get an average for my total daily activity time.

We'll call the total activity  $A$ , which means we're looking for  $\mu_A$  (or the expected value  $E(A)$ ). We know that  $A = W + B$ . Here's the rule we need to remember: when we want to find the mean of the sum, we just find the sum of the means. So because  $A = W + B$ ,

$$\mu_A = \mu_W + \mu_B$$

$$\mu_A = 1.1 + 0.6$$

$$\mu_A = 1.7$$



But if we want to find the standard deviation of these two variables, we can't simply add the standard deviations together. In other words,  $\sigma_A = \sigma_W + \sigma_B$  is not a valid equation.

Instead, to find the standard deviation of the total activity, we need to square the two standard deviations. Remember that this is really giving us the variation for both walking and biking.

$$\sigma_W^2 = 0.2^2 = 0.04$$

$$\sigma_B^2 = 0.1^2 = 0.01$$

Then we add these together to get the sum of the variances, which gives us the variance for total activity.

$$\sigma_A^2 = \sigma_W^2 + \sigma_B^2$$

$$\sigma_A^2 = 0.04 + 0.01$$

$$\sigma_A^2 = 0.05$$

Then to find standard deviation for total activity, we take the square root of both sides.

$$\sqrt{\sigma_A^2} = \sqrt{0.05}$$

$$\sigma_A \approx 0.22$$

Instead of the sum, we could also find the difference in my walking and biking times. We could define a new variable for the difference and call it



$D$ . Then the difference is  $D = W - B$ , and the expected value of the difference would be

$$E(D) = \mu_D = \mu_W - \mu_B$$

$$E(D) = \mu_D = 1.1 - 0.6$$

$$E(D) = \mu_D = 0.5$$

And the standard deviation of the difference would be

$$\sqrt{\sigma_D^2} = \sqrt{\sigma_W^2 + \sigma_B^2}$$

$$\sqrt{\sigma_D^2} = \sqrt{0.04 + 0.01}$$

$$\sqrt{\sigma_D^2} = \sqrt{0.05}$$

$$\sigma_D \approx 0.22$$

One important thing to note is that, regardless of whether we're finding the sum of the variables, or the difference of the variables, in both cases we take the sum of the variances  $\sigma_W^2 + \sigma_B^2$ . We don't use the sum of the variances for the sum, and the difference of the variances for the difference; we always use the sum for both.

When we find the mean of the sum or difference of variables, it doesn't matter whether or not the variables are dependent or independent. In other words, if the variables are dependent, we can find a valid mean of their sum or difference. And if the variables are independent, we can find a valid mean of their sum or difference.



But in order to find the standard deviation of the sum or difference of two variables, the variables must be independent. So we can summarize what we know about the formulas this way:

	<b>Combination</b>	<b>Mean</b>	<b>Variance</b>
Sum	$S = X + Y$	$\mu_S = \mu_X + \mu_Y$	$\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$
Difference	$D = X - Y$	$\mu_D = \mu_X - \mu_Y$	$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$

## Combinations of normally distributed variables

When we combine variables that are both normally distributed, the combination will be normally distributed as well.

So if we're given the mean and standard deviation of two normally distributed variables, we can calculate the mean and standard deviation of the new combination.

But then, since the combination is normally distributed, we can use what we know about the probability under normal distributions to answer probability questions about the combination.

### Example

A popcorn company fills each of its variety popcorn tins with three flavors of popcorn: white cheddar, caramel, and chocolate covered. The amount of each flavor of popcorn that gets packed in the tin is normally distributed with a mean of 1 pound and a standard deviation of 0.1



pounds. The amount of each popcorn flavor is independent from the other flavors.

If  $W$  is the total weight of popcorn in a randomly selected tin, find the probability that the tin contains less than 3.25 pounds.

We have three normally distributed variables, one for each flavor. Their means are

$$\text{White cheddar} \quad \mu_D = 1$$

$$\text{Caramel} \quad \mu_M = 1$$

$$\text{Chocolate covered} \quad \mu_C = 1$$

Therefore, the mean of the combination (the mean weight of a full tin) is

$$\mu_W = \mu_D + \mu_M + \mu_C$$

$$\mu_W = 1 + 1 + 1$$

$$\mu_W = 3$$

The standard deviations of the three normally distributed variables are

$$\text{White cheddar} \quad \sigma_D = 0.1$$

$$\text{Caramel} \quad \sigma_M = 0.1$$

$$\text{Chocolate covered} \quad \sigma_C = 0.1$$



To find the standard deviation of the combination (the standard deviation of the weight of a full tin), we'll find the variance of the combination.

$$\sigma_W^2 = \sigma_D^2 + \sigma_M^2 + \sigma_C^2$$

$$\sigma_W^2 = 0.1^2 + 0.1^2 + 0.1^2$$

$$\sigma_W^2 = 0.01 + 0.01 + 0.01$$

$$\sigma_W^2 = 0.03$$

So the standard deviation is

$$\sigma_W = \sqrt{0.03}$$

$$\sigma_W \approx 0.1732$$

Now that we have the mean  $\mu_W = 3$  and standard deviation  $\sigma_W \approx 0.1732$  of the normally distributed weight of the full tin, we can answer probability questions about the combined normal distribution. We want to find the probability that the tin contains less than 3.25 pounds, so we'll calculate the  $z$ -score for 3.25.

$$z = \frac{x - \mu_W}{\sigma_W}$$

$$z = \frac{3.25 - 3}{0.1732}$$

$$z \approx 1.44$$

If we look up  $z = 1.44$  in a  $z$ -table, we find the value 0.9251, which means there's an approximately 93 % chance that the weight of the full tin is less than 3.25 pounds.

---



# Permutations and combinations

In order to answer many probability questions, we need to understand permutations and combinations. A **permutation** is the number of ways we can arrange a set of things, and the order matters.

The formula for a permutation is

$${}_n P_k = \frac{n!}{(n - k)!}$$

where  $n$  is the total number of items we have, and  $k$  is the number of items we want to arrange.

## Example

I have 4 scoops of ice cream: 1 chocolate, 1 strawberry, 1 vanilla, and 1 mint. I want to eat only 3 of the scoops. How many different ways can I eat 3 of the scoops if I consider both which scoops I eat and the order in which I eat them?

This is a permutation question, since I care about the order in which I eat the scoops. There are 4 total scoops, but I only want to eat 3 of them. Therefore, the number of ways I could eat three of the scoops of ice cream is

$${}_n P_k = \frac{n!}{(n - k)!}$$



$${}_4P_3 = \frac{4!}{(4-3)!}$$

$${}_4P_3 = \frac{4!}{1!}$$

$${}_4P_3 = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1}$$

$${}_4P_3 = 4 \cdot 3 \cdot 2$$

$${}_4P_3 = 24$$

There are 24 different ways that I could eat 3 of the 4 scoops. For example, chocolate-strawberry-vanilla would be 1 of the 24 options, but since order matters, chocolate-vanilla-strawberry would be another option.

---

On the other hand, a **combination** is the number of ways we can arrange a set of things, but the order doesn't matter. The formula for a combination is

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

where  $n$  is the total number of items we have, and  $k$  is the number of items we want to choose. Sometimes people write  ${}_nC_k$  as

$$\binom{n}{k}$$

which is called the binomial coefficient, and read as “ $n$  choose  $k$ .”



So to continue with the example from earlier, chocolate-strawberry-vanilla and chocolate-vanilla-strawberry would not count separately, because we don't care about the order when we're talking about combinations. All we care about is which items we picked, so chocolate-strawberry-vanilla and chocolate-vanilla-strawberry would count as the same thing.

### Example

I have the same 4 scoops of ice cream: 1 chocolate, 1 strawberry, 1 vanilla, and 1 mint. I want to eat only 3 of the scoops, and I don't care about the order in which I eat my 3 scoops. How many different combinations of 3 scoops can I create?

This is a combination question, since I don't care about the order in which I eat the scoops. There are 4 total scoops, but I only want to eat 3 of them. Therefore, the number of ways I could eat three of the scoops of ice cream is

$${}_nC_k = \frac{n!}{k!(n-k)!}$$

$${}_4C_3 = \frac{4!}{3!(4-3)!}$$

$${}_4C_3 = \frac{4!}{(3!)(1!)}$$

$${}_4C_3 = \frac{4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(1)}$$

$${}_4C_3 = \frac{4}{1}$$

$${}_4C_3 = 4$$

There are 4 different ways that I could eat 3 of the 4 scoops. For example, chocolate-strawberry-vanilla would be 1 of the 4 options, but since order doesn't matter, chocolate-vanilla-strawberry would be the same combination, and wouldn't count as another one of the 4 combinations.

---

# Binomial random variables

Remember that “bi” means two, so a **binomial variable** is a variable that can take on exactly two values. A coin is the most obvious example of a binomial variable because flipping a coin can only result in two values: heads or tails. On the other hand, “rolling a die until a 3 appears” can’t be represented by a binomial random variable. We’ll learn more about why when we look at the required characteristics of a binomial random variable.

The two outcomes that the binomial random variable can take do not have to be equally probable. The probability of getting heads when we flip a fair coin is a binomial random variable where the probability of “success” is 50 %. But the probability of choosing a girl from our math class when we randomly choose one student might be a binomial random variable where the probability of “success” is 70 % (if 70 % of the students in our class are female).

## Binomial random variables

In order for a variable  $X$  to be a **binomial random variable**,

- each trial must be independent,
- each trial can be called a “success” or “failure,”
- there are a fixed number of trials, and
- the probability of success on each trial is constant.



Let's think about  $X$  as how many heads we get when we flip a coin 10 times.

- Each trial is independent, because the result of one flip doesn't affect the result of any other flip.
- Each trial can be called a success or failure because "heads" is a success and "tails" is a failure.
- We're flipping the coin 10 times, so there are a fixed number of trials.
- The probability of getting heads on the first flip is 50 % , the probability of getting heads on the second flip is 50 % , the probability of getting heads on the third flip is 50 % , etc. The probability of heads in each trial is constant.

So if  $X$  is how many heads we get when we flip a coin 10 times, then we could call  $X$  a binomial random variable. On the other hand, let's think about  $J$  as how many jacks we get when we pull 3 cards from a deck without replacing the card after each pull.

- Each trial is not independent, because the result of the first pull affects the result of every pull thereafter. If we get a jack on the first pull, then the probability we get a jack on the second pull is 3/51. But if we get something other than a jack on the first pull, then the probability we get a jack on the second pull is 4/51.
- Each trial can be called a success or failure because a jack is a success and anything else is a failure.



- We're pulling a card 3 times, so there are a fixed number of trials.
- The probability of getting a jack on the first pull is  $4/52$ . But the probability of getting a jack on the second pull changes depending on what we got on the first pull. The probability of pulling a jack in each trial is not constant.

Since it fails two of the four conditions,  $J$  cannot be called a binomial random variable.

## Binomial probability

In binomial probability questions, we're often asked to figure out the probability that we get an exact number of “successes,” assuming we perform a specific number of independent trials. The formula we'll use for this is

$$P(k \text{ successes in } n \text{ attempts}) = \binom{n}{k} p^k (1-p)^{n-k}$$

where the binomial coefficient

$$\binom{n}{k}$$

is the combination  $_nC_k$ ,  $p$  is the probability of a success,  $k$  is the exact number of times we want the success, and  $n$  is the total number of independent trials we'll run.

---

### Example



Let's say there are three marbles in a bag: 2 are green and 1 is red. We're going to do 5 trials where we pull a marble, note the color, and then replace the marble. What is the probability that we get the red marble exactly 3 times?

First, we'll confirm that this is a binomial random variable.

- 1) Since we're replacing each marble we pull before pulling another, each pull is an independent trial.
- 2) We can classify a red marble as a success, and green marble as a failure.
- 3) There are a fixed number of trials: 5.
- 4) The probability of success (getting a red marble) is constant through each trial, since we're replacing the marbles. Since all four conditions are met, this is a binomial random variable.

We're trying to pull a red marble exactly 3 times in 5 pulls. To solve this problem, we need to first figure out how many possible combinations we can do this in.

$${}_5C_3 = \binom{5}{3} = \frac{5!}{3!(5-3)!}$$

$${}_5C_3 = \binom{5}{3} = \frac{5!}{3!2!}$$

$${}_5C_3 = \binom{5}{3} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 2 \cdot 1}$$

$${}_5C_3 = \binom{5}{3} = \frac{5 \cdot 4}{2 \cdot 1}$$

$${}_5C_3 = \binom{5}{3} = \frac{20}{2}$$

$${}_5C_3 = \binom{5}{3} = 10$$

Then to find the probability that we get exactly 3 reds on 5 pulls, we say that  $f$  is the probability of pulling a red marble, and therefore that the probability of pulling 3 red marbles in 5 pulls is

$$P(k \text{ successes in } n \text{ attempts}) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(3 \text{ reds in 5 pulls}) = (10) \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2$$

$$P(3 \text{ reds in 5 pulls}) = \frac{40}{243} \approx 16.5\%$$

In other words, we have approximately a 16.5% chance of getting exactly 3 red marbles when we pull a random marble from this bag 5 times.

## Probability distributions for binomial random variables



We can also create a probability distribution for binomial random variables. Using our example of pulling a marble 5 times, where we have a 1/3 probability of pulling a red marble and 2/3 probability of pulling a green marble on each pull, we could calculate the following probabilities.

$$P(0 \text{ red in 5 pulls}) = \binom{5}{0} 0.33^0 0.67^5 = (1)(1)(0.67^5) \approx 0.1350$$

$$P(1 \text{ red in 5 pulls}) = \binom{5}{1} 0.33^1 0.67^4 = (5)(0.33^1)(0.67^4) \approx 0.3325$$

$$P(2 \text{ red in 5 pulls}) = \binom{5}{2} 0.33^2 0.67^3 = (10)(0.33^2)(0.67^3) \approx 0.3275$$

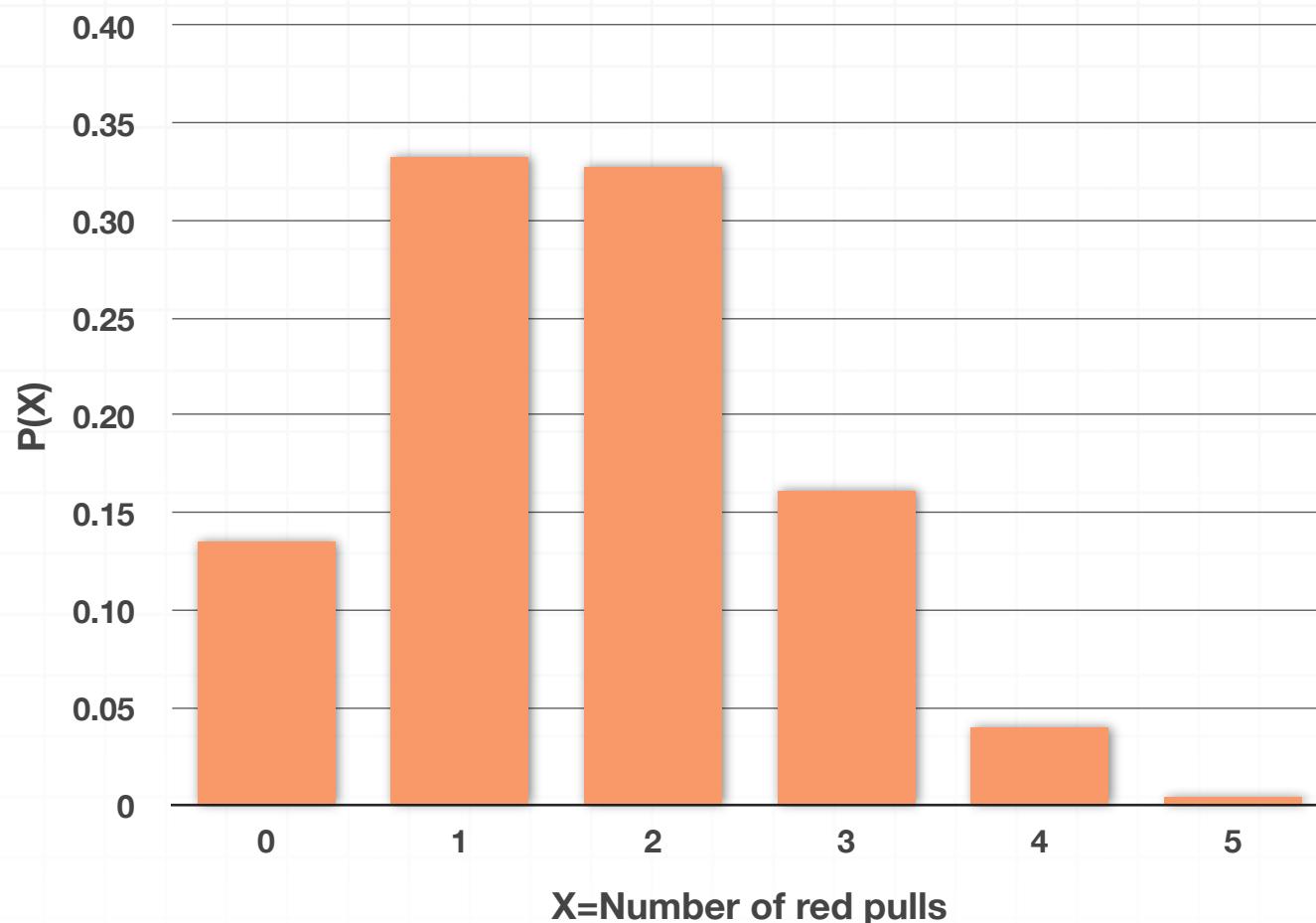
$$P(3 \text{ red in 5 pulls}) = \binom{5}{3} 0.33^3 0.67^2 = (10)(0.33^3)(0.67^2) \approx 0.1613$$

$$P(4 \text{ red in 5 pulls}) = \binom{5}{4} 0.33^4 0.67^1 = (5)(0.33^4)(0.67^1) \approx 0.0397$$

$$P(5 \text{ red in 5 pulls}) = \binom{5}{5} 0.33^5 0.67^0 = (1)(0.33^5)(0.67^0) \approx 0.0039$$

We could then plot this probability distribution to get a picture of the probability.





This probability distribution is a visual representation of the fact that we're most likely to get 1 or 2 red marbles when we pull 5 times from the bag of marbles. We're much less likely to get 4 or 5 red marbles when we pull 5 times.

# Poisson distributions

Like the binomial distribution, the Poisson distribution models a discrete random variable, and it's particularly useful for finding the probability that a specific number of events will occur in a given period of time.

## The Poisson process

A **Poisson process** calculates the number of times an event occurs in a period of time, or in a particular area, or over some distance, or within any other kind of measurement, and the process has particular characteristics:

1. The experiment counts the number of occurrences of an event over some other measurement,
2. The mean is the same for each interval,
3. The count of events in each interval is independent of the other intervals, and
4. The intervals don't overlap.
5. The probability of the event occurring is proportional to the period of time.

The Poisson process is useful in modeling many real-life events, such as radioactive decay, the number of visitors to a website, or even the number of trees in an acre of forest.



Let's say we want to use a Poisson process to model the number of cars that pass through an intersection each hour. In this case, because we're counting the number of occurrences of an event over time, we've met the first condition.

In order to meet the second condition, we need to be able to assume that the average number of cars that passes through the intersection each hour is the same. In other words, if the average number of cars passing through the intersection daily between 9 : 00 a.m. and 10 : 00 a.m. is 15, then the average for each other hour of the day needs to be 15 as well.

To meet the third condition, we need to be able to say that the number of cars that pass through the intersection this hour is not affected by the number of cars that came through during a previous hour, and that the number of cars passing through this hour will not affect the number of cars that pass through in the coming hours. In other words, the car count for each hour is independent of the count for any other hour.

To meet the fourth condition, we need our intervals to be non-overlapping. So if we take data from 9 : 00 a.m. to 10 : 00 a.m. and from 10 : 00 a.m. to 11 : 00 a.m., then we can't include a data set for 9 : 30 a.m. to 10 : 30 a.m., because that data would overlap with the data from the other intervals.

To meet the fifth condition, we need to scale the probability of the number of cars to scale based on the size of the interval. For instance, if the probability that  $C$  cars pass through the intersection in one hour is  $P(C)$ , then the probability that  $2C$  cars pass through the intersection in two hours is  $2P(C)$ .



Assuming we meet all of those conditions, then we'll be able to use a Poisson process. In this example, the discrete random variable would be the actual number of cars passing through the intersection.

## Probability formula for Poisson

The probability of exactly  $x$  occurrences of the event, when the mean number of occurrences in the interval is  $\lambda$ , is

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

So if we believe the mean number of cars passing through the intersection in any particular hour is  $\lambda = 15$ , and if we want to know the probability that  $x = 13$  cars will pass through it in the next hour, that probability will be

$$P(13) = \frac{15^{13} e^{-15}}{13!}$$

$$P(13) \approx 0.0956$$

$$P(13) \approx 9.56 \%$$

So there's an approximately 9.56 % chance that 13 cars will pass through the intersection in the next hour.

The Poisson formula can also be used to calculate cumulative probabilities. For instance, if we want to know the probability that at most 7 cars pass through the intersection in an hour, we calculate that as

$$P(x \leq 7) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$+P(x=4) + P(x=5) + P(x=6) + P(x=7)$$

$$P(x \leq 7) = \frac{15^0 e^{-15}}{0!} + \frac{15^1 e^{-15}}{1!} + \frac{15^2 e^{-15}}{2!} + \frac{15^3 e^{-15}}{3!}$$

$$+ \frac{15^4 e^{-15}}{4!} + \frac{15^5 e^{-15}}{5!} + \frac{15^6 e^{-15}}{6!} + \frac{15^7 e^{-15}}{7!}$$

$$P(x \leq 7) = \frac{1}{e^{15}} + \frac{15}{e^{15}} + \frac{15^2}{2e^{15}} + \frac{15^3}{6e^{15}}$$

$$+ \frac{15^4}{24e^{15}} + \frac{15^5}{120e^{15}} + \frac{15^6}{720e^{15}} + \frac{15^7}{5,040e^{15}}$$

$$P(x \leq 7) = \frac{1}{e^{15}} + \frac{15}{e^{15}} + \frac{15^2}{2e^{15}} + \frac{15^3}{6e^{15}} + \frac{15^4}{24e^{15}} + \frac{15^5}{120e^{15}} + \frac{15^6}{720e^{15}} + \frac{15^7}{5,040e^{15}}$$

$$P(x \leq 7) = \frac{1}{e^{15}} \left( 1 + 15 + \frac{15^2}{2} + \frac{15^3}{6} + \frac{15^4}{24} + \frac{15^5}{120} + \frac{15^6}{720} + \frac{15^7}{5,040} \right)$$

$$P(x \leq 7) \approx 0.018$$

So there's an approximately 1.8 % chance that 7 or fewer cars pass through the intersection in an hour.

## Poisson distribution to approximate the binomial distribution

We already know how to calculate binomial probabilities, but using the Poisson distribution instead can actually save us some time. The Poisson distribution very closely approximates the binomial distribution when the



number of binomial trials  $n$  is at least 20 and when the probability of success  $p$  in the binomial trial is at most 0.05.

So if those two conditions are met, then using the Poisson probability formula will give us a great approximation of the binomial probability. And the math for the Poisson probability is easier than the math for the binomial probability, so we'll save some time.

The probability of  $x$  successes in  $n$  attempts, given by the Poisson formula is

$$P(x) = \frac{(np)^x e^{-np}}{x!}$$

Notice that we've just substituted  $\lambda = np$  into the original Poisson formula. Let's do an example with the Poisson formula for a binomial random variable.

### Example

There are 30 students in a Kindergarten class and each one of them has a 4% chance of forgetting their lunch on any given day. What is the probability that exactly 5 of them will forget their lunch today.

This is a binomial experiment with  $n = 30$ ,  $p = 0.04$ , and  $x = 5$ . Because we have at least 20 “attempts,” and because the probability of a “success” is less than 5%, we can use the Poisson formula to estimate this binomial probability.



$$P(x) = \frac{(np)^x e^{-np}}{x!}$$

$$P(5) = \frac{(30 \cdot 0.04)^5 e^{-30 \cdot 0.04}}{5!}$$

$$P(5) = \frac{1.2^5 e^{-1.2}}{120}$$

$$P(5) \approx 0.006246$$

So the chance that exactly 5 of the Kindergarteners forget their lunch today is approximately 0.62%.

---

# “At least” and “at most,” and mean, variance, and standard deviation

We can do more than just calculate the probability of pulling exactly 3 red marbles in 5 total pulls. For any binomial random variable, we can also calculate something like the probability of pulling at least 3 red marbles, or the probability of pulling no more than 3 marbles.

What we want to know is that the probability of pulling at least 3 red marbles is the probability that we pull 3, or 4, or 5 red marbles, which is simply the probability of each of these, all added together.

$$P(\text{at least 3 reds in 5 pulls}) = P(3 \text{ reds}) + P(4 \text{ reds}) + P(5 \text{ reds})$$

$$\begin{aligned} P(\text{at least 3 reds in 5 pulls}) &= \binom{5}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 \\ &\quad + \binom{5}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 + \binom{5}{5} \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^0 \end{aligned}$$

$$\begin{aligned} P(\text{at least 3 reds in 5 pulls}) &= (10) \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 \\ &\quad + (5) \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 + (1) \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^0 \end{aligned}$$

$$P(\text{at least 3 reds in 5 pulls}) \approx 0.1646 + 0.0412 + 0.0041$$

$$P(\text{at least 3 reds in 5 pulls}) \approx 0.2099$$

$$P(\text{at least 3 reds in 5 pulls}) \approx 21\%$$

In the same way, the probability of pulling at most 3 red marbles would be the probability of pulling 0, 1, 2, or 3 red marbles, all added together.

If we're calculating the probability of at least one success or at least one failure, we can use these formulas:

$$P(\text{at least 1 success}) = 1 - P(\text{all failures})$$

$$P(\text{at least 1 failure}) = 1 - P(\text{all successes})$$

This is because all probability distribution functions must add up to 1.

### Example

Find the probability that we get at least 1 heads on 5 coin flips.

We can actually simplify this problem a lot by realizing that every single set of 5 coin flips will have at least one heads, unless every one of the 5 flips is tails: *TTTTT*. The probability of getting 5 tails in a row is

$$P(TTTTT) = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) \left(\frac{1}{2}\right) = \frac{1}{32}$$

The probability of getting “at least one heads” is the same as the probability of not getting “all tails.” Therefore, since total probability is always equal to 1, we can say that the probability of at least one heads is

$$P(\text{at least 1 heads}) = 1 - \frac{1}{32} = \frac{31}{32}$$



Let's do another example where we find an “at most” probability for a binomial random variable.

### Example

Let  $X$  be a binomial random variable with  $n = 10$  and  $p = 0.30$ . Find  $P(X \leq 5)$ .

The variable  $X$  follows a binomial distribution, but instead of finding the probability of exactly  $k$  successes in  $n$  trials, we’re asked to find the probability of  $k$  or fewer successes in  $n$  trials. Specifically, find the chance of 5 or fewer successes in 10 trials, where the probability of success on any one trial is  $p = 0.30$ .

Find the probability of 0 successes, 1 success, 2 successes, etc., up to 5 successes, and then find the sum of those probabilities.

$$P(X \leq 5) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

To find the probability for each value of  $k$ , we use the binomial probability formula.

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1 - p)^{n-k}$$

So the probability  $P(X \leq 5)$  is

$$P(X \leq 5) = \binom{10}{0} (0.30)^0 (1 - 0.30)^{10} + \binom{10}{1} (0.30)^1 (1 - 0.30)^9$$



$$+ \binom{10}{2} (0.30)^2 (1 - 0.30)^8 + \binom{10}{3} (0.30)^3 (1 - 0.30)^7$$

$$+ \binom{10}{4} (0.30)^4 (1 - 0.30)^6 + \binom{10}{5} (0.30)^5 (1 - 0.30)^5$$

$$P(X \leq 5) = 0.9527$$


---

## Mean, variance, and standard deviation

The mean of a binomial random variable  $X$  can be expressed as  $\mu_X$ . The mean is also called the expected value, and that's indicated as  $E(X)$ . Either way, the mean is given by

$$\mu_X = E(X) = np$$

where  $n$  is the fixed number of independent trials, and  $p$  is the probability of a success. The variance of a binomial random variable  $X$  is given by

$$\sigma_X^2 = np(1 - p)$$

Standard deviation is the square root of the variance and is therefore given by

$$\sqrt{\sigma_X^2} = \sqrt{np(1 - p)}$$

$$\sigma_X = \sqrt{np(1 - p)}$$

If we continue with our example of the number of heads we get on 5 coin flips, we can say that the number of trials  $n$  is 5, and the probability of success (getting heads) is  $p = 0.5$ . Therefore, the mean is

$$\mu_X = np$$

$$\mu_X = 5(0.5)$$

$$\mu_X = 2.5$$

The variance is

$$\sigma_X^2 = np(1 - p)$$

$$\sigma_X^2 = 5(0.5)(1 - 0.5)$$

$$\sigma_X^2 = 2.5(1 - 0.5)$$

$$\sigma_X^2 = 2.5(0.5)$$

$$\sigma_X^2 = 1.25$$

And the standard deviation is

$$\sigma_X = \sqrt{np(1 - p)}$$

$$\sigma_X = \sqrt{1.25}$$

$$\sigma_X \approx 1.12$$

# Bernoulli random variables

Earlier we defined a binomial random variable as a variable that takes on the discrete values of “success” or “failure.” For example, if we want heads when we flip a coin, we could define heads as a success and tails as a failure. We could model this scenario with a binomial random variable  $X$  where  $X$  is the number of times we get heads when we flip a coin a specified number of times.

A **Bernoulli random variable** is a special category of binomial random variables. Specifically, with a Bernoulli random variable, we have exactly one trial only (binomial random variables can have multiple trials), and we define “success” as a 1 and “failure” as a 0.

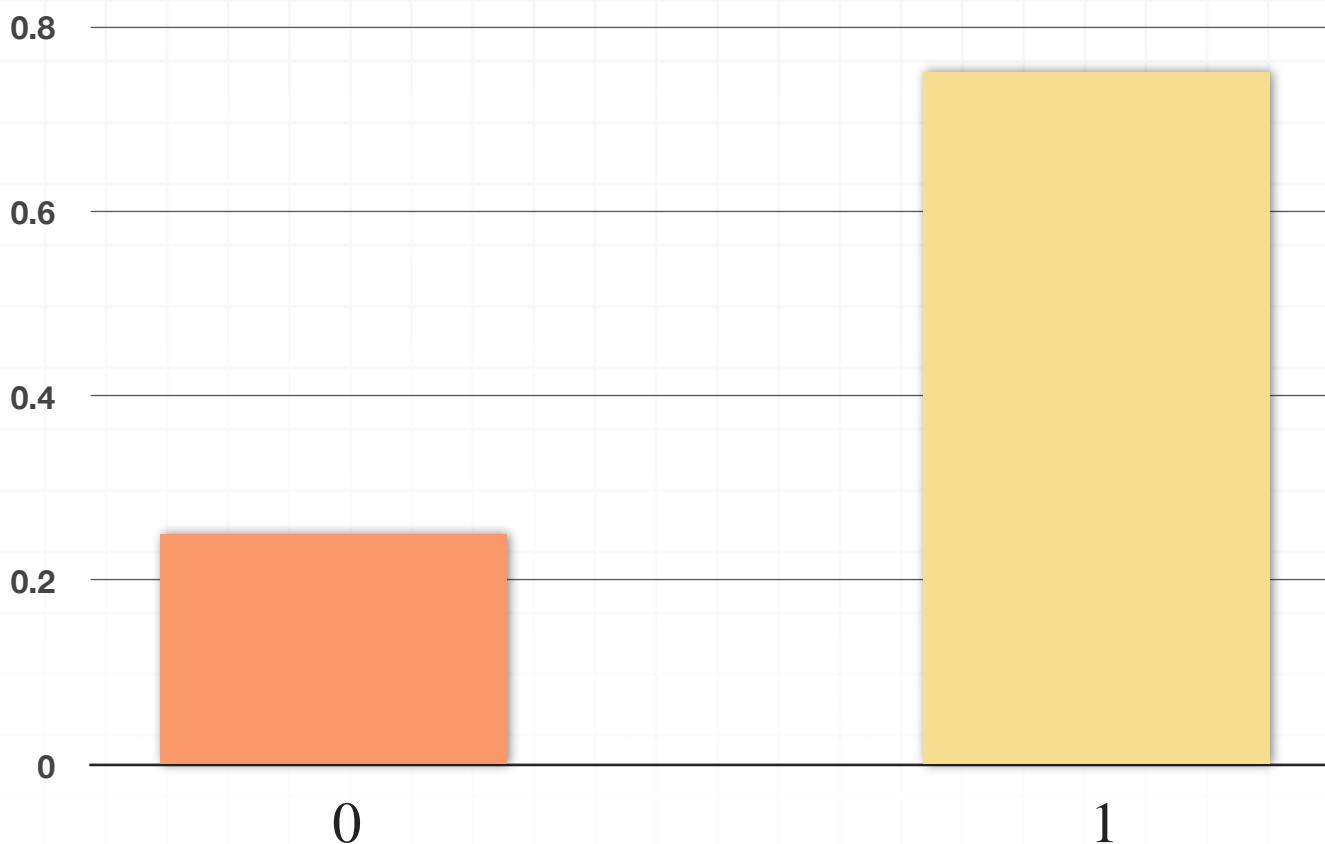
## Bernoulli distributions

Let’s say I want to know how many students in my school like peanut butter. I can’t survey the entire school, so I survey only the students in my class, using them as a sample. I ask them whether or not they like peanut butter, and I define “liking peanut butter” as a success with a value of 1 and “disliking peanut butter” as a failure with a value of 0. I find that 75 % of the students in my class like peanut butter.

Since everyone in our survey was forced to pick one choice or the other, 100 % of our population is represented in these two categories, which means that the probability of both options will always sum to 1.0 or 100 %. Therefore, since 75 % of the students in my class like peanut butter, that means  $100\% - 75\% = 25\%$  of the students dislike peanut butter.



I could represent this in a Bernoulli distribution as



## Mean, variance, and standard deviation

### Mean

Finding the mean of a Bernoulli random variable is a little counter-intuitive. It seems like we have discrete categories of “dislike peanut butter” and “like peanut butter,” and it doesn’t make much sense to try to find a mean and get a “number” that’s somewhere “in the middle” and means “somewhat likes peanut butter?” It’s all just a little bizarre.

How do we get around this? Well, we mentioned it before, but we assign a value of 0 to the failure category of “dislike peanut butter,” and a value of

1 to the success category of “like peanut butter.” Then we can take the probability weighted sum of the values in our Bernoulli distribution.

$$\mu = (\text{percentage of failures})(0) + (\text{percentage of successes})(1)$$

$$\mu = (0.25)(0) + (0.75)(1)$$

$$\mu = 0 + 0.75$$

$$\mu = 0.75$$

This is the mean of the Bernoulli distribution. Notice how the value we found for the mean is equal to the percentage of “successes.” We said that “liking peanut butter” was a “success,” and then we found that 75 % of our class liked peanut butter, so the mean of the distribution was going to be  $\mu = 0.75$ .

If we want to create a general formula for finding the mean of a Bernoulli random variable, we could call the probability of success  $p$ , and then call the probability of failure  $1 - p$  (since total probability always sums to 1, and  $p + (1 - p) = p + 1 - p = 1$ ). Then with failure represented by 0 and success represented by 1, the mean (also called the expected value) will always be

$$\mu = (1 - p)(0) + (p)(1)$$

$$\mu = 0 + p$$

$$\mu = p$$

And we see again that the mean is the same as the probability of success,  $p$ . Realize too that, even though we found a mean of  $\mu = 0.75$ , the



distribution is still discrete. No one in the population is going to take on a value of  $\mu = 0.75$ ; everyone will either be exactly a 0 or exactly a 1.

## Variance and standard deviation

We'll use a similar weighting technique to calculate the variance for a Bernoulli random variable. We'll find the difference between both 0 and the mean and 1 and the mean, square that distance, and then multiply by the "weight."

$$\sigma^2 = (0.25)(0 - \mu)^2 + (0.75)(1 - \mu)^2$$

$$\sigma^2 = (0.25)(0 - 0.75)^2 + (0.75)(1 - 0.75)^2$$

$$\sigma^2 = (0.25)(-0.75)^2 + (0.75)(0.25)^2$$

$$\sigma^2 = (0.25)(0.5625) + (0.75)(0.0625)$$

$$\sigma^2 = 0.140625 + 0.046875$$

$$\sigma^2 = 0.1875$$

The standard deviation of a Bernoulli random variable is still just the square root of the variance, so the standard deviation is

$$\sqrt{\sigma^2} = \sqrt{0.1875}$$

$$\sigma = 0.4330$$

The general formula for variance is always given by

$$\sigma^2 = (1 - p)(0 - p)^2 + (p)(1 - p)^2$$

$$\sigma^2 = (1 - p)(-p)^2 + (p)(1 - p)^2$$

$$\sigma^2 = (1 - p)(p^2) + (p)(1 - 2p + p^2)$$

$$\sigma^2 = p^2 - p^3 + p - 2p^2 + p^3$$

$$\sigma^2 = p^2 + p - 2p^2$$

$$\sigma^2 = p - p^2$$

$$\sigma^2 = p(1 - p)$$

Notice that this is just the probability of success  $p$  multiplied by the probability of failure  $1 - p$ . Therefore, standard deviation of the Bernoulli random variable is always given by

$$\sqrt{\sigma^2} = \sqrt{p(1 - p)}$$

$$\sigma = \sqrt{p(1 - p)}$$



# Geometric random variables

Remember that for a binomial random variable  $X$ , we're looking for the number of successes in a finite number of trials.

For a **geometric random variable**, most of the conditions we put on the binomial random variable still apply:

- each trial must be independent,
- each trial can be called a “success” or “failure,”
- the probability of success on each trial is constant.

The difference is that for a geometric random variable, we're looking at how many trials we have to use until we get a certain success. For a binomial random variable, we decided ahead of time on a certain number of trials. But for a geometric random variable, we'll run an infinite number of trials until we get a success.

For example, “flipping a coin until we get heads” could be described by a geometric random variable. It might take just one flip to get heads, but it could take us 5, 10, or (though very, very unlikely) 10,000 flips.

To find the probability that a success  $S$  occurs on the  $n$ th attempt, when a success has a probability of  $p$ , and therefore a failure has a probability of  $1 - p$ , we use this formula:

$$P(S = n) = p(1 - p)^{n-1}$$



If we look closely at this formula, we see that we're really just multiplying the probability of failure over and over again until the trial right before we have a success, and then multiplying by the probability of a success.

In other words, if we want to find the probability that we get our first success on the 7th trial, then the probability will be

$$P(\text{success on the 7th trial}) = (\text{probability of failure})^6(\text{probability of success})^1$$

Notice that the exponents add to 7, since we needed 7 trials to get the first success.

### Example

I'm playing a game where the probability of winning a prize is 0.7. What is the probability that I don't win a prize until the 4th time I play the game, assuming each game is independent?

We're looking for the probability that I don't "succeed" until the 4th "trial," so we can represent this with a geometric random variable.

Since the probability of success is 0.7, it means the probability of failure is 0.3. Since I fail 3 times, and then succeed once on the 4th game, the probability of this happening is

$$P(S = 4) = (0.3)^3(0.7)^1$$

$$P(S = 4) = (0.027)(0.7)$$

$$P(S = 4) = 0.0189$$

$$P(S = 4) \approx 2\%$$

There's an approximately 2% chance that I don't win a prize until the fourth game.

---

## More than and less than

### Less than

Sometimes we can be asked to find the probability that it takes less than a specific number of trials in order to get our first success. For instance, continuing with the example we just worked through, we could be asked to find the probability that it takes us less than 4 games to win a prize.

This is the same as saying that we win a prize on game 1, 2, or 3. If we call a success  $S$ , that means we want either  $S < 4$  or  $S \leq 3$ , which mean the same thing in the case of a geometric random variable.

$$P(S < 4) = P(S = 1) + P(S = 2) + P(S = 3)$$

The probability of success is 0.7 and the probability of failure is 0.3. When  $S = 1$ , that means we have 0 failures before we then have 1 success. When  $S = 2$ , that means we have 1 failure and then 1 success. When  $S = 3$ , that means we have 2 failures and then 1 success.

$$P(S < 4) = (0.3)^0(0.7)^1 + (0.3)^1(0.7)^1 + (0.3)^2(0.7)^1$$

$$P(S < 4) = (1)(0.7) + (0.3)(0.7) + (0.09)(0.7)$$



$$P(S < 4) = 0.7 + 0.21 + 0.063$$

$$P(S < 4) = 0.973$$

$$P(S < 4) = 97.3 \%$$

## At most

This is slightly different than being asked the probability that it takes us less than 4 games to win a prize. If it takes less than 4 games to win, that means we get a prize in the third game, or earlier. But if it takes us at most 4 games to win, that means we could win a prize in the fourth game. We could write that as  $S < 5$  or as  $S \leq 4$ . But either way, we fail no more than 3 times and then succeed in the fourth game, at the latest.

## More than

Similarly, we'll be asked to find the probability that it takes more than a specific number of trials in order to get our first success. For instance, continuing with the same example, we could be asked to find the probability that it takes more than 2 games for us to win a prize.

Remember that all probability distributions add to 1. If we're looking for the probability that it takes more than 2 trials to win a prize, we can find the probability of winning on the first trial and the probability of winning on the second trial, and then subtract those probabilities from 1, which will give us all the total probability of all outcomes, other than winning on the first or second game.

So the probability that it takes more than 2 games to win is

$$P(S > 2) = 1 - P(S \leq 2)$$



$$P(S > 2) = 1 - [(0.3)^0(0.7)^1 + (0.3)^1(0.7)^1]$$

$$P(S > 2) = 1 - [(1)(0.7) + (0.3)(0.7)]$$

$$P(S > 2) = 1 - (0.7 + 0.21)$$

$$P(S > 2) = 1 - 0.91$$

$$P(S > 2) = 0.09$$

$$P(S > 2) = 9\%$$

Keep in mind that we also could have written  $S > 2$  as  $S \geq 3$ , or  $S \leq 2$  as  $S < 3$ .

### At least

This is slightly different than being asked the probability that it takes us more than 2 games to win a prize. If it takes more than 2 games to win, that means we don't get a prize until the third game. But if it takes us at least 2 games to win, that means we could win a prize in the second game. We could write that as  $S > 1$  or as  $S \geq 2$ . But either way, we failed once and then succeeded sometimes in the second game or later.

$$P(S \geq 2) = 1 - P(S \leq 1)$$

$$P(S \geq 2) = 1 - P(S = 1)$$

$$P(S \geq 2) = 1 - (0.3)^0(0.7)^1$$

$$P(S \geq 2) = 1 - (1)(0.7)$$

$$P(S \geq 2) = 1 - 0.7$$

$$P(S \geq 2) = 0.3$$

$$P(S \geq 2) \approx 30\%$$

## Mean, variance, and standard deviation

### Mean

The mean  $\mu_X$  of a geometric random variable, which can also be called the expected value  $E(X)$  is given by

$$\mu_X = E(X) = \frac{1}{p}$$

where the probability of a success on a trial is  $p$ , and  $X$  is the number of independent trials required to get the first success.

So in our example from this section where we have a 70% chance of winning a prize, the mean is

$$\mu_X = \frac{1}{0.7} \approx 1.43$$

This means we should expect to win the game if we play about one or two times.

### Variance and standard deviation

The variance  $\sigma_X^2$  of a geometric random variable is given by

$$\sigma_X^2 = \frac{1-p}{p^2}$$

and standard deviation is the square root of the variance. Therefore, the variance of the geometric random variable we've been working with is

$$\sigma_X^2 = \frac{1 - 0.7}{0.7^2}$$

$$\sigma_X^2 = \frac{0.3}{0.49}$$

$$\sigma_X^2 \approx 0.61$$

and the standard deviation is

$$\sqrt{\sigma_X^2} \approx \sqrt{0.61}$$

$$\sigma_X \approx 0.78$$

# Types of studies

If we want to put to good use everything we've learned so far about data, we'll need to know how to run studies in a way that gives us good, reliable data. In this section we'll talk about different kinds of studies we can use to collect data, including observational studies and experiments.

In the next section we'll talk about how to make sure these studies are producing reliable results.

## The goal of collecting samples

The purpose of statistics is to gather information (data) about the world around us and analyze it in some way to help make sense of it.

Because collecting data for an entire population is usually difficult or impossible, we instead choose a smaller sample of the larger population, and then analyze the data for the sample, hoping that our results will translate to the larger population.

Characteristics like mean and standard deviation are called **statistics** when we calculate them for a sample. A **parameter** is the corresponding characteristic of the population that the statistic is trying to estimate. So we could choose a sample, calculate the sample mean (a statistic), and then use what we know about the sample mean to make inferences about the population mean (a parameter).

## Observational study

In an **observational study**, we're just looking at the information that's already there, or measuring it in some way, but we're adding nothing to the population that will change it in any way. In statistics, something that changes a population is called a **treatment**, so for an observational study, no treatment is applied.

### One-way tables

For example, let's say we want to know whether all the students at our school prefer peanut butter or jelly. We may choose to use the students in our classroom as a sample in order to estimate the preferences of the entire population (all the students at our school).

We ask every student in your classroom if they prefer peanut butter or jelly, and they give us an answer of “peanut butter” or an answer of “jelly.” We now have data for a one-way table in which the individuals are the students in our classroom, and the variable is “Peanut butter or jelly?” If we find that 70% of our classmates prefer peanut butter, we might infer that 70% of all the students at our school also prefer peanut butter.

Notice that we didn't do anything here except ask a question and record the responses. We didn't do anything that would change anyone's mind in any way, because we just wanted to make an observation about what was already going on.

Keep in mind that it's only true that 70% of the students in our school prefer peanut butter if the students in our classroom make up a random, representative, unbiased sample of the whole school, which they may not.



In fact, we'd say that we introduced bias into our study by convenience sampling, which we'll talk about soon.

## Two-way tables

Sometimes we might want to collect data in an observational study for a two-way table, (as opposed to just a one-way table in the above example) and understand how two parameters might move together in a population.

Maybe this time we want to know how height affects peanut butter and jelly preference. In other words, this time we'll survey all the students in our school, asking them whether they prefer peanut butter or jelly, and record this information along with their height.

We're looking to see how much height and peanut butter/jelly preference are correlated, if at all.

Keep in mind that even if we found that peanut butter/jelly preference and height were positively correlated, such that taller students were more likely to prefer peanut butter, and shorter students were the more likely to prefer jelly, we could only show correlation, not causation.

Two variables are **correlated** when they move together predictably. The variables are **positively correlated** when they increase together or decrease together. Variables are **negatively correlated** when they increase and decrease in opposite directions: one goes down while the other goes up, or one goes up while the other goes down.

On the other hand, **causation** means that one variable *causes* another variable to change. But just because we show correlation does not mean that we've proven causation.



For example, even if height and peanut butter/jelly preference are correlated, we don't know if being taller causes someone to like peanut butter more, or if liking peanut butter more causes someone to be tall. We don't know which variable causes which. Nor do we know if there's a **confounding variable**, which is a third variable that leads to both of the variables that were correlated. For example, being male might cause someone to be both taller and to prefer peanut butter.

## Experimental studies

In an **experiment**, we're manipulating what's happening, and trying to establish causality, not just correlation.

To run an experiment, we assign people into at least two different groups, hopefully using good random sampling techniques, so that our groups aren't biased in some way.

One group acts as the **control group**, which is the group that does nothing, receives nothing, or isn't manipulated, and the other is the **treatment group** (also called the experimental group), which is the group that does something, receives something, or is treated in some way. The classic example of this is in medical studies, where the treatment group receives some kind of new drug, and the control group receives a **placebo**, or sugar pill.

In an experiment, we're looking to see whether one or more **explanatory variables** (the treatment) has an effect on the **response variable** (whatever is expected to be effected). If we're testing to see whether a new drug decreases blood pressure, the new drug would be the explanatory

variable (the thing that explains the change), and blood pressure would be the response variable (the thing that might decrease as a result of the drug).

Even if our experiment shows a change in the response variable, we still may need to be skeptical of our conclusion. Did we run a good experiment? Could the results have been biased in some way? Was the effect we saw simply due to random chance or the placebo effect?

There are other things we can do to make our experiment more reliable. For example, we could make our experiment blind or double-blind. A **blind experiment** is when the participants don't know whether they're in the control group or the treatment group. A **double-blind experiment** is when neither the participants nor the people administering the experiment know which group anyone is in.

### Matched pairs

When researchers separate participants into like groups, it's called **blocking**. For example, researchers might choose to block on gender by randomly selecting an equal number of men and women, instead of a truly random sample in which the number of men and women isn't controlled.

If they then treat half of the men and half of the women with the drug, and give a placebo to the other half of the men and the other half of the women, the blocking on gender helps them to see if the drug effects men and women differently.



A **matched pairs experiment** is a more specific kind of blocking where we make sure that the participants in our experimental group and control group are matched based on similar characteristics.

Maybe these researchers want to see how both gender and age change the effect of the blood pressure drug. They could match the ages and genders in the control group with the ages and genders in the experimental group. For example, they could put one 18-year-old woman in the treatment group, and put her matched pair (another 18-year-old woman) in the control group.

A matched pairs experiment design is an improvement over a completely randomized design, because participants are still randomly assigned into the treatment and control groups, but potentially confounding variables, like age and gender, are controlled for.

## Replication

We also want to make sure that other people can replicate our experiment. If other people can run the same experiment in the same way, and they get the same results that we do, that provides more evidence that our results are legitimate.



# Sampling and bias

No matter what kind of study we're doing (observational or experimental), we always want to make sure that the subjects in our study are picked randomly if we're using a sample.

Remember that a population  $N$  always consists of the entire group of subjects we're interested in. A sample  $n$ , on the other hand, is a smaller group within the larger population. So if we're able to study the entire population, then we don't have to worry about the group we look at, because we're looking at everybody, or everything. But if we're only going to be using a sample to represent a larger population, then how we pick the subjects that will be in our sample is very important.

In a perfect world, we want the sample to be representative of the population. If it were in fact perfectly representative, we might call it a **representative sample**, because the information we collect about the sample would “scale up” to the population.

For example, we might want to know how many strawberries are in our fruit salad. The fruit salad fills a 20-cup bowl, so we don't want to pick through the entire thing and count every strawberry. Instead, we'll take a 1-cup scoop, and hope that it's a representative sample of the entire bowl.

In the 1-cup scoop, we count 2 strawberries. If the sample (the cup) perfectly represents the population (the entire bowl), we could “scale up” the strawberry count for 1 cup to 20 cups to get a count of the number of strawberries in the entire bowl.



$$\frac{2 \text{ strawberries in our sample}}{1 \text{ cup in our sample}} = \frac{x \text{ strawberries in the bowl}}{20 \text{ cups in the bowl}}$$

$$\frac{2}{1} = \frac{x}{20}$$

$$2 = \frac{x}{20}$$

$$40 = x$$

Therefore, based on the sample, we guess that there are 40 strawberries in the whole bowl. If there are in fact 40 strawberries in the whole bowl, then the single 1-cup scoop would have been a perfectly representative sample of the entire fruit salad.

But this won't always be the case. Maybe the fruit salad wasn't perfectly mixed, and a lot of the strawberries had sunk to the bottom. If we knew that there were actually 100 strawberries in the bowl, then the sample we picked wasn't a very good one.

Many times a sample won't do a good job representing the population because of bias.

## Bias in sampling

Bias, by definition, is showing favor toward something over something else. When we talk about **bias** in statistics, we're basically talking about something that skews our results and makes them inaccurate.



When we collect data for a sample, and we're using that sample to represent the population, we mentioned before that we want a representative sample. We also call this an **unbiased sample**.

To get a representative, or unbiased, sample, we try to avoid introducing bias into the data. Unfortunately, it's really easy to introduce all different kinds of bias into a data set and skew our results:

### Response bias

**Measurement bias:** There's something wrong with the tool we're using to collect the data, so our method of collecting observations or responses from the sample results in false values. For example, if we calibrated a scale improperly before taking measurements, then all the results would suffer from measurement bias.

**Social desirability bias:** If our survey asks "Have you ever stolen something?" people may not answer truthfully. If anyone participating in our survey lies when they answer this question, then we have some social desirability bias in our data. This is similar to measurement bias because in both measurement and social desirability bias, there's something wrong with the tool we're using to collect data.

**Leading questions:** Leading questions are questions that are framed in a way that push respondents toward a particular response. For example, if we ask "Are you more likely to purchase Coca-Cola?" it may cause respondents to answer differently than if we'd simply asked "Are you more likely to purchase another cola brand?" because we're leading them specifically toward Coca-Cola.



## Undercoverage bias

**Selection bias:** Also called undercoverage, this is when we don't collect data from an entire group of subjects that should have been included in our data. For example, let's say I own a daycare center and want to find out the mean household income of the families whose children I look after. If I watch 20 children, and I choose to sample only the parents who pick up their kids before 5 : 00 p.m., I might be dramatically skewing my data. What if the parents who work later have significantly higher incomes, because anyone who picks up their child before 5 : 00 only works part time? I'm not representing all the parents who work late, so that part of the population is under-represented.

**Voluntary response sampling:** This is when people voluntarily respond to my survey or participate in my study, which means that voluntary response sampling can be a cause of selection bias. People who voluntarily participate may have different habits, tendencies, opinions, or backgrounds than people who tend not to participate. So the data we collect from a sample of voluntary respondents may be biased.

**Convenience sampling:** This is when we choose a sample simply because it's convenient, not because we're trying to get a good, random representative sample. Therefore, this can be another cause of selection bias. There's almost always some aspect of convenience to sampling, but a good example would be if we're trying to collect data about the people in our city, and we just ask the neighbors who live on our street. It's really convenient to collect



data for our street only, but it certainly doesn't give us an unbiased sample for the entire city, so this convenience sample may cause a big problem.

## Non-response bias

**Non-response bias:** This is when we get a large number of people who don't respond to our survey. There may be bias in our data because we didn't collect answers from everyone who didn't respond, and we don't know what they may have said. For example, state representatives often send surveys to all of their constituents to ask them how much they care about different political issues. If they only get a response rate of 5%, that means 95% of constituents didn't bother to send back the survey. Which means the representative only collected opinions from 5% of the population they were interested in, so nonresponse bias may be a big issue.

## Direction of bias

Based on the bias that we suspect may exist in our sample (response bias, undercoverage, and non-response bias), we always want to be able to make an educated guess about whether our results are more likely producing an overestimate or an underestimate.

For example, in the response bias section, we talked about the survey question “Have you ever stolen something?” If that question exists in our survey, it could actually lead to response bias or undercoverage, depending on whether people choose to skip the question or lie.



Or in the undercoverage section, we talked about surveying only the parents who pick up their children from daycare before 5 : 00 p.m. We might admit that we have some undercoverage in our data, and would guess that our estimate for household income is low, since parents who work late might make more money than those who get off work earlier.

Either way, we always want to be thinking about the kind of bias we're introducing, and whether the bias in our data has caused us to overestimate or underestimate the value we're looking at.

## Sampling techniques

So how can we avoid bias in our sample data? Well, there are few techniques we can use to divide subjects into groups that help ensure that our data stays as random (unbiased) as possible.

First, we could assign a number to each subject in the population, then pick numbers out of a hat, assigning every other subject to the same group. Or, we could use a random number generator on a computer to do the same thing. We could also get a computer generated string of digits, and then pick out numbers in order from the random number string.

When we assign subjects to groups in a totally random way like this, we call it a **simple random sample**. But even if we assign subjects to groups in a totally random way, we can still end up with a skewed sample. For example, it's possible that we could use a random number generator on a computer to randomly put 50 men and 50 women into two different groups, and yet end up with 40 men and 10 women in one group, and 10 men and 40 women in another group.



To fix problems like this, instead of doing a **simple random sample**, we can try to take a **stratified random sample**, where we put some parameter on the sample where we require an even number of subjects from different groups. For example, if we want to have one group of 25 men and one group of 25 women in our sample, then we'll treat men as one population and women as another. Then we'll take a random sample of 25 men from the male population and a random sample of 25 women from the female population. These two groups (men and women) are called the **strata** of the stratified random sample.

Earlier we used the term blocking to describe this. These concepts are the same thing. The strata in sampling are what we called the blocks in an experiment with a randomized block design.

We could also take a **clustered random sample**, where we break our population into clusters, and then either 1) take a random sample within each cluster to be our total sample, or 2) randomly pick some clusters and then sample everyone in those clusters.

In a cluster sample we want each cluster to be similar to the population as a whole. For example, say we have a nicely mixed fruit salad that's been divided into 12 portions. Then each of these portions is a cluster. Once the population is divided into representative clusters you can take a simple random sample of clusters, say 3 out of the 12 portions to analyze the sample.

In the case of a stratified sample, the fruit salad would've had to be separated back into strawberries, bananas, watermelon etc., and a sample from each group selected. So often we will hear a stratified sample is the



same *within groups*, (like each fruit is the same) and a cluster sample is the same *between groups*, (like each portion of fruit salad should be representatively the same).

Finally, we could use **systematic sampling**, which is really similar to simple random sampling. The difference is that we assign numbers to individuals in a population and choose them at some specified interval. For example, we could list all students in a school in alphabetical order, choose student #5 as our starting point, and then select every 10th student on the list from there.



# Sampling distribution of the sample mean

We already know how to find parameters that describe a population, like mean, variance, and standard deviation. But we also know that finding these values for a population can be difficult or impossible, because it's not usually easy to collect data for every single subject in a large population.

So, instead of collecting data for the entire population, we choose a subset of the population and call it a “sample.” We say that the larger population has  $N$  subjects, but the smaller sample has  $n$  subjects.

In the same way that we'd find parameters for the population, we can find statistics for the sample. Then, based on the statistic for the sample, we can infer that the population parameter might be similar to its corresponding sample statistic.

## Sampling distribution of the sample mean

Consider the fact though that pulling one sample from a population could produce a statistic that isn't a good estimator of the corresponding population parameter.

For example, maybe the mean height of girls in our statistics class is 65 inches. Let's say there are 30 girls in the class, and we take a sample of 3 of them. If we happened to pick the three tallest girls, then the mean of that sample wouldn't be a good estimate of the population mean, because the mean height from the sample would be significantly higher than the mean



height of the population. Similarly, if we instead happened to choose the three shortest girls, the sample mean would be much lower than the actual population mean.

So how do we correct for this? How do we adjust for the fact that individual samples might produce sample statistics that are bad estimates of their corresponding population parameters?

Well, instead of taking just one sample from the population, think about what happens when we take every possible sample of 3 girls from the population of girls in our class. The total number all of possible samples is  $N^n$ , where  $N$  is the total population from which we take our samples, and  $n$  is the sample size.

So if we take a sample of 3 girls from a population of 30 girls, the total number of possible samples is

$$N^n = 30^3 = 27,000$$

Keep in mind that, in this scenario, we're **sampling with replacement**, which means we pick a random sample of three girls, and then "put them back" into the population and pick another random sample of three girls. We'd keep doing this over and over until we've taken every unique 3-girl sample.

What we want to realize is that every one of these samples has its own mean, so now we have a data set of 27,000 sample means. Here's the key point: This set of sample means actually forms its own distribution around the real population mean. In other words, if we look at these means as a probability distribution, it turns out that this probability distribution of



sample means is always normal (as long as we're taking large enough samples, more on this later), and this normal distribution is called the **sampling distribution of the sample mean (SDSM)**.

Just think about the sampling distribution of the sample mean as the probability distribution of all possible values of the sample mean  $\bar{x}$ .

Because the sampling distribution of the sample mean is normal (assuming the original population was normal, and/or we used a large enough sample size  $n \geq 30$ ), we can of course find a mean and standard deviation for the distribution, and therefore answer probability questions about it.

## Central Limit Theorem

We just said that the sampling distribution of the sample mean is normal, but let's clarify. In actuality,

- If the original population is normally distributed, then the SDSM will also be normally distributed, regardless of the sample size  $n$  that we use.
- If the original population is not normally distributed, or if we don't know the shape of the population distribution, then the SDSM is only guaranteed to be normally distributed when we use a sample size of at least  $n = 30$ .

This conclusion about the normality of the SDSM is the Central Limit Theorem. In reality, many populations don't follow a normal distribution, meaning that they don't approximate the bell-shaped-curve of a normal



distribution. Real-life distributions are all over the place, because real-life phenomena don't always follow a perfectly normal distribution.

The **Central Limit Theorem (CLT)** is how we turn a non-normal population into a normal SDSM. It tells us that, even if a population distribution is non-normal, as long as we use a large enough sample ( $n \geq 30$ ), that we can make inferences about our sample statistics, because of the fact that the SDSM will be a normal distribution.

So the Central Limit Theorem is useful because it lets us apply what we know about normal distributions, like the properties of mean, variance, and standard deviation, to non-normal populations.

## Mean, variance, and standard deviation

The Central Limit Theorem also states that the mean of the sampling distribution of the sample mean will always be the same as the mean of the original distribution.

$$\mu_{\bar{x}} = \mu$$

If the population is infinite, or if we're sampling with replacement, then the variance of the SDSM is equal to the population variance divided by the sample size.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

If population variance is unknown (which will almost always be the case), then we can use sample variance as an estimate of population variance.



$$s_{\bar{x}}^2 \approx \frac{s^2}{n}$$

The standard deviation of the sampling distribution, also called the **standard error**, the standard deviation of sample means, or the standard error of the mean, is therefore given by

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is population standard deviation and  $n$  is sample size. When we don't know population standard deviation, we can use sample standard deviation as an estimate of population standard deviation.

$$SE = s_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

The standard error tells us how accurate the mean of any given sample is likely to be as an estimate of the actual population mean. “Error” doesn’t mean there’s been a mistake, it just refers to the distance between any particular sample mean and the mean of the population.

When the standard error is larger, it indicates that the sample means in the SDSM are more spread out, so it’s less likely that any given sample mean is an accurate representation of the true population mean. But when the standard error is smaller, the sample means are less spread out, so it’s more likely that any given sample mean is an accurate representation of the true population mean.

- The standard error will be larger when population standard deviation is larger, and/or when the sample size is smaller.



- The standard error will be smaller when population standard deviation is smaller, and/or when the sample size is larger.

## Finite population correction factor

We've already said that the standard deviation of the sampling distribution of the sample mean (standard error) is given by  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ . In fact, this formula leaves something out. The complete formula is actually

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

This extra  $\sqrt{(N-n)/(N-1)}$  is what we call the **finite population correction factor (FPC)**. When the population we're sampling from is

- infinite, or
- when we're sampling with replacement, or
- when the population is finite but large in comparison to a smaller sample (the sample size is less than or equal to 5% of the population,  $n/N \leq 0.05$ ),

then the value of the FPC is 1 or very close to 1. And when the FPC's value is 1 or very close to 1, it'll have no or little impact on the value of the standard error, which is when we can simplify the standard error formula to just  $SE = \sigma_{\bar{x}} = \sigma/\sqrt{n}$ .

In other words, as long as the population is infinite, or we're sampling with replacement, or we're sampling from no more than 5% of a finite



population, then we can use the simplified formula  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$  for standard error. Otherwise, if we're sampling without replacement or sampling from more than 5 % of a finite population, we should use

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

And of course, given this formula for the standard error, we know that the variance of the SDSM is given by

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right)$$

Keep in mind that there's some debate among statisticians and statistics textbooks about whether the FPC should be applied only when  $n/N > 0.10$ , instead of using the stricter threshold  $n/N > 0.05$ . For our purposes, we'll stick to using the  $n/N > 0.05$  rule.

Let's do an example so that we can see how these formulas work.

### Example

A group of 4 people have the following weights in pounds: 150, 156, 158, 164. Find all possible random samples of size 2 if we're sampling with replacement. Then find the sample mean for every sample. Determine the probability distribution of the sample mean, the mean of the SDSM  $\mu_{\bar{x}}$ , and the standard error  $\sigma_{\bar{x}}$ .



Let's first determine the total number of possible samples, using  $N^n$ , given  $N = 4$  and  $n = 2$ .

$$N^n = 4^2 = 16$$

The complete sample space, and the mean for each sample, is

Sample	Sample mean
150, 150	150
150, 156	153
150, 158	154
150, 164	157
156, 150	153
156, 156	156
156, 158	157
156, 164	160
158, 150	154
158, 156	157
158, 158	158
158, 164	161
164, 150	157
164, 156	160
164, 158	161
164, 164	164

Build a table for the probability distribution of the sample mean. Because there are 16 total samples, the probability of each sample mean will be



given by the number of times that sample mean occurs, divided by the total number of possible samples, so “count/16.”

Sample mean	$P(x_i)$
150	1/16
153	2/16
154	2/16
156	1/16
157	4/16
158	1/16
160	2/16
161	2/16
164	1/16

Now we can calculate the mean of the sampling distribution of the sample mean,  $\mu_{\bar{x}}$ , where  $\bar{x}_i$  is a given sample mean,  $P(\bar{x}_i)$  is the probability of that particular sample mean occurring, and  $N$  is the number of samples.

$$\mu_{\bar{x}} = \sum_{i=1}^N \bar{x}_i P(\bar{x}_i)$$

$$\mu_{\bar{x}} = 150 \left( \frac{1}{16} \right) + 153 \left( \frac{2}{16} \right) + 154 \left( \frac{2}{16} \right) + 156 \left( \frac{1}{16} \right) + 157 \left( \frac{4}{16} \right)$$

$$+ 158 \left( \frac{1}{16} \right) + 160 \left( \frac{2}{16} \right) + 161 \left( \frac{2}{16} \right) + 164 \left( \frac{1}{16} \right)$$

$$\mu_{\bar{x}} = \frac{2,512}{16}$$

$$\mu_{\bar{x}} = 157$$

Because we're sampling with replacement, we would expect this mean of the SDSM to be equivalent to the mean of the population,  $\mu_{\bar{x}} = \mu$ , and we can see that it is if we calculate the mean of the population.

$$\mu = \frac{150 + 156 + 158 + 164}{4} = \frac{628}{4} = 157$$

Both means are  $\mu_{\bar{x}} = \mu = 157$ . The population variance is

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{(150 - 157)^2 + (156 - 157)^2 + (158 - 157)^2 + (164 - 157)^2}{4}$$

$$\sigma^2 = \frac{(-7)^2 + (-1)^2 + 1^2 + 7^2}{4}$$

$$\sigma^2 = \frac{49 + 1 + 1 + 49}{4}$$

$$\sigma^2 = \frac{100}{4}$$

$$\sigma^2 = 25$$

which means that the population standard deviation is

$$\sigma = \sqrt{25}$$

$$\sigma = 5$$

Because we're sampling with replacement, we can use the simplified formula for standard error (the one *without* the FPC).

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{5}{\sqrt{2}}$$

$$\sigma_{\bar{x}} \approx 3.54$$

Instead of calculating population variance and standard deviation, and using those values to find standard error, we could have calculated the variance of the SDSM directly,

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^N (\bar{x}_i - \mu)^2 P(\bar{x}_i)$$

$$\sigma_{\bar{x}}^2 = (150 - 157)^2 \left( \frac{1}{16} \right) + (153 - 157)^2 \left( \frac{2}{16} \right) + (154 - 157)^2 \left( \frac{2}{16} \right)$$

$$+ (156 - 157)^2 \left( \frac{1}{16} \right) + (157 - 157)^2 \left( \frac{4}{16} \right) + (158 - 157)^2 \left( \frac{1}{16} \right)$$

$$+ (160 - 157)^2 \left( \frac{2}{16} \right) + (161 - 157)^2 \left( \frac{2}{16} \right) + (164 - 157)^2 \left( \frac{1}{16} \right)$$

$$\sigma_{\bar{x}}^2 = (-7)^2 \left( \frac{1}{16} \right) + (-4)^2 \left( \frac{2}{16} \right) + (-3)^2 \left( \frac{2}{16} \right) + (-1)^2 \left( \frac{1}{16} \right)$$

$$+ 1^2 \left( \frac{1}{16} \right) + 3^2 \left( \frac{2}{16} \right) + 4^2 \left( \frac{2}{16} \right) + 7^2 \left( \frac{1}{16} \right)$$

$$\sigma_{\bar{x}}^2 = \frac{49}{16} + \frac{32}{16} + \frac{18}{16} + \frac{1}{16} + \frac{1}{16} + \frac{18}{16} + \frac{32}{16} + \frac{49}{16}$$

$$\sigma_{\bar{x}}^2 = \frac{200}{16}$$

$$\sigma_{\bar{x}}^2 = \frac{25}{2}$$

$$\sigma_{\bar{x}}^2 = 12.5$$

and then found standard error directly.

$$\sigma_{\bar{x}} = \sqrt{12.5}$$

$$\sigma_{\bar{x}} \approx 3.54$$

# Conditions for inference with the SDSM

Now that we've defined the sampling distribution for the sample mean and the Central Limit Theorem, we want to be able to use it to make inferences about the population. After all, the whole point of statistics is being able to use samples to make educated guesses about the population.

So, even when we don't know the population mean, our goal now will be to take a sample, find the mean of the sample, and then compare the sample mean to the mean of the SDSM. That comparison will allow us to draw conclusions about the population mean.

But, in order for this process to be valid, there are always three conditions we need to meet when we're sampling: the sample needs to be random, it needs to be large enough for the Central Limit Theorem to kick in and ensure normality, and our samples need to be independent.

## Random

Any sample we take needs to be a simple random sample. Often we'll be told in the problem that sampling was random.

## Normal (large counts)

We need to know that our sample size is large enough. In the case of the sampling distribution of the sample mean, 30 is a magic number for the sample size we need to use to make the sampling distribution normal. In other words, the sample size needs to be at least 30 in order for the CLT to



create a normal SDSM for a non-normally distributed population (unless we're sampling with replacement, in which case we can get away with a sample size smaller than 30).

If the original population is normally distributed, then this rule doesn't apply because the sampling distribution will also be normal, even if our sample size is less than 30.

If our population is finite, and we're sampling without replacement or taking a sample larger than 5% of the population, then we just have to remember to use the finite population correction factor that we talked about earlier.

## Independent (10% rule)

If we're sampling with replacement, then the 10% rule tells us that we can assume the independence of our samples. But if we're sampling without replacement (we're not "putting our subjects back" into the population every time we take a new sample), then we need to keep our sample size below 10% of the total population.

For example, if the original population is 2,000 subjects, we need to make sure that the sample size is no more than 200 subjects so that we stay under the  $200/2,000 = 1/10 = 10\%$  threshold.

In other words, as long as we keep the sample size to 10% or less of the total population, we can "get away with" a sample that isn't truly independent (we can get away with sampling without replacement),

because this 10% threshold is small enough to approximate an independent sample.

If our sample meets these conditions, then we can use the sampling distribution of the sample mean to answer questions about the probability that any given sample mean will fall within some distance of the population mean.

For these types of problems, we'll need to use

$$z_{\bar{x}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

to determine  $z$ -scores that apply to the sampling distribution of the sample mean. Let's do an example.

### Example

A sports equipment company produces approximately 1,000,000 soccer balls per year, and the pressure in the soccer balls is normally distributed with a mean of 8.7 PSI (pounds per square inch), and a standard deviation of 0.4 PSI. They randomly select 25 soccer balls (without replacement) to check their pressure. Find the probability that the sample mean  $\bar{x}$  is within 0.2 PSI of the population mean.

Before we can answer this probability question, we need to check our conditions for inference. We were told in the problem that the sample was taken randomly, so we can assume we've met the “random” condition.



We were also told that the PSI in the population of soccer balls is normally distributed, so our SDSM will also be normal, even though our sample size is smaller than 30, and we've therefore met the “normal” condition.

We're sampling without replacement, which means our sample needs to be at most 10% of the population, but 25 soccer balls is a significantly smaller sample than 10% of the population, so we've met the “independent” condition.

To answer the probability question, we'll start by finding the mean of the SDSM. The Central Limit Theorem tells us that it'll be equal to the population mean, so  $\mu_{\bar{x}} = 8.7$ . The standard error will be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{0.4}{\sqrt{25}}$$

$$\sigma_{\bar{x}} = \frac{0.4}{5}$$

$$\sigma_{\bar{x}} = 0.08$$

Remember here that we're sampling without replacement from a finite population, but our sample size is significantly smaller than 5% of the population, so we didn't need to apply the finite population correction factor when we found the standard error.



We want to know the probability that the sample mean  $\bar{x}$  is within 0.2 PSI of the population mean, 8.7. A 0.2 interval around 8.7 gives us the interval 8.5 to 8.9, so

$$P(8.5 < \bar{x} < 8.9) = P\left(\frac{8.5 - 8.7}{0.08} < z_{\bar{x}} < \frac{8.9 - 8.7}{0.08}\right)$$

$$P(8.5 < \bar{x} < 8.9) = P\left(\frac{-0.2}{0.08} < z_{\bar{x}} < \frac{0.2}{0.08}\right)$$

$$P(8.5 < \bar{x} < 8.9) = P(-2.50 < z_{\bar{x}} < 2.50)$$

In the  $z$ -table, a  $z$ -value of 2.50 gives 0.9938,

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	<b>.9938</b>	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964

and a  $z$ -value of  $-2.50$  gives 0.0062.

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	<b>.0062</b>	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

Which means the probability under the normal curve between these  $z$ -scores is

$$P(-2.50 < z_{\bar{x}} < 2.50) = 0.9938 - 0.0062$$

$$P(-2.50 < z_{\bar{x}} < 2.50) = 0.9876$$

$$P(-2.50 < z_{\bar{x}} < 2.50) \approx 98.8\%$$

So there's an approximately 98.8 % chance that the mean  $\bar{x}$  of the 25-ball sample the company takes will fall within 0.2 PSI of the population mean of  $\mu = 8.7$  PSI.

---



# Sampling distribution of the sample proportion

In the same way that we were able to find a sampling distribution for the sample mean (SDSM), we can find a sampling distribution for the sample proportion (SDSP).

In other words, if we're dealing with a population proportion  $p$ , instead of a population mean  $\mu$ , then we'll be trying to create a sampling distribution of the sample proportion, as opposed to a sampling distribution of the sample mean.

## Sampling distribution of the sample proportion

Often we'll want to calculate a **population proportion**  $p$ , which is the number of subjects in our population that meet a certain condition.

For example, maybe we want to know how many students in our school have brown hair. If there are 5,000 students who attend our school, it might not be possible to survey everybody. So instead we could take a random sample of 100 students and see how many of them have brown hair. This is the **sample proportion**, since it's the proportion of students in the sample with brown hair, which is given by

$$\hat{p} = \frac{x}{n}$$

where  $\hat{p}$  is the sample proportion,  $x$  is the number of students in the sample with brown hair (the number of “successes”), and  $n$  is the sample size (we surveyed 100 students for our sample).



Just like for the SDSM, the **sampling distribution of the sample proportion (SDSP)** is created when we take every possible sample from our population, calculate the sample proportion for each sample, and then plot all of those sample proportions into a probability distribution.

In other words, the SDSP is the probability distribution of all possible sample proportions  $\hat{p}$ .

## Central Limit Theorem

Remember that, whenever we're dealing with a proportion, the distribution is a binomial distribution, since every outcome is either a “success” or a “failure.” So in the case of a population proportion, the original population will be modeled by a binomial distribution, not a normal distribution.

But even though the population follows a binomial distribution, we can still use the Central Limit Theorem to create a sampling distribution of the sample proportion. Just like the SDSM, the CLT tells us that the SDSP is only guaranteed to be normally distributed when we use a sample size of at least  $n = 30$ .

## Mean, variance, and standard deviation

The mean of the sampling distribution of the sample proportion  $\mu_{\hat{p}}$  will be equal to the population proportion  $p$ .

$$\mu_{\hat{p}} = p$$

The standard deviation of the sampling distribution of the sample proportion  $\sigma_{\hat{p}}$ , also called the **standard error of the proportion**, will be

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

where  $p$  is the population proportion and  $n$  is the sample size. We use this formula for standard error of the proportion if our population is infinite, or if the population is finite but large in comparison to our sample size (if sample size is no more than 5% of the population,  $n/N \leq 0.05$ ).

Of course, based on this formula for standard error, we can say that the variance of the sampling distribution of the sample proportion is

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

If the population proportion is unknown, we estimate the population proportion  $p$  using the next best thing, the sample proportion  $\hat{p}$ , and the formulas for standard error and variance become

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\sigma_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n}$$

## Finite population correction factor



In the case where the population is finite and  $n/N > 0.05$ , we have to apply the finite population correction factor, and in that case the correct formula for the standard error of the proportion is then

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$$

where  $p$  is the population proportion,  $n$  is the sample size, and  $N$  is the size of the population. If we're applying the FPC, then the formula for variance of the SDSP is

$$\sigma_{\hat{p}}^2 = \left( \frac{p(1-p)}{n} \right) \left( \frac{N-n}{N-1} \right)$$

And again, if the population proportion is unknown, we approximate it with the sample proportion, and our formulas for standard error and variance with the FPC are

$$SE = \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\hat{p}}^2 = \left( \frac{\hat{p}(1-\hat{p})}{n} \right) \left( \frac{N-n}{N-1} \right)$$

Let's do an example so that we can see how these formulas work.

### Example

A group of 4 people have the following hair color: brown, brown, brown, blonde. Find all possible random samples of size 2 if we're sampling with replacement. If we define brown hair as a “success,” then find the sample



proportion for every sample. Determine the probability distribution of the sample proportion, the mean of the SDSP  $\hat{p}$ , and the standard error  $\sigma_{\hat{p}}$ .

Let's first determine the total number of possible samples, using  $N^n$ , given  $N = 4$  and  $n = 2$ .

$$N^n = 4^2 = 16$$

The complete sample space, and the proportion for each sample, is



Sample	Sample proportion
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
brown, brown	1
brown, brown	1
brown, brown	1
brown, blonde	1/2
blonde, brown	1/2
blonde, brown	1/2
blonde, brown	1/2
blonde, blonde	0

Build a table for the probability distribution of the sample proportion.

Because there are 16 total samples, the probability of each sample proportion will be given by the number of times that sample proportion occurs, divided by the total number of possible samples, so “count/16.”

Sample proportion	P( $p_i$ )
0	1/16
1/2	6/16
1	9/16

Now we can calculate the mean of the sampling distribution of the sample proportion,  $\mu_{\hat{p}}$ , where  $\hat{p}_i$  is a given sample proportion,  $P(\hat{p}_i)$  is the probability of that particular sample proportion occurring, and  $N$  is the number of samples.

$$\mu_{\hat{p}} = \sum_{i=1}^N \hat{p}_i P(\hat{p}_i)$$

$$\mu_{\hat{p}} = 0 \left( \frac{1}{16} \right) + \frac{1}{2} \left( \frac{6}{16} \right) + 1 \left( \frac{9}{16} \right)$$

$$\mu_{\hat{p}} = \frac{3}{16} + \frac{9}{16}$$

$$\mu_{\hat{p}} = \frac{12}{16}$$

$$\mu_{\hat{p}} = \frac{3}{4}$$

Because we're sampling with replacement, we would expect this mean of the SDSP to be equivalent to the population proportion,  $\mu_{\hat{p}} = p$ , and we can see that it is if we calculate the population proportion.

$$p = \frac{3 \text{ people with brown hair}}{4 \text{ people in the population}} = \frac{3}{4}$$



Both proportions are  $\mu_{\hat{p}} = p = 3/4$ . The variance of the SDSP would be

$$\sigma_{\hat{p}}^2 = \sum_{i=1}^N (\hat{p}_i - p)^2 P(\hat{p}_i)$$

$$\sigma_{\hat{p}}^2 = \left(0 - \frac{3}{4}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{1}{2} - \frac{3}{4}\right)^2 \left(\frac{6}{16}\right) + \left(1 - \frac{3}{4}\right)^2 \left(\frac{9}{16}\right)$$

$$\sigma_{\hat{p}}^2 = \left(-\frac{3}{4}\right)^2 \left(\frac{1}{16}\right) + \left(-\frac{1}{4}\right)^2 \left(\frac{6}{16}\right) + \left(\frac{1}{4}\right)^2 \left(\frac{9}{16}\right)$$

$$\sigma_{\hat{p}}^2 = \frac{9}{16} \left(\frac{1}{16}\right) + \frac{1}{16} \left(\frac{6}{16}\right) + \frac{1}{16} \left(\frac{9}{16}\right)$$

$$\sigma_{\hat{p}}^2 = \frac{9}{256} + \frac{6}{256} + \frac{9}{256}$$

$$\sigma_{\hat{p}}^2 = \frac{24}{256}$$

$$\sigma_{\hat{p}}^2 = \frac{3}{32}$$

and then the standard error would be

$$\sigma_{\hat{p}} = \sqrt{\frac{3}{32}}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{3}}{4\sqrt{2}}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{6}}{8}$$

$$\sigma_{\hat{p}} \approx 0.31$$

---



# Conditions for inference with the SDSP

Just like we did with the sampling distribution of the sample mean, we have to meet specific sampling conditions in order to be able to use the sampling distribution of the sample proportion to make inferences about the population proportion.

The conditions for inference that apply to the sampling distribution of the sample proportion are similar to the conditions we applied to the sampling distribution of the sample mean.

## Random

Any sample we take needs to be a simple random sample. Often we'll be told in the problem that sampling was random.

## Normal (large counts)

The sampling distribution of the sample proportion can only be guaranteed to be normal if  $np \geq 5$  and  $n(1 - p) \geq 5$ , where  $n$  is the sample size and  $p$  is the population proportion. If  $np \geq 5$  is true, it tells us that we can expect to have at least 5 “successes” in the sample, and if  $n(1 - p) \geq 5$  is true, it tells us that we can expect to have at least 5 “failures.”

So if our sample size is  $n = 100$  and the population proportion is  $p = 60\%$ , then we multiply 100 by 0.6 and by  $1 - 0.6 = 0.4$  to make sure both values are at least 5.



$$100 \cdot 0.6 = 60 > 5$$

$$100 \cdot 0.4 = 40 > 5$$

Since both values are at least 5, the sampling distribution of the proportion is approximately normal.

If we don't know the population proportion  $p$ , then we use the sample proportion  $\hat{p}$  as its best estimate, and use  $\hat{p}$  to check that our sample has at least 5 "successes" and at least 5 "failures."

Note that there's some debate among statisticians and statistics textbooks about whether the large counts condition should be  $np \geq 10$  and  $n(1 - p) \geq 10$  or  $np \geq 5$  and  $n(1 - p) \geq 5$ . For our purposes, we'll stick to using the  $np \geq 5$  and  $n(1 - p) \geq 5$  rule.

## Independent (10 % rule)

If we're sampling with replacement, then the 10 % rule tells us that we can assume the independence of our samples. But if we're sampling without replacement (we're not "putting our subjects back" into the population every time we take a new sample), then we need keep our sample size below 10 % of the total population.

If our sample meets these conditions, then we can use the sampling distribution of the sample proportion to answer questions about the probability that any given sample proportion will fall within some distance of the population proportion.

For these types of problems, we'll need to use



$$z_{\hat{p}} = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

to determine  $z$ -scores that apply to the sampling distribution of the sample proportion. Let's work through an example.

### Example

An ice cream shop claims that 40% of their 1,000 customers order their ice cream in a waffle cone. We want to verify this claim, so we take a random sample of 90 customers and see whether or not they order a waffle cone.

What is the probability that our results are within 5% of the ice cream shop's 40% claim?

In our case,  $n = 90$  and  $p = 0.4$ , which means  $1 - p = 0.6$ , so

$$np = (90)(0.4) = 36 \geq 5$$

$$n(1 - p) = (90)(1 - 0.4) = (90)(0.6) = 54 \geq 5$$

and we've verified normality. We were told in the question that our sample was random, and our sample is 90 of the total population of 1,000, which means it's  $90/1,000 = 9\%$  of the population, so we're not violating the 10% rule.

With the conditions for inference out of the way, we can calculate the mean and standard deviation of the sampling distribution of the sample proportion. The mean is



$$\mu_{\hat{p}} = p$$

$$\mu_{\hat{p}} = 0.4$$

To find the standard deviation of the sampling distribution, we need to apply the finite population correction factor, given that our population is finite and we're sampling from more than 5 % of the population.

$$\sigma_{\hat{p}} = \sqrt{\frac{0.4(1 - 0.4)}{90}} \sqrt{\frac{1,000 - 90}{1,000 - 1}}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.4(0.6)}{90}} \sqrt{\frac{910}{999}}$$

$$\sigma_{\hat{p}} \approx 0.0516(0.9544)$$

$$\sigma_{\hat{p}} \approx 0.0492$$

The question asks us for the probability that our sample proportion is within 5 % of population proportion  $p = 40\%$ . In other words, how likely is it that the sample proportion falls between 35 % and 45 %?

$$P(0.35 < \hat{p} < 0.45) \approx P\left(\frac{0.35 - 0.4}{0.0492} < z < \frac{0.45 - 0.4}{0.0492}\right)$$

$$P(0.35 < \hat{p} < 0.45) \approx P\left(\frac{-0.05}{0.0492} < z < \frac{0.05}{0.0492}\right)$$

$$P(0.35 < \hat{p} < 0.45) \approx P(-1.02 < z < 1.02)$$

In a  $z$ -table, a  $z$ -value of 1.02 gives 0.8461,



<b>z</b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>0.9</b>	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
<b>1.0</b>	.8413	.8438	<b>.8461</b>	.8485	.8508	.8531	.8554	.8577	.8599	.8621
<b>1.1</b>	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

and a value of  $-1.02$  gives 0.1539.

<b>z</b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>-1.1</b>	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
<b>-1.0</b>	.1587	.1562	<b>.1539</b>	.1515	.1492	.1469	.1446	.1423	.1401	.1379
<b>-0.9</b>	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611

Which means the probability under the normal curve of the sampling distribution of the sample proportion between these  $z$ -scores is

$$P(-1.02 < z < 1.02) \approx 0.8461 - 0.1539$$

$$P(-1.02 < z < 1.02) \approx 0.6922$$

$$P(-1.02 < z < 1.02) \approx 69\%$$

Which means there's an approximately 69% chance that our sample proportion will fall within 5% of the ice cream shop's claim. In other words, approximately 69% of our samples will produce a sample proportion that's within 5% of the population proportion.

# The student's t-distribution

So far we've been working with the normal distribution, which is that perfectly symmetrical bell-shaped distribution with a mean  $\mu$  and a standard deviation  $\sigma$ .

The **standard normal distribution** (also called the  $z$ -distribution) is the normal distribution that specifically has mean  $\mu = 0$  and standard deviation  $\sigma = 1$ , and we've been looking up  $z$ -values for the  $z$ -distribution in the  $z$ -table.

Of course, by the Empirical rule, about 68 % of the area under the standard normal distribution is found between the  $z$ -scores  $z = -1$  and  $z = 1$ , about 95 % of the area is found between  $z = -2$  and  $z = 2$ , and about 99.7 % of the area is found between  $z = -3$  and  $z = 3$ .

Now we want to turn our attention toward the  $t$ -distribution, and  $t$ -scores that we look up in the  $t$ -table.

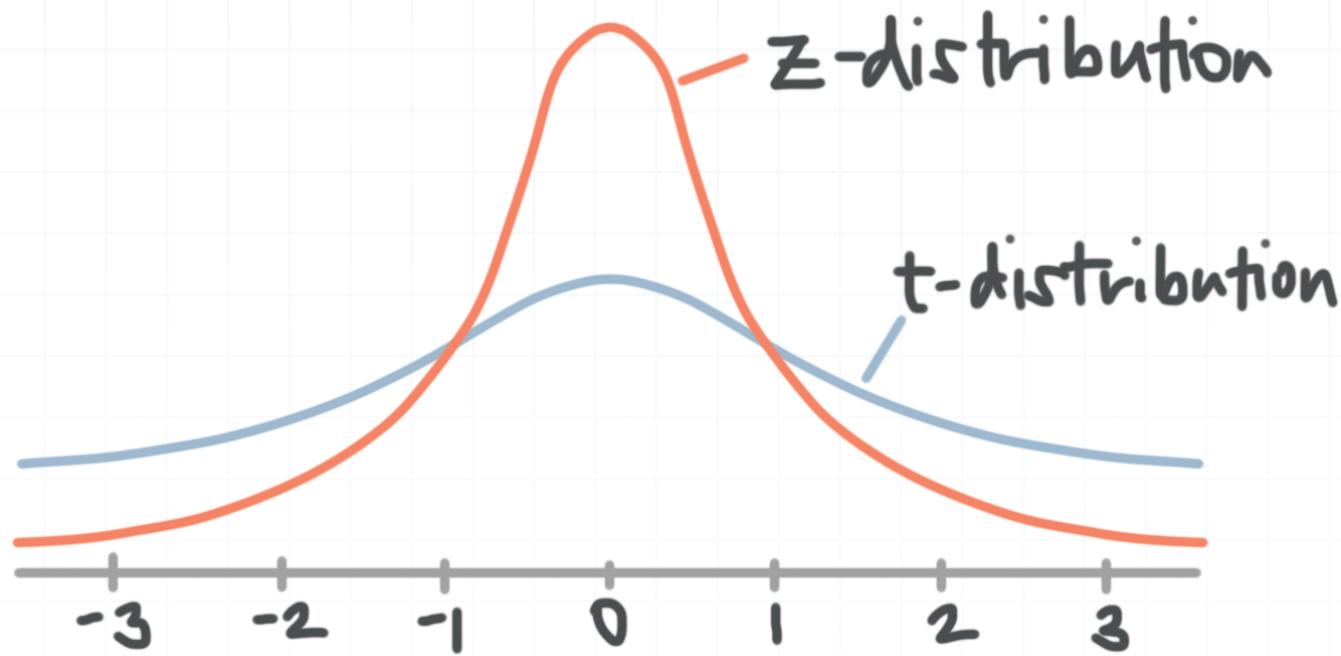
## The student's t-distribution

The **student's  $t$ -distribution** is similar to the standard normal distribution ( $z$ -distribution) in the sense that it's symmetrical, bell-shaped, and centered around the mean  $\mu = 0$ .

But the  $t$ -distribution is flatter and wider than the standard normal distribution, so more of the area under the  $t$ -distribution is pushed out toward the tails. That means that the standard deviation of the  $t$



$t$ -distribution is larger than the standard deviation of the standard normal distribution.



Keep in mind that there isn't one generic  $t$ -distribution. The specific shape of a  $t$ -distribution, and therefore the  $t$ -scores we find in the  $t$ -table, will depend on the number of **degrees of freedom**, which is given by  $n - 1$ , where  $n$  is our sample size.

So the  $t$ -distribution for a sample of size  $n = 10$  (with  $n - 1 = 10 - 1 = 9$  degrees of freedom) will look a little different than the  $t$ -distribution for a sample of size  $n = 20$  (with  $n - 1 = 20 - 1 = 19$  degrees of freedom).

In what way does the  $t$ -distribution for  $n = 10$  look different than the  $t$ -distribution for  $n = 20$ ? Well, regardless of the sample size (and therefore the number of degrees of freedom), the  $t$ -distribution is always normal. But the larger the sample size, the taller and narrower the  $t$ -distribution gets. The smaller the sample size, the shorter and wider the  $t$ -distribution gets.

In other words, as the sample size gets larger, the data becomes more tightly clustered around the mean, and the standard deviation gets smaller.

When the sample size reaches  $n = 30$ , the  $t$ -distribution gets just tall enough that its values very closely approximate the values of the  $z$ -distribution. So for sample sizes  $n \geq 30$ , the values from the  $t$ - and  $z$ -distributions will be almost identical. But for smaller samples  $n < 30$ , the  $t$ -distribution will do a better job estimating probability than the  $z$ -distribution.

Therefore, if we're using a sample size  $n < 30$ , we should calculate a  $t$ -score and look up that  $t$ -score in a  $t$ -table, instead of calculating a  $z$ -score and looking up that  $z$ -score in a  $z$ -table.

$n \geq 30$       Find a  $z$ -score, look it up in the  $z$ -table

$n < 30$       Find a  $t$ -score, look it up in the  $t$ -table

Because the  $z$ - and  $t$ -table return almost identical values when  $n \geq 30$ , such that either the  $z$ - or  $t$ -table could really be used for larger samples, and because we must use the  $t$ -table for smaller samples, some statisticians prefer to simplify things and always use the  $t$ -table, regardless of sample size. But many still choose to use a  $t$ -table for smaller samples and switch to the  $z$ -table for larger samples.

For our purposes, we'll follow this second approach, using a  $t$ -table for smaller samples and a  $z$ -table for larger samples.

There are also other conditions, besides sample size, that dictate when we should use the  $z$ -distribution vs. the  $t$ -distribution. For instance, if



population standard deviation is unknown, we'll always use the *t*-table, regardless of sample size. For example, given a sample of size  $n = 40$ , if population standard deviation is unknown, we'd use a *t*-table.

## Finding values in the *t*-table

Looking up *t*-values in a *t*-table is similar to looking up *z*-values in a *z*-table. To find values in the *t*-table, we need to know degrees of freedom ( $n - 1$ , when the sample size is  $n$ ) and either the upper-tail probability or the confidence level.

We'll talk more later about upper-tail probability and confidence levels, but for now, we just want to understand that we can use either one (along with degrees of freedom) to locate values in the *t*-table.

In the *t*-table below, we see values for upper-tail probability along the top of the table, and confidence levels along the bottom of the table. So knowing either of these lets us locate the correct column of the table. And the degrees of freedom, which we see down the left side of the table, lets us locate the correct row.



df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	0.987	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									



For example, let's say we know that upper-tail probability is 0.01, and that our sample size is  $n = 20$ . Then we find  $n - 1 = 20 - 1 = 19$  degrees of freedom on the left of the table, and upper-tail probability 0.01 along the top of the table. The intersection of those two values is the value we want to pull from the  $t$ -table.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Keep in mind that if we'd instead been given 19 degrees of freedom and a confidence level of 98 %, we'd have found the exact same value from the  $t$ -table. That's because an upper-tail probability of 0.01 and a confidence level of 98 % both put us in the same column of the  $t$ -table. The reason this is true is because "an upper-tail probability of 0.01" and "a confidence level of 98 %" are just two different ways of saying exactly the same thing. Similarly,

- An upper-tail probability of 0.05 = a confidence level of 90 %
- An upper-tail probability of 0.025 = a confidence level of 95 %
- An upper-tail probability of 0.005 = a confidence level of 99 %

Remember, the reason the full  $t$ -table above only extends to 30 degrees of freedom is because, when the sample size is  $n \geq 30$ , values in the  $z$ -table and  $t$ -table are approximately equivalent. Therefore, once we reach more

than 30 degrees of freedom, using a  $z$ -table would give us a good enough approximation of the value we need.



# Confidence interval for the mean

We've learned how to find the sample mean and sample proportion, and we understand that these are sample statistics that we can use to estimate the values of their associated population parameters.

But as we mentioned before, a sample mean or sample proportion might be a great estimate of the population parameter, or it might be a really bad estimate. So it would be really helpful to be able to say how confident we are about how well the sample statistic is estimating the population parameter. That's where confidence intervals come in.

## Point and interval estimates

The sample mean and sample proportion are both examples of a **point estimate**, because they estimate a particular point. The point estimate for the population mean,  $\mu$ , is the sample mean,  $\bar{x}$ , and the point estimate for population standard deviation,  $\sigma$ , is sample standard deviation,  $s$ .

The benefit of using a point estimate is that it's easy to calculate. The drawback is that calculating a point estimate doesn't tell us how good or bad the estimate really is. The point estimate could be a really good estimate or a really bad estimate, and we wouldn't know one way or the other.

In contrast, we can find an **interval estimate**, which gives us a range of values in which the population parameter may lie. It's a little harder to calculate than a point estimate, but it gives us much more information.



With an interval estimate, we're able to make statements like "I'm 95 % confident that the population mean lies in the interval  $(a, b)$ ," or "I'm 99 % confident that the population proportion lies in the interval  $(a, b)$ ."

These 95 % and 99 % values we're referring to are called confidence levels. A **confidence level** is the probability that an interval estimate will include the population parameter. It's most common to choose 90 %, 95 %, or 99 % as the confidence level, and then find the interval associated with that confidence level.

It's important to clarify what we mean when we talk about a particular confidence level. To use an example, if we choose a 95 % confidence level, what we're saying is that 95 % of all confidence intervals that we find will contain the population parameter.

## Alpha $\alpha$ and the region of rejection

To take the inverse of the last statement, for a 95 % confidence level, we're saying that 5 % of the confidence intervals we find won't contain the population parameter. This 5 % (or 10 % for a 90 % confidence level, or 1 % for a 99 % confidence level) is called the **alpha value**,  $\alpha$ . We also call it the **level of significance**, or the probability of making a Type I error (more on Type I and Type II errors later). So

$$\alpha = 1 - \text{confidence level}$$

Put another way, a  $1 - \alpha$  confidence interval has a significance level of  $\alpha$ .

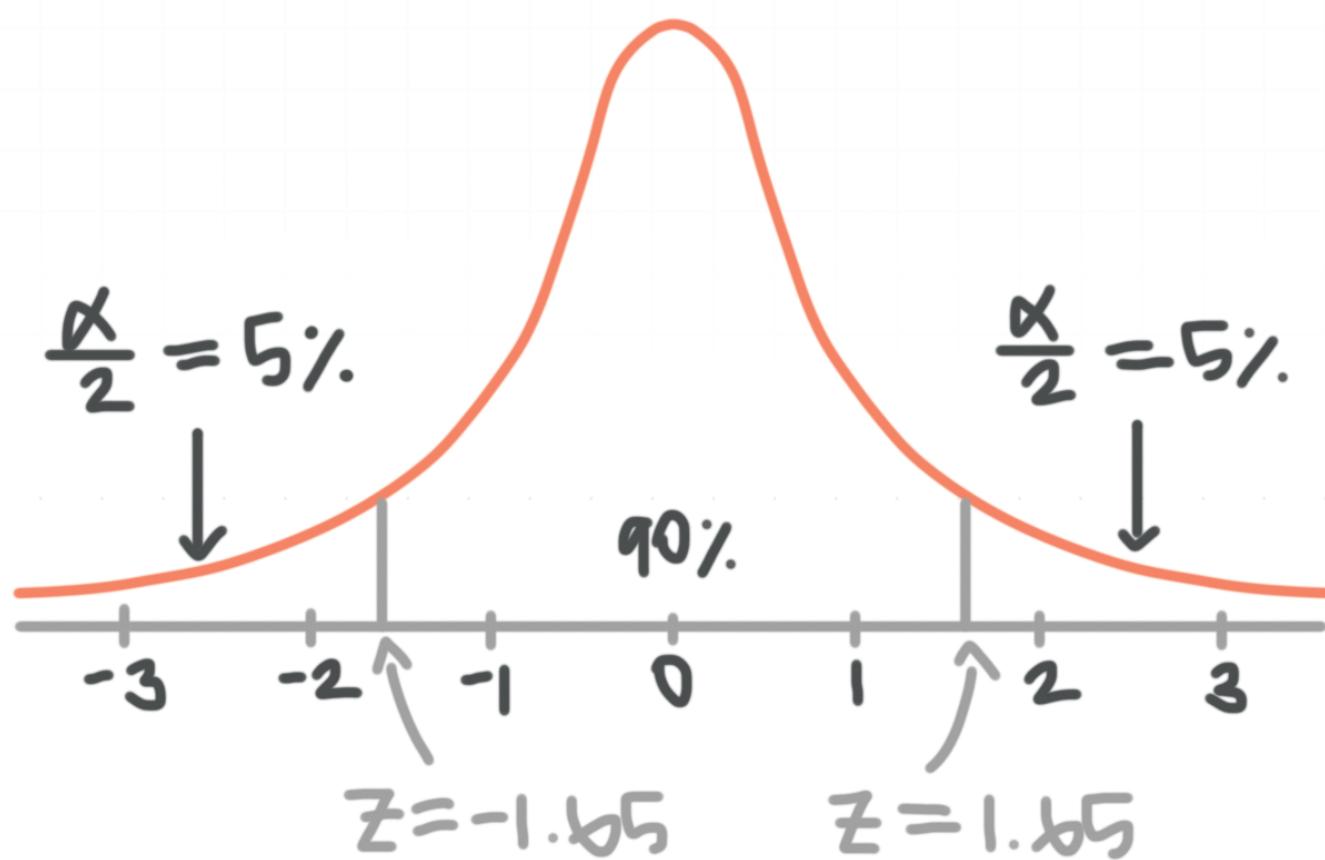


We can visualize  $\alpha$  as the total area under the normal distribution outside of the confidence interval. For instance, given a 90 % confidence level, the alpha value is

$$\alpha = 1 - 0.90$$

$$\alpha = 0.10$$

Since the confidence interval is always centered around the mean of the normal distribution, we can show the central 90 % of the distribution, with half of  $\alpha$  in the lower tail to the left of the confidence interval, and the other half of  $\alpha$  in the upper tail to the right of the confidence interval.



In other words, at a 90 % confidence level, we can expect the smallest 5 % and largest 5 % of values to fall outside the confidence interval, because  $\alpha$  is split evenly into the upper and lower tails, and  $\alpha/2 = 0.10/2 = 0.05$ .

Using a  $z$ -table, the  $z$ -values associated with  $-0.05$  and  $+0.05$  are  $-1.65$  and  $+1.65$ , respectively. Which means the boundaries of the  $90\%$  confidence interval are  $-z_{\alpha/2} = -1.65$  and  $z_{\alpha/2} = +1.65$ .

From this, we can conclude that any  $z$ -value outside of  $z = \pm 1.65$  will put us outside the  $90\%$  confidence interval, and inside the **region of rejection**. So  $\pm z_{\alpha/2}$  are the boundaries of the region of rejection.

Since we'll use them all the time, it's a good idea to know the  $z$ -values that will give us the boundaries of the region of rejection for these common confidence levels.

For a  $90\%$  confidence level,  $z = \pm 1.65$

For a  $95\%$  confidence level,  $z = \pm 1.96$

For a  $99\%$  confidence level,  $z = \pm 2.58$

## The confidence interval when $\sigma$ is known

When population standard deviation  $\sigma$  is known, the **confidence interval** is given as  $(a, b)$  by

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where  $(a, b)$  is the confidence interval,  $\bar{x}$  is the sample mean,  $z^*$  is the **critical value** (which is the  $z$ -score for the confidence level we've chosen),  $\sigma$  is population standard deviation, and  $n$  is the sample size. Since the



standard deviation of the sampling distribution of the sample mean (standard error) is  $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ , we'll also see this formula written as

$$(a, b) = \bar{x} \pm z^* \sigma_{\bar{x}}$$

In both of these versions of the formula for the confidence interval, we'll sometimes use  $z_{\alpha/2}$  instead of  $z^*$ . They mean the same thing, so using one versus the other isn't actually changing the formula. Using the  $z_{\alpha/2}$  notation is an easy way to remember that the  $\alpha$  value gets cut in half, with half of  $\alpha$  in the lower tail and half of  $\alpha$  in the upper tail, to form the region of rejection.

No matter how we write the formula, the confidence interval is always given by the sample mean  $\bar{x}$ , plus or minus the **margin of error**, so the margin of error is

$$z^* \frac{\sigma}{\sqrt{n}} = z^* \sigma_{\bar{x}} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = z_{\alpha/2} \sigma_{\bar{x}}$$

Now that we know the confidence interval formula, what's the formula actually telling us? Well, if we examine the confidence interval formula, we see that the confidence interval is related to the confidence level (as given by  $z^*$ ), the population standard deviation  $\sigma$ , and the sample size  $n$ .

From the formula, we can see that:

- The higher the confidence level, the wider the confidence interval (because as  $z^*$  gets larger, the margin of error will get larger, which makes the entire confidence interval wider).



- The larger the population standard deviation  $\sigma$ , the wider the confidence interval (because as  $\sigma$  gets larger, the margin of error will get larger, which makes the entire confidence interval wider).
- The larger the sample size  $n$ , the narrower the confidence interval (because as  $n$  gets larger, the margin of error will get smaller, which makes the entire confidence interval narrower).

In general, we want the smallest confidence interval we can get, because the smaller the confidence interval, the more accurately we can estimate the population parameter.

Keep in mind that the finite population correction factor applies to the confidence interval formula in the same way that it applied to the formula for standard error.

If sampling is done without replacement from a finite population, then the confidence interval formula we use is

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

## When $\sigma$ is unknown and/or we have a small sample

When  $\sigma$  is unknown, we use the best available approximation, which is sample standard deviation,  $s$ . But because  $s$  is a less reliable predictor of  $\sigma$  than  $\sigma$  itself, we have to use a more conservative  $t$ -value, instead of a  $z$ -value, to find the confidence interval.



$$(a, b) = \bar{x} \pm t^* \frac{s}{\sqrt{n}} = \bar{x} \pm t^* s_{\bar{x}}$$

Similarly, if our sample size is small ( $n < 30$ ), then we don't have enough data for the Central Limit Theorem to reliably apply, and we'll again have to use the more conservative  $t$ -value, instead of a  $z$ -value, in our confidence interval formula. So our confidence interval formula is

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \quad \text{whenever } \sigma \text{ is known}$$

$$(a, b) = \bar{x} \pm t^* \frac{s}{\sqrt{n}} \quad \text{whenever } \sigma \text{ is unknown, and/or } n < 30$$

Let's do an example where we find the confidence interval around the mean when population standard deviation is unknown, and the sample size is small.

### Example

The mean exam score of a sample of 10 randomly selected students is 86.7, with a sample standard deviation of 5.72. Determine the confidence interval of the true mean at a confidence level of 99 % .

Because population standard deviation is unknown, and our sample size is small, we'll have to use the confidence interval formula with a  $t$ -score instead of with a  $z$ -score.



We're given the sample mean and the sample standard deviation. The only thing we need to find is the  $t$ -value, which depends on the degrees of freedom and the confidence level. In our case,

$$df = n - 1 = 10 - 1 = 9$$

Since the confidence level is 99 %, the confidence interval will leave out 0.5 % of the area under the  $t$ -distribution in the left tail, and 0.5 % of the area under the  $t$ -distribution in the right tail.

Look up the critical  $t$ -value in the  $t$ -table.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

We see that  $t = 3.250$ . Substitute the values we've found into the formula for the confidence interval.

$$(a, b) = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$(a, b) = 86.7 \pm 3.250 \cdot \frac{5.72}{\sqrt{10}}$$

$$(a, b) \approx 86.7 \pm 5.88$$

Therefore, we can say that the confidence interval is



$$(a, b) \approx (86.7 - 5.88, 86.7 + 5.88)$$

$$(a, b) \approx (80.82, 92.58)$$

We're 99 % certain that the mean exam score for the population falls between 80.82 and 92.58.

---

Let's do an example in which population standard deviation  $\sigma$  is known.

### Example

A machine is filling water bottles, and the amount of water in the bottles has a standard deviation of  $\sigma = 1$  ounce. We take a sample of 100 bottles and find that the bottles are filled with an average of 16 ounces of water. What is the confidence interval for a confidence level of 90 % ?

Because population standard deviation is known, we can use the confidence interval formula with a  $z$ -score.

We're asking for the amount of water in ounces that correspond to an upper and lower limit for an area of 90 % in the center of the normal distribution. Which means the confidence interval will leave out 5 % of the area under the distribution in the left tail, and 5 % of the area under the distribution in the right tail.



If we look up  $z$ -scores that correspond to 5% on the lower end, and 95% on the upper end, we get  $z = \pm 1.65$ . Now plug everything we know into the confidence interval formula.

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

$$(a, b) = 16 \pm 1.65 \cdot \frac{1}{\sqrt{100}}$$

$$(a, b) = 16 \pm 1.65(0.1)$$

$$(a, b) = 16 \pm 0.165$$

Therefore, we can say that the confidence interval is

$$(a, b) = (16 - 0.165, 16 + 0.165)$$

$$(a, b) = (15.835, 16.165)$$

We're 90% certain that the actual population mean of the amount of water in the bottles is between 15.835 and 16.165 ounces.

## Required sample size for fixed margin of error

Often we'll want to determine the smallest possible sample we can take in order to stick to a specific margin of error. We can easily find the sample size by manipulating the margin of error formula and then plugging in a few values. The margin of error formula is

$$ME = z^* \frac{\sigma}{\sqrt{n}}$$

Since we want to find a sample size, we'll solve this for  $n$ .

$$ME\sqrt{n} = z^*\sigma$$

$$\sqrt{n} = \frac{z^*\sigma}{ME}$$

$$n = \left( \frac{z^*\sigma}{ME} \right)^2$$

Now let's say, for example, that we're solving a problem where we want a 95% confidence interval (corresponding to a  $z$ -score of 1.96), that the standard deviation is 5.14, and that we want a margin of error of  $\pm 2$ . Then the smallest possible sample size we can take to ensure that margin of error is

$$n = \left( \frac{1.96 \cdot 5.14}{2} \right)^2$$

$$n = 5.0372^2$$

$$n \approx 25.37$$

To meet that threshold, and keep a margin of error of  $\pm 2$  at 95% confidence, we'd need to take a sample size of at least  $n = 26$ .



# Confidence interval for the proportion

In real life, when we're interested in a proportion, we usually won't know the population proportion  $p$ , because we won't be able to survey or test every subject within our population. For instance, we might want to know the proportion of people in our country who support a particular political candidate, but, since we can't ask every person in the country, we can't truly know the population proportion,  $p$ .

Instead, we have to take a smaller sample of our larger population, and then compute the sample proportion  $\hat{p}$ . Once we find  $\hat{p}$ , we can use it to make inferences about the value of  $p$ .

We'll find the sample proportion  $\hat{p}$  by taking a sample with  $n$  subjects and surveying the number of those subjects that meet our criteria. Out of that sample, the percentage of subjects that meet our criteria will be  $\hat{p}$ .

$$\hat{p} = \frac{\text{number of subjects that meet our criteria}}{n}$$

The confidence interval for the sample proportion is given by

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where  $z^*$  comes from the  $z$ -table,  $\hat{p}$  is the sample proportion, and  $n$  is the sample size. We use  $\hat{p}$  in the confidence interval formula (instead of  $p$ ) because, if we're constructing a confidence interval, by definition that means we're trying to use a sample proportion to estimate the population proportion, which means we don't know the population proportion. Since



we don't know the population proportion, we use the sample proportion  $\hat{p}$  in our formula, instead of the population proportion  $p$ .

In order to be able to use this formula, we need to have  $n\hat{p} \geq 5$  and  $n(1 - \hat{p}) \geq 5$ , which will almost always be the case in real life, as long as our sample size is reasonably large.

The finite population correction factor applies here as well, so if we're sampling without replacement from more than 5 % of a population of finite size  $N$  ( $n/N > 0.05$ ), then the confidence interval for the population proportion is given by

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \sqrt{\frac{N - n}{N - 1}}$$

So if we know how we're sampling, what confidence level we want to use, and we know the sample proportion, then we can plug these values into the correct formula, find the critical value associated with the confidence level, and then calculate the confidence interval directly.

Let's do an example so that we can see how to calculate a confidence interval for the population proportion.

### Example

There are 500 sea turtles that live in a bay off of Maui, Hawaii, and we want to estimate the proportion that are male. Let's say we take a random sample of 50 turtles and find that 20 of them are male.

Based on this sample, what is a 90 % confidence interval for the proportion of male sea turtles in the bay.

As always, we have to first check for normality. We were told that the sample we took was random.

Based on the sample proportion  $\hat{p} = 20/50 = 0.4$ , we'll get at least 5 “successes” ( $50 \cdot 0.4 = 20$ ) and at least 5 “failures” ( $50 \cdot 0.6 = 30$ ), so we've met the normal condition. It looks like we're sampling without replacement, using  $50/500 = 10\%$  of the population, so we'll need to use the finite population correction factor in our confidence interval formula.

$$(a, b) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \sqrt{\frac{N - n}{N - 1}}$$

$$(a, b) = 0.4 \pm z^* \sqrt{\frac{0.4(0.6)}{50}} \sqrt{\frac{500 - 50}{500 - 1}}$$

For a 90% confidence interval, we're looking in a normal distribution at the middle 90% of probability, which means we'll only have 10% probability in the two tails, or just 5% in the upper tail, which means we're interested in the  $z$ -score that puts us at 95% probability. If we look for approximately 0.9500 in the  $z$ -table, we get about 1.65. So the critical value  $z^*$  is approximately 1.65, and we can say that our confidence interval is

$$(a, b) = 0.4 \pm 1.65 \sqrt{\frac{0.4(0.6)}{50}} \sqrt{\frac{500 - 50}{500 - 1}}$$

$$(a, b) = 0.4 \pm 1.65 \sqrt{\frac{0.24}{50}} \sqrt{\frac{450}{499}}$$

$$(a, b) \approx 0.4 \pm 1.65 \sqrt{0.0048} \sqrt{0.9018}$$

$$(a, b) \approx 0.4 \pm 0.1086$$

$$(a, b) \approx (0.4 - 0.1086, 0.4 + 0.1086)$$

$$(a, b) \approx (0.2914, 0.5086)$$

We interpret this to mean that about 90 % of the confidence intervals we construct this way (with 50-turtle samples) will contain the actual population proportion  $p$  of male sea turtles in the bay.

---

## Margin of error

Just like for the confidence interval for the mean, the margin of error for the proportion is simply the part of the confidence interval formula that comes after the  $\pm$  sign.

$$z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

If we want to keep our margin of error at or below a certain value, then we can set up an inequality that will allow us to find the minimum possible sample size we'd need to use in order to keep the margin of error fixed to that preset maximum.

---

### Example

We want the margin of error in our sea turtle study (from the previous example) to be no more than  $\pm 4\%$  at a 90 % confidence level. Find the



smallest possible sample size we can use to stay within that margin of error.

First, we'll set up the inequality.

$$z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \sqrt{\frac{N - n}{N - 1}} \leq 0.04$$

If we want to find the smallest possible sample size  $n$  that keeps us within this margin of error, then we need to optimize  $\hat{p}(1 - \hat{p})$ , since making the numerator of a fraction as large as possible will make the entire fraction as large as possible. In turn, that will make the value of the square root as large as possible, which will make the entire value on the left side of the inequality as large as possible, thereby minimizing the value of  $n$ .

We could prove this algebraically, but the value of  $\hat{p}$  that optimizes  $\hat{p}(1 - \hat{p})$  is always  $\hat{p} = 0.5$ . Therefore, we'll plug everything we know into the margin of error inequality, remembering that a 90% confidence level has a critical value of approximately 1.65.

$$1.65 \sqrt{\frac{0.5(0.5)}{n}} \sqrt{\frac{500 - n}{500 - 1}} \leq 0.04$$

$$\frac{\sqrt{0.5^2}}{\sqrt{n}} \cdot \frac{\sqrt{500 - n}}{\sqrt{499}} \leq \frac{0.04}{1.65}$$

$$\frac{0.5}{\sqrt{n}} \cdot \frac{\sqrt{500 - n}}{\sqrt{499}} \leq \frac{0.04}{1.65}$$



$$\frac{0.5}{\sqrt{499}} \cdot \frac{\sqrt{500-n}}{\sqrt{n}} \leq \frac{0.04}{1.65}$$

Solve the inequality for  $n$ .

$$\frac{\sqrt{500-n}}{\sqrt{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500-n}{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500}{n} - \frac{n}{n}} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\sqrt{\frac{500}{n} - 1} \leq \frac{0.04}{1.65} \cdot \frac{\sqrt{499}}{0.5}$$

$$\frac{500}{n} - 1 \leq \left( \frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2$$

$$\frac{500}{n} \leq \left( \frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2 + 1$$

We can invert both sides if we flip the inequality sign.

$$\frac{n}{500} \geq \frac{1}{\left( \frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2 + 1}$$

$$n \geq \frac{500}{\left( \frac{0.04\sqrt{499}}{1.65 \cdot 0.5} \right)^2 + 1}$$

$$n \geq 230.092$$

If we need to sample more than 230.092 members of our population, that means we need to sample at least 231 of them, because sampling only 230 wouldn't quite be enough to meet our threshold. Therefore, our minimum sample size has to be

$$n \geq 231$$

---



# Inferential statistics and hypotheses

One thing we like to do in statistics is make a statement about a population parameter, collect a sample from that population, investigate the sample, and then make a clear statement about whether or not that sample supports our original statement about the population.

We'll cover each part of this process throughout this section. Here's where we're headed: There are five steps for **hypothesis testing**:

1. State the null and alternative hypotheses.
2. Determine the level of significance.
3. Calculate the test statistic.
4. Find critical value(s) and determine the regions of acceptance and rejection.
5. State the conclusion.

This process is also called **inferential statistics**, because we're using information we have about the sample to make *inferences* about the population.

For instance, we might have been told that 40 % of cars in our town are blue. We could try to take a random sample of cars in our town, look at the proportion of cars in that sample which are blue (maybe 37 %), and then build a confidence interval to state how confident we are that our sample, which produced  $\hat{p} = 0.37$ , supports the claim that  $p = 0.40$ .



## Proof vs. support

But it's important to make the distinction right up front between proving a claim and providing support for a claim.

When we do inferential statistics, we're usually not able to prove something with certainty (our  $\hat{p} = 0.37$  doesn't *prove* that  $p = 0.40$ , even though it might lend some support for that claim). Instead, we use the data to support a theory that we have. Hopefully, if the data is strong enough, we can provide strong, or confident support for our theory, but we still can't necessarily prove it.

## Hypotheses for means and proportions

Before we can use inferential statistics, we first need a **hypothesis**, which is a statement of expectation about a population parameter that we develop for the purpose of testing it (40% of cars in our town are blue).

In any hypothesis test, the first thing we always want to do is state what are called the null and alternative hypotheses. Every hypothesis test contains this set of two opposing statements about a population parameter.

The **alternative hypothesis**  $H_a$  is the abnormality we're looking for in the data; it's the significance we're hoping to find. Once we have an alternative hypothesis, we always want to state the opposite claim, which we call the **null hypothesis**,  $H_0$ .



For example, if our city releases data stating that 40% of the cars in our town are blue, then this 40% figure is the status quo; it's our normal baseline; it should be our null hypothesis. If we then want to test this claim, to see if we can find interesting new data that shows that the city's claim is wrong, then we'd be looking for data that goes against their stated figure. So the alternative hypothesis will be that the proportion of blue cars is contradictory to what the city has stated.

$H_a$ : the proportion of blue cars in our town is not 40%

$H_0$ : 40% of cars in our town are blue

Interestingly enough, we always test the null hypothesis  $H_0$ , not the alternative hypothesis  $H_a$ . We say that, if our sample gives us good enough evidence, then we can *reject* the null hypothesis, and therefore provide evidence that supports our alternative hypothesis.

In this section we'll focus on hypothesis tests about two population parameters: the population mean  $\mu$  and the population proportion  $p$ .

### For population means

Remember that the population mean is the mean value of some characteristic that we're interested in. For example, the mean height of American females might be  $\mu = 65$  inches if the average American woman is 5'5" tall.

Whether we're investigating a population proportion or a population mean, the null hypothesis states the status quo that the population parameter is  $\leq$ ,  $=$ , or  $\geq$  the claimed value. The null hypothesis always says



that the population mean (or parameter) is normal; nothing new or different is happening.

So if we're testing the claim that the mean height of American females is  $\mu = 65$  inches, the null hypothesis is  $H_0 : \mu = 65$ .

If we think the mean height of American females is different than this claim, then we state that in the alternative hypothesis as

- The mean height of American females is different than  $\mu = 65$ :

$$H_a : \mu \neq 65$$

If, on the other hand, we'd started with a null hypothesis of  $H_0 : \mu \leq 65$ , then our alternative hypothesis would be

- The mean height of American females is greater than  $\mu = 65$ :

$$H_a : \mu > 65$$

and if we'd started with a null hypothesis of  $H_0 : \mu \geq 65$ , then our alternative hypothesis would be

- The mean height of American females is less than  $\mu = 65$ :

$$H_a : \mu < 65$$

## For population proportions

On the other hand, the population proportion is the proportion that meets some sort of criteria we've established. For example, the proportion of American females with blue eyes might be  $p = 0.15$  if 15% of American



females have blue eyes. The null hypothesis would be  $H_0 : p = 0.15$ , stating that the population proportion is 15%.

If we think the population proportion is different than this 15% claim, then we state that in the alternative hypothesis as

- The proportion of American females with blue eyes is different than  $p = 0.15$ :

$$H_a : p \neq 0.15$$

If, on the other hand, we'd started with a null hypothesis of  $H_0 : p \leq 0.15$ , then our alternative hypothesis would be

- The proportion of American females with blue eyes is greater than  $p = 0.15$ :

$$H_a : p > 0.15$$

and if we'd started with a null hypothesis of  $H_0 : p \geq 0.15$ , then our alternative hypothesis would be

- The proportion of American females with blue eyes is less than  $p = 0.15$ :

$$H_a : p < 0.15$$

As we can see for both population means and population proportions, the alternative hypothesis states the opposite of the null hypothesis. We find support for the alternative hypothesis only because we find a reason to reject the null hypothesis. Because the null hypothesis always includes a  $\leq$ ,



=, or  $\geq$  sign, the alternative hypothesis always includes a  $>$ ,  $\neq$ , or  $<$  sign, respectively. In summary,

If  $H_0$  is  $\mu =$  or  $p =$ , then  $H_a$  is  $\mu \neq$  or  $p \neq$

If  $H_0$  is  $\mu \leq$  or  $p \leq$ , then  $H_a$  is  $\mu >$  or  $p >$

If  $H_0$  is  $\mu \geq$  or  $p \geq$ , then  $H_a$  is  $\mu <$  or  $p <$

Let's practice writing pairs of hypothesis statements.

### Example

Write different sets of hypothesis statements to test the claims that students at Springdale High School perform 1) differently than, 2) better than, and 3) worse than students at Greenville High School on the SAT test.

1) To test the claim that students at SHS perform differently on the SAT than students at GHS, we would write these hypothesis statements:

**Null:** Students at Springdale High School do not perform differently on the SAT than students at Greenville High School:

$$H_0 : \mu_S = \mu_G$$

**Alternative:** Students at Springdale High School perform differently on the SAT than students at Greenville High School:

$$H_a : \mu_S \neq \mu_G$$



2) To test the claim that students at SHS perform better on the SAT than students at GHS, we would write these hypothesis statements:

**Null:** Students at Springdale High School perform no better on the SAT than students at Greenville High School:

$$H_0 : \mu_S \leq \mu_G$$

**Alternative:** Students at Springdale High School perform better on the SAT than students at Greenville High School:

$$H_a : \mu_S > \mu_G$$

3) To test the claim that students at SHS perform worse on the SAT than students at GHS, we would write these hypothesis statements:

**Null:** Students at Springdale High School perform no worse on the SAT than students at Greenville High School:

$$H_0 : \mu_S \geq \mu_G$$

**Alternative:** Students at Springdale High School perform worse on the SAT than students at Greenville High School:

$$H_a : \mu_S < \mu_G$$

---

Keep in mind that some statisticians and statistics textbooks will use the convention where the null hypothesis always and only includes an = sign, with the alternative hypothesis stated with <, ≠, or >.

We can use either convention, we just have to stay consistent with whichever one we choose. For our purposes, we'll always stick with the matching pairs we outlined earlier:

If  $H_0$  is  $\mu =$  or  $p =$ , then  $H_a$  is  $\mu \neq$  or  $p \neq$

If  $H_0$  is  $\mu \leq$  or  $p \leq$ , then  $H_a$  is  $\mu >$  or  $p >$

If  $H_0$  is  $\mu \geq$  or  $p \geq$ , then  $H_a$  is  $\mu <$  or  $p <$

# Significance level and type I and II errors

Whenever we're using hypothesis testing, we always run the risk that the sample we chose isn't representative of the population. Even if the sample was random, it might not be representative.

For instance, if we've been told that 15% of American females have blue eyes, and we've set up null and alternative hypotheses to test this claim,

$H_0$ : 15% of American females have blue eyes

$H_a$ : the percentage of American females with blue eyes is not 15%

then when we take a sample to investigate our null hypothesis, we still run the risk of committing two types of errors.

## Type I and Type II errors

Let's assume that the null hypothesis is true and that the percentage of American females with blue eyes is 15%. But let's say we take a sample of 100 women, find that 40 of them have blue eyes, and therefore calculate a sample proportion of  $\hat{p} = 40\%$ .

If, based on the large difference between the sample proportion and the hypothesize proportion (40% versus 15%), we reject the null hypothesis, we've just made a **Type I error**. In other words, we make a Type I error when we mistakenly reject a null hypothesis that's actually true. The probability of making a Type I error is alpha,  $\alpha$ , also called the **level of significance**.



Now let's consider the opposite situation and assume that the null hypothesis is false, such that the percentage of American females with blue eyes is not 15 %. But imagine that we take a sample of 100 women and find a sample proportion of  $\hat{p} = 15 \%$ .

If, based on the equality of the sample proportion and the hypothesized portion, we accept the null hypothesis, we've just made a **Type II error**. In other words, we make a Type II error when we mistakenly accept the null hypothesis when it's actually false. The probability of making a Type II error is beta,  $\beta$ .

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

There are lots of other ways to describe Type I and Type II errors, including

Type I error: Supporting the alternative hypothesis when the null hypothesis is true.

Type II error: Not supporting the alternative hypothesis when the null hypothesis is false.

Thinking about Type I and Type II errors can get people a little twisted around sometimes, so if we find that there's one description of them that makes more sense to us than the others, we can stick with that one.

Because  $\alpha$  is literally the probability of making a Type I error, and  $\beta$  is literally the probability of making a Type II error, we can say that the **alpha level** is

- the probability of making the wrong decision when the null hypothesis is true, or
- the probability of rejecting the null hypothesis when it's true, or
- the probability of making a Type I error

and the **beta level** is

- the probability of making the wrong decision when the null hypothesis is false, or
- the probability of accepting the null hypothesis when it's false, or
- the probability of making a Type II error

Since the probability of committing a Type I error is  $\alpha$ , the probability of making a correct decision when  $H_0$  is true is  $1 - \alpha$ . And since the probability of committing a Type II error is  $\beta$ , the probability of making a correct decision when  $H_0$  is false is  $1 - \beta$ .

Let's do an example where we look at the  $\alpha$  and  $\beta$  levels in a hypothesis test.

### Example

Lynnie is testing the hypothesis that people in her town spend more money on coffee on Monday than they do on Tuesday. She doesn't know



it, but her hypothesis is false: people *don't* spend more money on coffee on Monday. She picks a random sample of people in her town and asks them how much money they spent on coffee each day. Say whether Lynnie will make a Type I or Type II error.

	Monday	Tuesday
Average spend	\$6.75	\$5.45

Lynnie's null and alternative hypotheses are

$H_0$ : People spend no more on coffee on Monday than they do on Tuesday

$$C_M \leq C_T$$

$H_a$ : People spend more on coffee on Monday than they do on Tuesday

$$C_M > C_T$$

In reality, her alternative hypothesis is false. But her sample data is showing that people spend more on Monday than they do on Tuesday. Which means she's in danger of rejecting the null hypothesis when she shouldn't, since the null hypothesis is true.



	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

From the table we looked at earlier, the intersection of “reject the null” and “the null is true” is a Type I error. Lynnie is in danger of committing a Type I error.

---

## Power

Sometimes we say that the **power** of a hypothesis test is the probability that we'll reject the null hypothesis when it's false, which is a correct decision. Rejecting the null hypothesis when it's false is exactly what we want to do.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error $P(\text{Type I error})=\alpha$	CORRECT Power
Accept $H_0$	CORRECT	Type II error $P(\text{Type II error})=\beta$

So, the higher the power of our test, the better off we are. Power is also equal to  $1 - \beta$ .

## Confidence levels and the $\alpha$ value

This  $\alpha$  value, or level of significance, is the same  $\alpha$  value we talked about when we looked at confidence levels and confidence intervals.

Remember that we usually pick a confidence level of 90 %, 95 %, or 99 %, and these correspond to  $\alpha$  values of

At 90 % confidence, the alpha value is  $\alpha = 1 - 90\% = 10\%$

At 95 % confidence, the alpha value is  $\alpha = 1 - 95\% = 5\%$

At 99 % confidence, the alpha value is  $\alpha = 1 - 99\% = 1\%$

So, in the same way that we said we normally pick a confidence level of 90 %, 95 %, or 99 %, we could equivalently say that we normally pick an  $\alpha$  value of 10 %, 5 %, or 1 %.

When we decide on  $\alpha$ , we're actually deciding how much we want to risk committing a Type I error. In other words, if we choose  $\alpha = 0.05$ , we're saying that, 5 % of the time, or 1 out of 20 times, we'll reject the null hypothesis when the null hypothesis is actually true.

Choosing a significance level of 1 % means we want to be more confident about the result than if we'd picked  $\alpha = 10\%$ . While it's true that we always want to be as confident as possible about our result, remember that picking a higher confidence level (and therefore lower alpha value) comes at a cost. The lower the alpha value, the wider the confidence interval and the larger the margin of error. This means that it'll be less likely that we detect a true difference between our sample statistic and the hypothesized value if that difference actually exists.



Similarly, since  $\alpha$  is the probability of making a Type I error, and  $\beta$  is the probability of making a Type II error, we'd obviously like to minimize  $\alpha$  and  $\beta$  as much as possible, because of course we always want to minimize the possibility that we'll make an error.

However, keep in mind that if we decrease  $\alpha$ , it becomes more difficult to reject the null hypothesis, because the region of acceptance grows while the region of rejection shrinks. And if the null hypothesis is false while we decrease  $\alpha$ , the risk of committing a Type II error increases because  $\beta$  gets larger and the power of the test decreases.

Alternatively, if we increase  $\alpha$ , it becomes easier to reject the null hypothesis, because the region of rejection grows while the region of acceptance shrinks; the power of the test increases. We're at risk of committing a Type I error whenever we reduce the risk of committing a Type II error.

In other words, reducing the  $\alpha$  value increases the  $\beta$  value, and vice versa. The only way to reduce them both simultaneously is to increase the sample size. If we could increase the sample size until it's as big as the population, the values of  $\alpha$  and  $\beta$  would be 0.

Because of the inverse relationships between  $\alpha$  and  $\beta$ , we're always trying to decide which type of error is more dangerous, and the answer to that depends on the situation. The question we need to ask ourselves is “What's the worst-case scenario?” For example, let's say a factory produces car parts and has a strict quality-control process in place. They assume that their production meets the minimum quality requirements, so that's their null hypothesis. The factory wants a low  $\alpha$ , because that means

they have to reject fewer parts as defective, which saves them money. But this lower  $\alpha$  value might mean that more defective parts make it into cars, which could lead to cars that are less safe for consumers.

On the other hand, if we're a consumer who purchases a car made with these parts, we might prefer that the factory uses a higher  $\alpha$ , rejects more defective car parts, thereby making sure our car is as safe as possible. However, if the factory uses a higher  $\alpha$  value to keep the car safer, we may have to pay more for the car to account for the increased number of wasted defective parts.

So increasing the  $\alpha$  level will decrease the Type II error risk for the consumer, but increase the Type I error risk for the producer. In other words, there are competing interests that are affected by changing the  $\alpha$  value, and we have to decide exactly what  $\alpha$  value gives us the balance we want.

# Test statistics for one- and two-tailed tests

Remember that our steps for any hypothesis test are

1. State the null and alternative hypotheses.
2. Determine the level of significance.
3. Calculate the test statistic.
4. Find critical value(s) and determine the regions of acceptance and rejection.
5. State the conclusion.

We've already covered the first two steps, and now we want to talk about how to calculate the test statistic. But any test statistic we calculate will depend on whether we're running a two-tailed test or a one-tailed test. Whether we run a one- or two-tailed test is dictated by the hypothesis statements we wrote in the first step.

So let's define one- and two-tailed tests, and start over with the hypothesis statements to show when we'll use each test type.

## One- and two-tailed tests

We've already learned that, for both means and proportions, we can write the null and alternative hypothesis statements in three ways:

$$H_0 \text{ with an } = \text{ sign and } H_a \text{ with a } \neq \text{ sign}$$

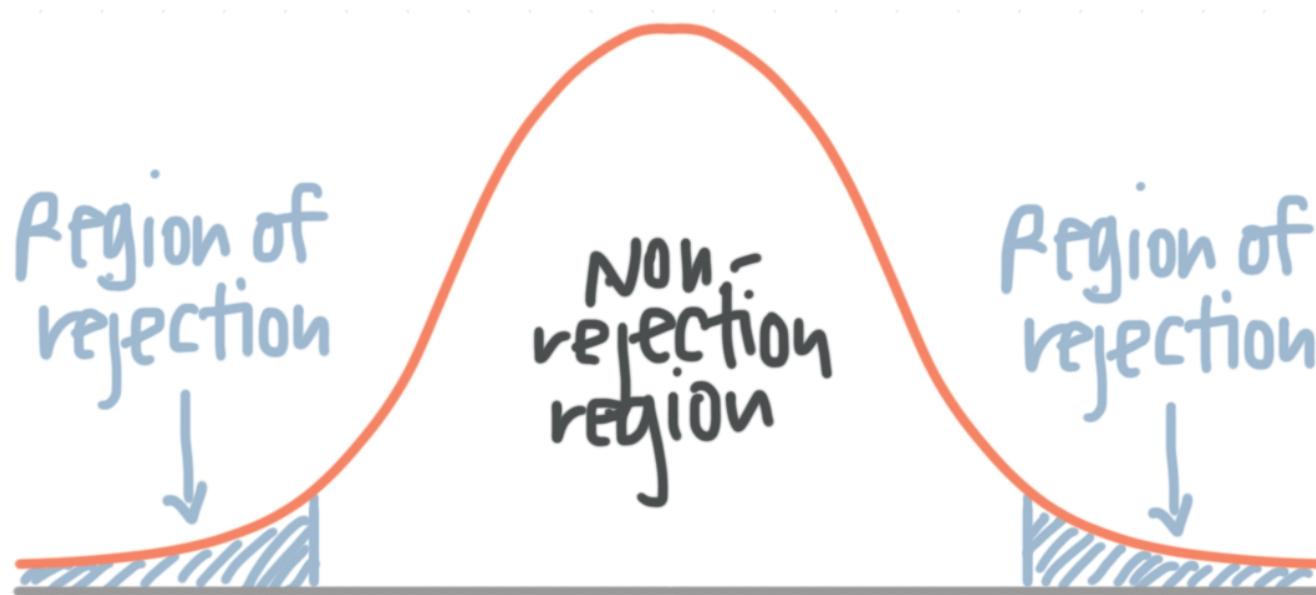


$$H_0 \text{ with a } \leq \text{ sign and } H_a \text{ with a } > \text{ sign}$$

$$H_0 \text{ with a } \geq \text{ sign and } H_a \text{ with a } < \text{ sign}$$

When the null and alternative hypotheses use the  $=$  and  $\neq$  signs, we'll use a **two-tailed test** (also called a two-sided or nondirectional test). But in the other two cases, with either  $\leq$  and  $>$ , or  $\geq$  and  $<$ , we'll use a **one-tailed test** (also called a one-sided test or direction test).

Here's the way to understand one- and two-tailed tests. When we state in the null hypothesis that the population mean or population proportion is equal to some value, and then state with the alternative hypothesis that the mean or proportion is not equal to that value, we're not predicting any direction between the variables. We're not saying that one value is greater or less than the other, we're just saying that they're different. Which means the difference could be in either direction (less than or greater than), which means we'll have a region of rejection in each tail of the distribution.



We call it a “two-tailed test” because we have two regions of rejection, one in each tail.

On the other hand, if we hypothesize that the mean or proportion is greater than or less than the stated value, then it means we'll be performing a one-tailed test. If we predict that the population parameter is greater than the stated value, then the only rejection region will be in the upper tail, so we call this an **upper-tailed test**, or a **right-tailed test**.



But if we predict that the population parameter is less than the stated value, then the only rejection region will be in the lower tail, so we call this a **lower-tailed test**, or a **left-tailed test**.



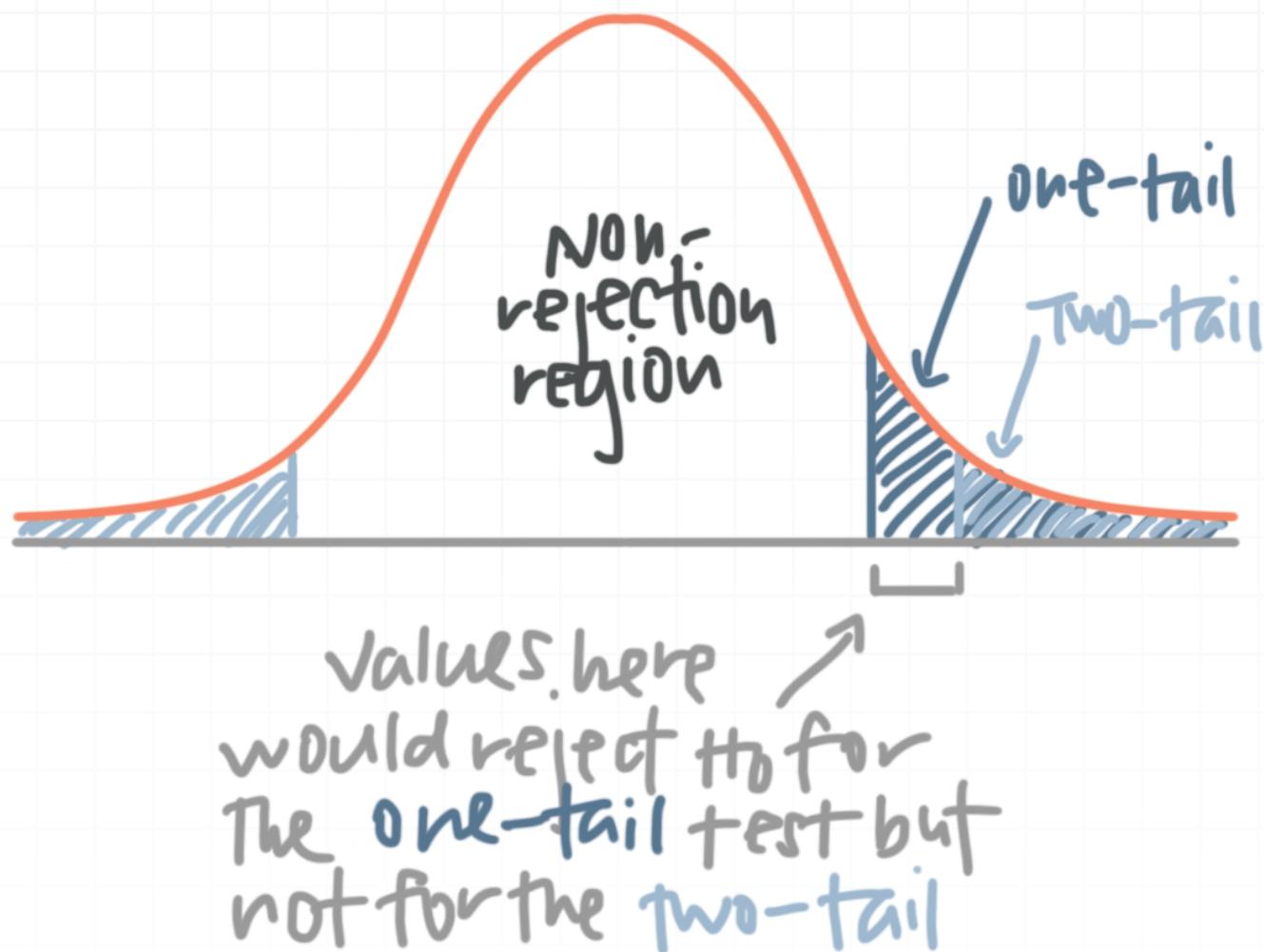
## Choosing a one-tailed or two-tailed test

Whether we use a one- or two-tailed test is determined by the hypothesis statements we choose. With that in mind, we really want to think ahead when we're writing our hypothesis statements, and consider which kind of test we want to set ourselves up for.

A one-tailed test has a larger region of rejection, because all of the area that represents the region of rejection is consolidated into one tail. A two-tailed test, on the other hand, has the region of rejection split into two tails, which means each individual rejection region for the two-tailed test is smaller than the single rejection region from the one-tailed test.



Inherently, this means that a two-tailed test is always more conservative than a one-tailed test. Looking at this last figure, we could get a whole range of results that fall within the rejection region of the one-tailed test, but fail to reach the rejection region of the two-tailed test.



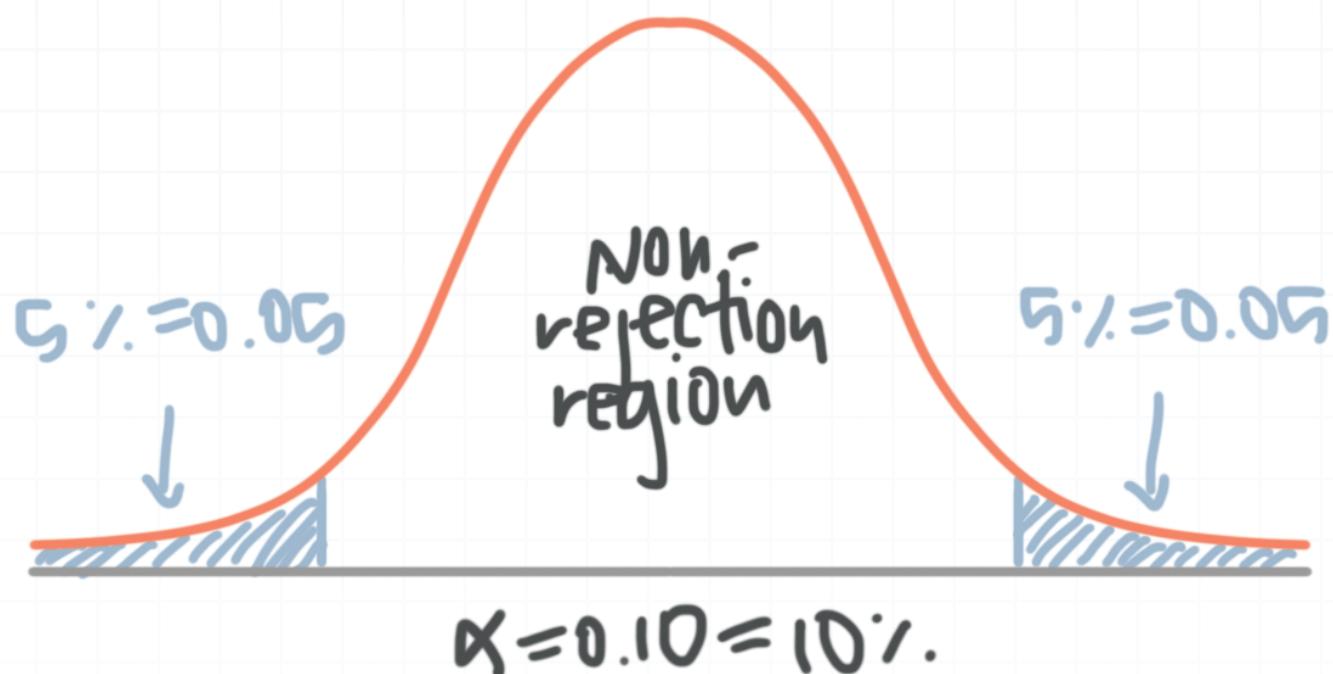
The two-tailed test really forces us to find a more extreme result in order to reach the region of rejection and reject the null hypothesis.

That being said, we should only use a one-tailed test when we have good reason to believe that the difference between the means or proportions is in the specific direction that we think it's in. If we're not extremely confident about directionality, we should play it safe and use the more conservative two-tailed test.

## The $\alpha$ value for one- and two-tailed tests

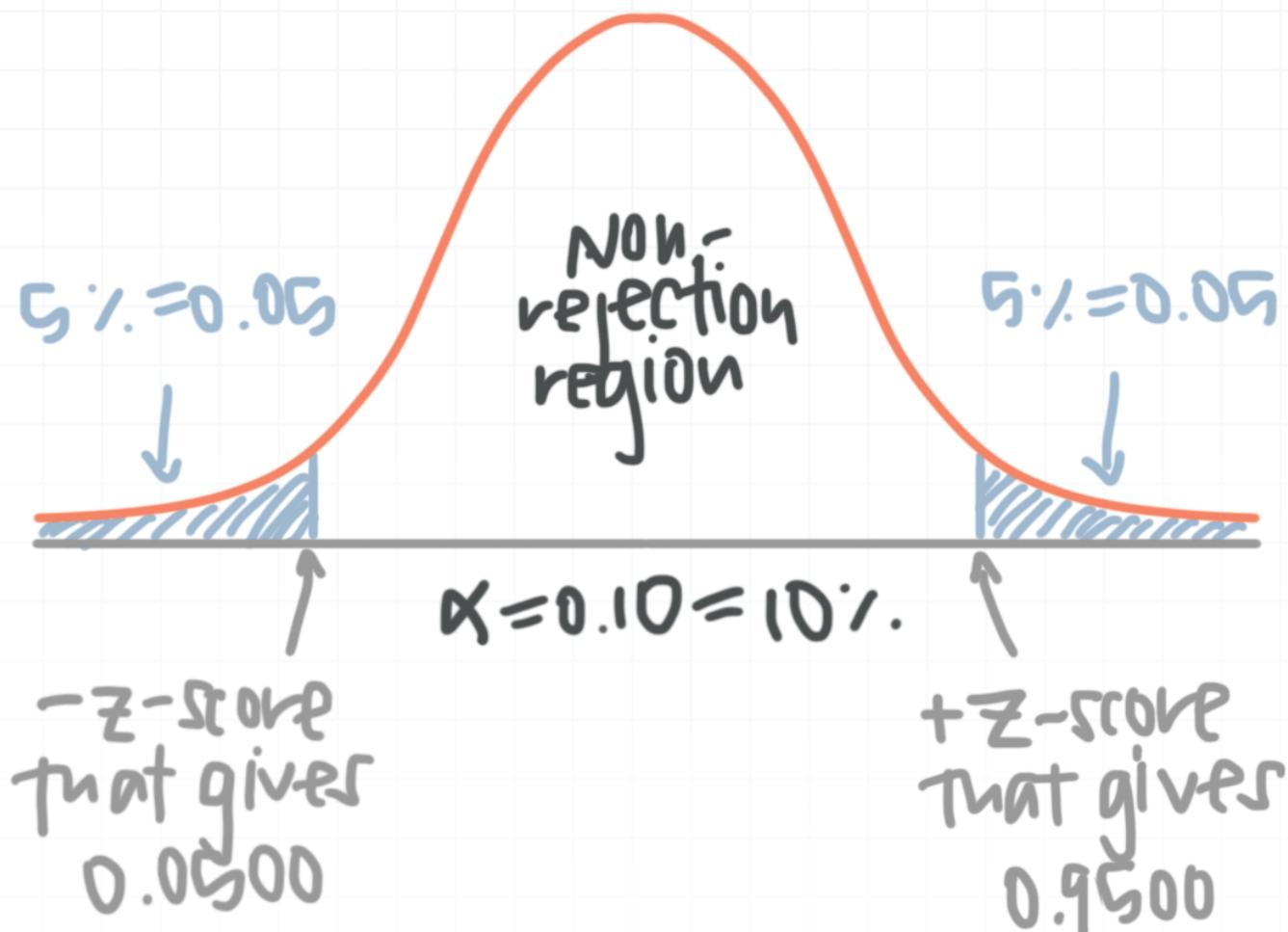
Let's say we're running two different tests. One is two-tailed, and the other is one-tailed. If we set the significance level at  $\alpha = 0.10$ , then in the

two-tailed test we'll split that 10% evenly into two tails, 5% in the lower tail and 5% in the upper tail.



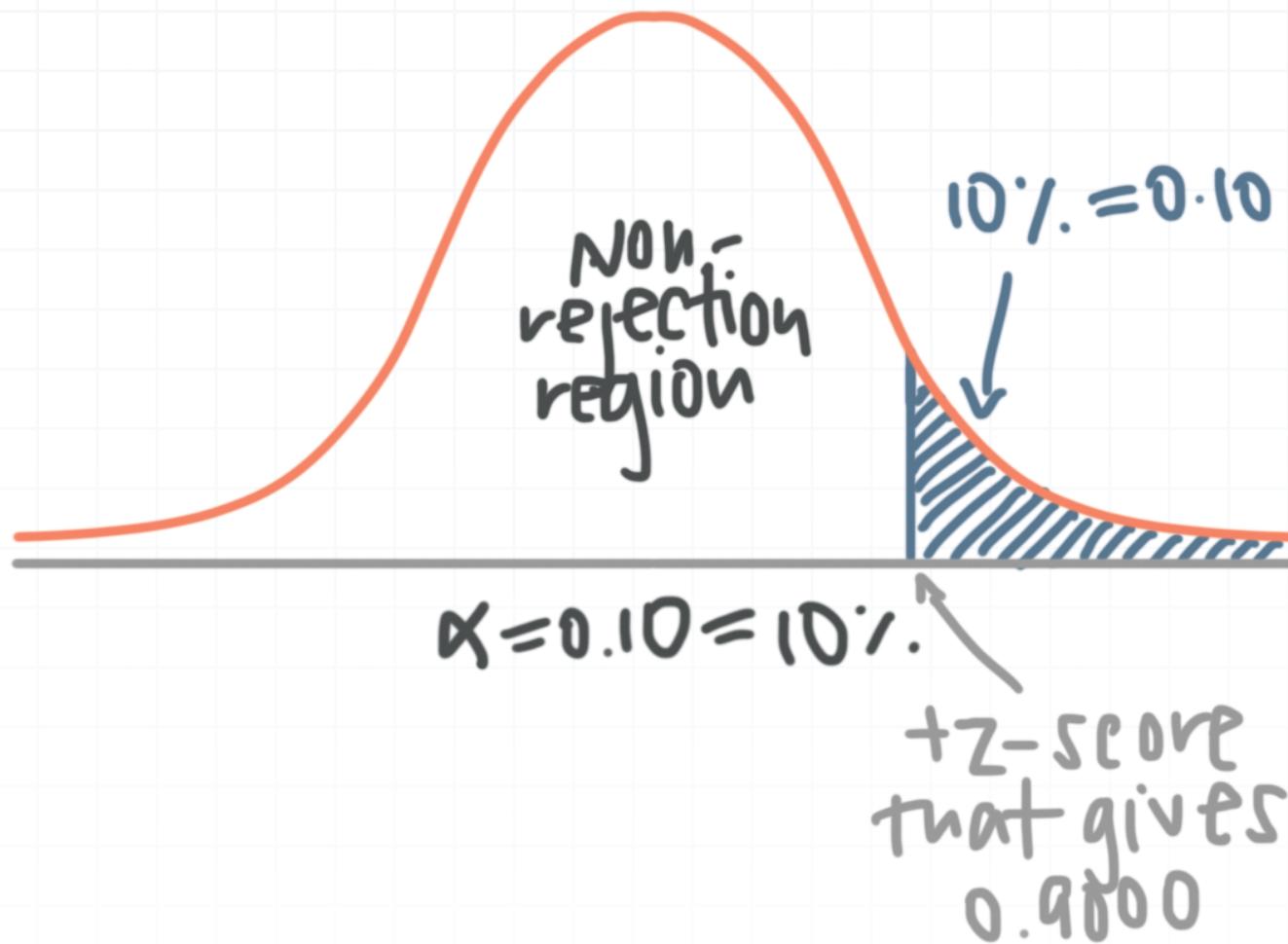
Here's something that's a little tricky: When we go to look up the  $z$ -score that corresponds to the boundary of the upper rejection region, realize that only 5% of the rejection area is in that upper tail, not 10%. Which means that, when we look for the  $z$ -score, we need to look for a  $z$ -score that corresponds to 0.9500, not 0.9000, even though  $\alpha = 0.10$ .

And the same goes for the lower tail. When we look for the negative  $z$ -score that corresponds to the boundary of the rejection region in the lower tail, we need to look for a  $z$ -score that corresponds to 0.0500, not 0.1000, even though  $\alpha = 0.10$ .



Many people get tripped up here, because they assume that, with  $\alpha = 0.10$ , they're looking for the  $z$ -score associated with 0.9000. But because the rejection region is split into both tails, we only have  $10\% / 2 = 5\%$  of the area in each tail, and that 5% corresponds to 0.9500 in the upper tail and 0.0500 in the lower tail.

But if instead we're using a one-tailed test with the same  $\alpha = 0.10$ , the entire 10% of area that represents the rejection region is consolidated into one tail of the distribution. So if we're running an upper-tailed test, then we'll find the boundary of the rejection region at a  $z$ -score for 0.9000.



For a lower-tailed test, we'll find the boundary of the rejection region at a  $z$ -score for 0.1000.

## Calculating the test statistic

At this point, we know how to write the hypothesis statements, determine the alpha level we want to use, and set up a one- or two-tailed test based on the hypothesis statements. We can also find the boundary or boundaries of the rejection region based on the alpha value and which test type we're using.

The next step is always to calculate the test statistic, and then determine whether that value lies inside or outside the region of rejection.

The formula we'll use to calculate the test statistic depends on the information we have and the size of our sample, but the general formula for the test statistic is

$$\text{test statistic} = \frac{\text{observed} - \text{expected}}{\text{standard deviation}}$$

The specific formula for the test statistic will depend on whether or not the population standard deviation  $\sigma$  is known or unknown. When  $\sigma$  is known, we'll use

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

When  $\sigma$  is unknown, and/or if we're using a small sample  $n < 30$ , we'll use the sample standard deviation  $s_{\bar{x}}$  instead of the population standard deviation  $\sigma$ , but we'll use the more conservative  $t$ -table to compensate for using  $s_{\bar{x}}$ .

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

These are test statistics for the mean, but the test statistic for the proportion, as long as  $n\hat{p} \geq 5$  and  $n(1 - \hat{p}) \geq 5$ , will be

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Let's do an example where we calculate the test statistic for the mean, where population standard deviation is unknown (which will usually be the case).

### Example

We're testing the claim that a car dealership averages \$1,000,000 in monthly sales. We take a random sample of 40 months of sales data and find  $\bar{x} = \$985,000$  and  $s = \$200,000$ . Find the hypothesis statements, choose a one- or two-tailed test, and calculate a test statistic.

There's nothing in the problem to indicate we're very confident about directionality, so we should choose a two-tailed test, and the hypothesis statements will therefore be

$H_0$ : The dealership sells \$1,000,000

$$\mu = \$1,000,000$$

$H_a$ : The dealership makes sales other than \$1,000,000

$$\mu \neq \$1,000,000$$

Because population standard deviation is unknown, the test statistic will be

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$



$$t = \frac{\$985,000 - \$1,000,000}{\frac{\$200,000}{\sqrt{40}}}$$

$$t = -\$15,000 \cdot \frac{\sqrt{40}}{\$200,000}$$

$$t = -\frac{3\sqrt{40}}{40}$$

$$t = -\frac{3\sqrt{10}}{20}$$

$$t \approx -0.4743$$

In the next section, we'll look at what to do with this test statistic once we have it.

# The p-value and rejecting the null

Technically, we define the **p-value** (or the observed level of significance) as the smallest level of significance at which we can reject the null hypothesis, assuming the null hypothesis is true.

We can also think about the *p*-value as the total area of the region of rejection. Remember that in a one-tailed test, the region of rejection is consolidated into one tail, whereas in a two-tailed test, the region of rejection is split between two tails.

So, as we might expect, calculating the *p*-value as the area of the rejection region will be slightly different depending on whether we're using a one-tailed test or a two-tailed test, and whether the one-tailed test is an upper-tailed test or lower-tailed test.

## Calculating the *p*-value

### For a one-tailed, lower-tailed test

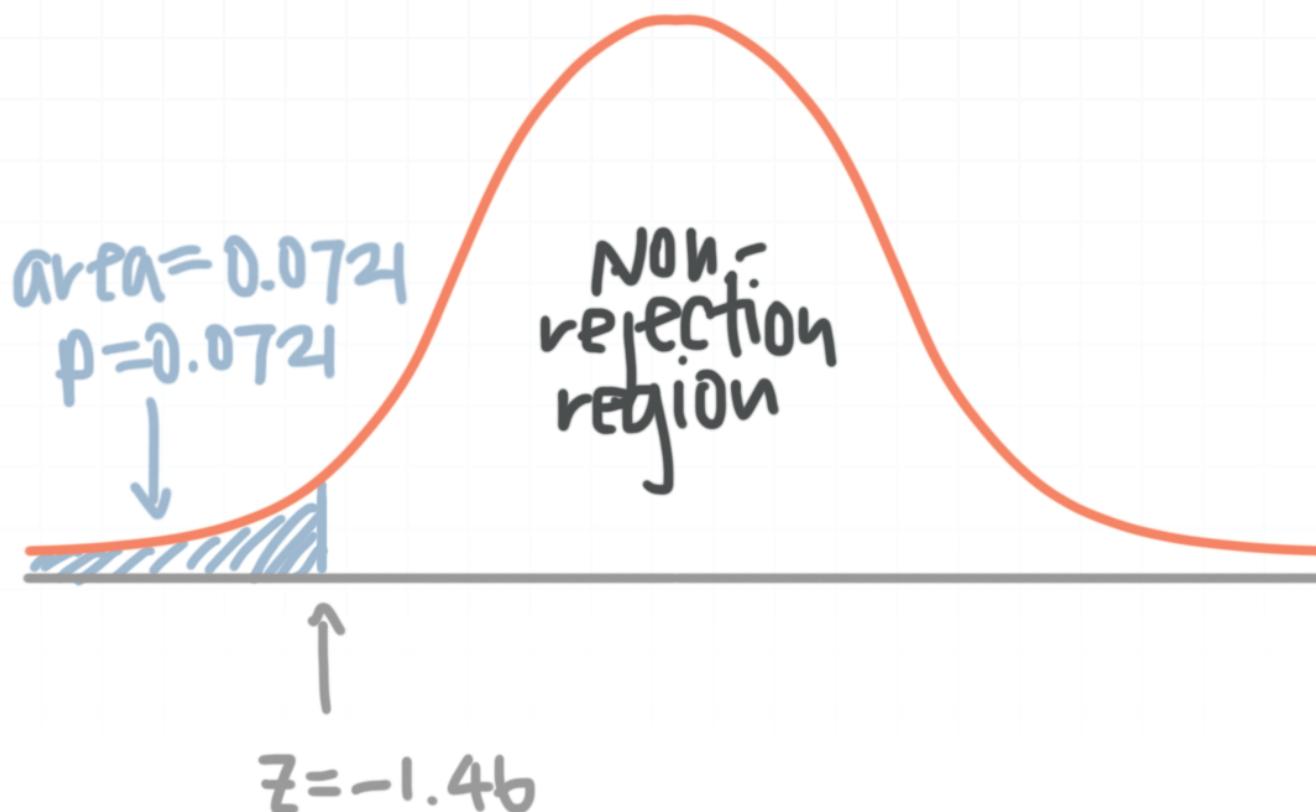
For a one-tailed test, we'll first calculate the  $z$ -test statistic. For a lower-tailed test,  $z$  will be negative. Then the value we find from the negative  $z$ -table represents the area under the probability distribution to the left of the negative  $z$ -value.

For instance, let's assume we calculated a test statistic of  $z = -1.46$ . In a  $z$ -table, we'd find



$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	<b>.0721</b>	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

So 0.0721 is the area under the curve to the left of  $z = -1.46$ , and this is the  $p$ -value also. So  $p = 0.0721$ .



### For a one-tailed, upper-tailed test

For a one-tailed test, we'll first calculate the  $z$ -test statistic. For an upper-tailed test,  $z$  will be positive. Then the value we find from the positive  $z$ -table represents the area under the probability distribution to the left of the positive  $z$ -value.

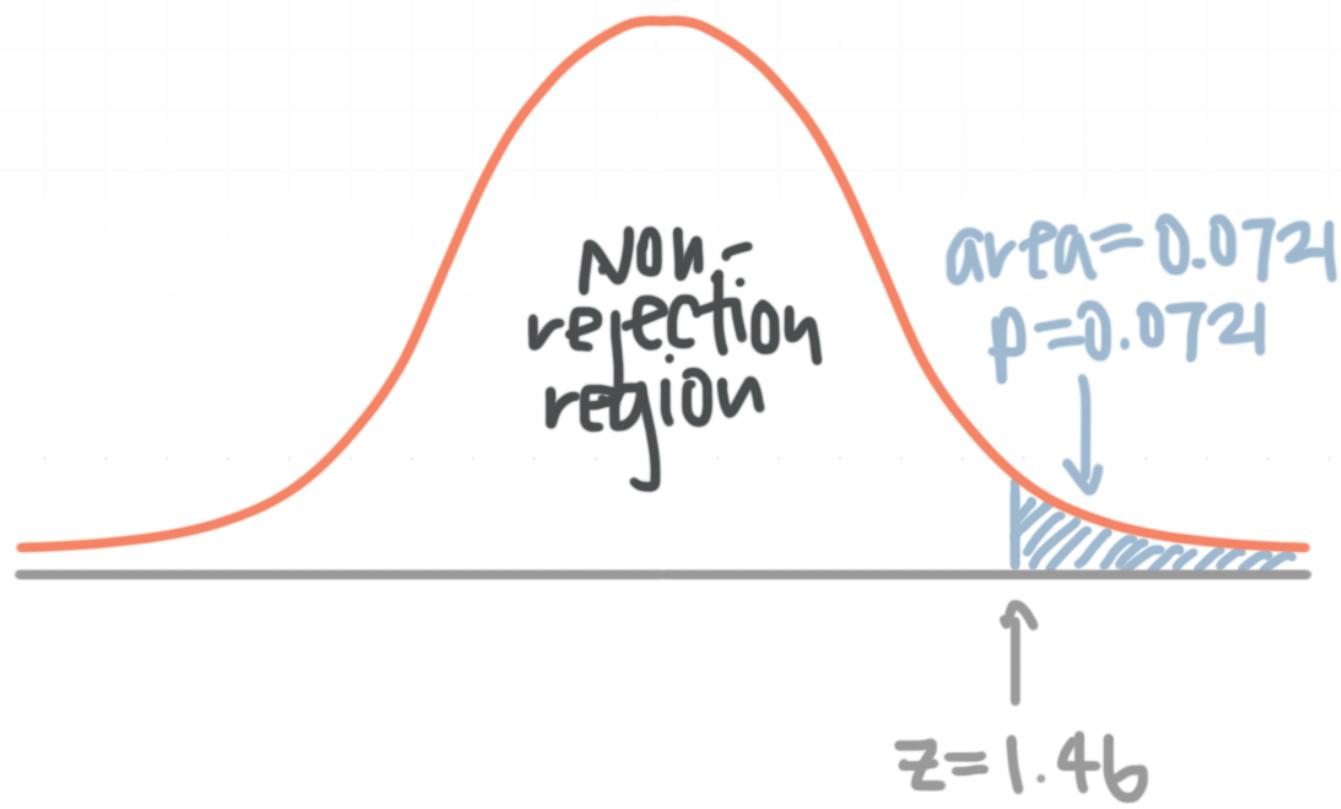
For instance, let's assume we calculated a test statistic of  $z = 1.46$ . In a  $z$ -table, we'd find

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	<b>.9279</b>	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441

But in an upper-tailed test, we're interested in the area to the right of the *z*-value, not the area to the left. To find the area to the right, we need to subtract the value in the *z*-table from 1.

$$1 - 0.9279 = 0.0721$$

So 0.0721 is the area under the curve to the right of  $z = 1.46$ , and this is the *p*-value also. So  $p = 0.0721$ .



### For a two-tailed test

For a two-tailed test, we'll first calculate the *z*-test statistic. For a two-tailed test, *z* could be either positive or negative. Then the value we find from the *z*-table represents the area under the probability distribution to the left of the *z*-value.

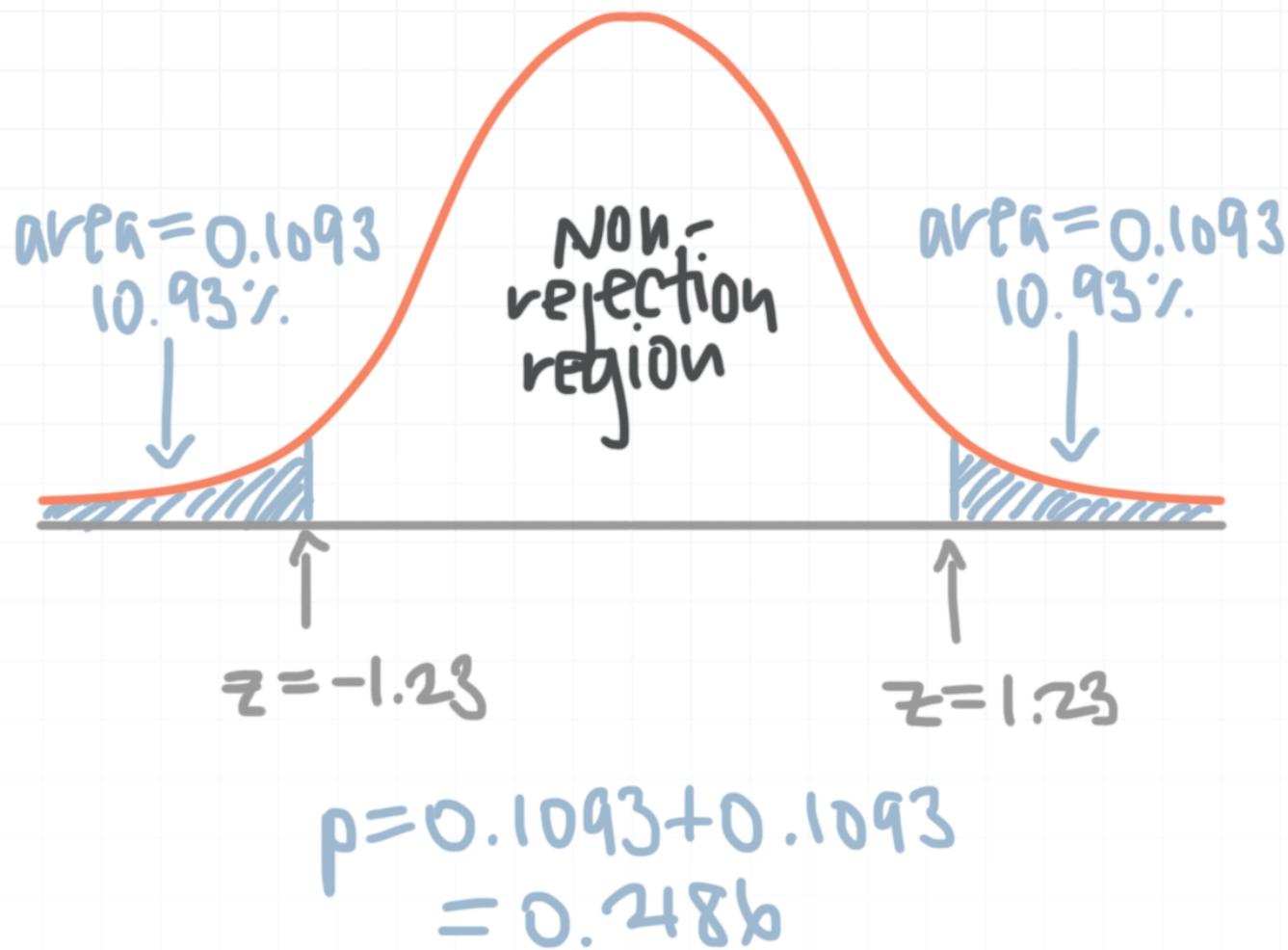
For instance, assume we found  $z = 1.23$ . In a  $z$ -table, we'd find

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

But for a positive  $z$ -value, we're interested in the area to the right of the  $z$ -value, not the area to the left. To find the area to the right, we need to subtract the value in the  $z$ -table from 1.

$$1 - 0.8907 = 0.1093$$

So 0.1093 is the area under the curve to the right of  $z = 1.23$ . Because this is a two-tailed test, the region of rejection is not only the 10.93 % of area in the upper tail, but also the symmetrical 10.93 % of area in the lower tail. So we'll double 0.1093 to get  $2(0.1093) = 0.2186$ , and this is the  $p$ -value also. So  $p = 0.2186$ .



## Rejecting the null using the $p$ -value

The reason we've gone through all this work to understand the  $p$ -value is because using a  $p$ -value is a really quick way to decide whether or not to reject the null hypothesis.

Whether or not we should reject the null hypothesis  $H_0$  can be determined by the relationship between the  $\alpha$  level and the  $p$ -value.

If  $p \leq \alpha$ , reject the null hypothesis

If  $p > \alpha$ , do not reject the null hypothesis

In our earlier examples, we found

$p = 0.0721$  for the lower-tailed one-tailed test

$p = 0.0721$  for the upper-tailed one-tailed test

$p = 0.2186$  for the two-tailed test

With these in mind, let's say for instance we set the confidence level of our hypothesis test at 90 %, which is the same as setting the  $\alpha$  level at  $\alpha = 0.10$ . In that case,

$$p = 0.0721 \leq \alpha = 0.10$$

$$p = 0.2186 > \alpha = 0.10$$

So we would have rejected the null hypothesis for both one-tailed tests, but we would have failed to reject the null hypothesis in the two-tailed test. If, however, we'd picked a more rigorous  $\alpha = 0.05$  or  $\alpha = 0.01$ , we would have failed to reject the null hypothesis every time. So to summarize the  $p$ -value approach when  $\sigma$  is known,

Lower-tailed test

Reject  $H_0$  when  $p \leq \alpha$

Upper-tailed test

Reject  $H_0$  when  $p \leq \alpha$

Two-tailed test

Reject  $H_0$  when  $p \leq \alpha$

Lastly, remember that this entire  $p$ -value approach applies in exactly the same way, whether we're using a  $z$ -test statistic or a  $t$ -test statistic. So when our test statistic is a  $t$ -value, we look up the  $t$ -value in the  $t$ -table. For a lower-tailed test, the value we find becomes the  $p$ -value; for an upper-tailed test, we subtract the value we find from 1 to get the  $p$ -value; and for a two-tailed test, we double the value we find to get the  $p$ -value. Then we use the same rules as the  $z$ -test statistic to compare  $p$  to  $\alpha$ .



to determine whether or not to reject the null hypothesis. So to summarize the  $p$ -value approach when  $\sigma$  is unknown and/or when we have a small sample,

Lower-tailed test

Reject  $H_0$  when  $p \leq \alpha$

Upper-tailed test

Reject  $H_0$  when  $p \leq \alpha$

Two-tailed test

Reject  $H_0$  when  $p \leq \alpha$

## Rejecting the null using the critical value

When deciding whether or not to reject the null hypothesis, we can also determine **critical values**, instead of taking the  $p$ -value approach. Critical values are like cut-off values that bound the rejection region(s). The critical values depend on the test statistic itself,  $z$  or  $t$ , and the level of significance  $\alpha$ .

For example, if we perform an upper-tailed  $z$ -test and the level of significance is  $\alpha = 0.05$ , then we can first look up the probability that's closest to 0.95 in the  $z$ -table (since the test is upper-tailed), and then find the corresponding critical  $z$ -value from the table that's equivalent to 1.96. The region of rejection is to the right of 1.96, so if the test statistic is larger than 1.96, we need to reject the null hypothesis. Otherwise, we accept the null.

So to summarize the critical value approach,

Lower-tailed test

Reject  $H_0$  when  $z \leq -z_\alpha$



Upper-tailed test

Reject  $H_0$  when  $z \geq z_\alpha$

Two-tailed test

Reject  $H_0$  when  $z \leq -z_{\frac{\alpha}{2}}$  or  $z \geq z_{\frac{\alpha}{2}}$

## Significance

The **significance** (or **statistical significance**) of a test is the probability of obtaining the result by chance. The less likely it is that we obtained a result by chance, the more significant our results.

Hopefully by now it's not too surprising that all of these are equivalent statements:

- The finding is significant at the 0.01 level
- The confidence level is 99 %
- The Type I error rate is 0.01
- The alpha level is 0.01,  $\alpha = 0.01$
- The area of the rejection region is 0.01
- The  $p$ -value is 0.01,  $p = 0.01$
- There's a 1 in 100 chance of getting a result as extreme, or more extreme, as than this one

The smaller the  $p$ -value, or the smaller the alpha value, or the lower the Type I error rate, and the smaller the region of rejection, the higher the confidence level, and the less likely it is that we got our result by chance.



So, to take a different example, an alpha level of 0.10 (or a  $p$ -value of 0.10, or a confidence level of 90 %) is a pretty low bar to clear. At that significance level, there's a 1 in 10 chance that the result we got was just by chance. And therefore there's a 1 in 10 chance that we'll reject the null hypothesis when we really shouldn't have, thinking that we provided support for the alternative hypothesis when we shouldn't have.

But a stricter alpha level of 0.01 (or a  $p$ -value of 0.01, or a confidence level of 99 %) is a higher bar to clear. At that significance level, there's only a 1 in 100 chance that the result we got was just by chance. And therefore there's only a 1 in 100 chance that we'll reject the null hypothesis when we shouldn't, thinking that we provided support for the alternative hypothesis when we shouldn't have.

If we find a result that clears the bar we've set for ourselves, then we reject the null hypothesis and we say that the finding is significant at the  $p$ -value that we find. Otherwise, we fail to reject the null.

# Hypothesis testing for the population proportion

Up to now we've been focused mostly on hypothesis testing for the mean, but we can also perform hypothesis testing for a proportion. In order for the test to work, we'll need  $np \geq 5$  and  $n(1 - p) \geq 5$ , where  $p$  is the population proportion (or we can substitute the sample proportion  $\hat{p}$  when we don't know the population proportion  $p$ ).

We want the number of successes and failures to be at least 5, because that threshold means that the probability distribution will approximate the normal curve. Keep in mind that we have to be careful about distinguishing between the sample proportion  $\hat{p}$  and the  $p$ -value of the test.

## One-tailed test

Just like for the mean, a one-tailed test for the proportion indicates directionality, so our hypothesis statements will either be

$$H_0: p \leq k$$

$$H_a: p > k$$

if we suspect that  $p > k$ , or

$$H_0: p \geq k$$

$$H_a: p < k$$

when we suspect that  $p < k$ . Once we set the hypothesis tests, we'll pick a significance level, calculate the test statistic, and state the conclusion.



## Example

We want to test the hypothesis that more than 32 % of Americans watch the Super Bowl, so we collect a random sample of 1,000 Americans and find that 350 of them watched the game. What can we conclude at a significance level of  $\alpha = 0.05$ ?

First build the hypothesis statements.

$H_0$ : At most 32 % of Americans watched the Super Bowl,  $p \leq 0.32$

$H_a$ : More than 32 % of Americans watched the Super Bowl,  $p > 0.32$

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{350}{1,000} = 0.35$$

Then find the standard error of the proportion.

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.32(1 - 0.32)}{1,000}} = \sqrt{\frac{0.2176}{1,000}} \approx 0.0148$$

Now we have enough to find the test statistic.

$$z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{0.35 - 0.32}{0.0148} \approx 2.03$$

The critical value for 95 % confidence with an upper-tailed test is  $z = 1.65$ .



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	<b>.9505</b>	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

Using the critical value approach, we can say that our  $z$ -value exceeds  $z = 1.65$ ,

$$z \approx 2.03 > z_{\frac{\alpha}{2}} = 1.65$$

and therefore falls in the region of rejection, which means we'll reject the null hypothesis and conclude that more than 32% of Americans watch the Super Bowl.

We know our findings are significant at  $\alpha = 0.05$ , but we can find the  $p$ -value to state a higher level of significance that corresponds to  $z \approx 2.03$  and not just  $z = 1.65$ . The test statistic  $z \approx 2.03$  gives a value of 0.9788 in the  $z$ -table.

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	<b>.9788</b>	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

Which means the conclusion isn't only significant at  $\alpha = 0.05$ , but it's actually significant at

$$1 - 0.9788 = 0.0212$$

The result is significant at the 0.0212 level, so as long as  $\alpha \geq 0.0212$ , we'll be able to reject  $H_0$ .



## Two-tailed test

The two-tailed test for the proportion follows the same steps as the one-tailed test, other than the fact that we split the alpha value into both tails.

Let's continue with the same Super Bowl example we were using, but this time we'll say that we don't have a guess about directionality, and instead will simply hypothesize that the proportion of Americans who watch the Super Bowl is some proportion other than 32 % .

### Example (cont'd)

We want to test the hypothesis that 32 % of Americans watch the Super Bowl, so we collect a random sample of 1,000 Americans and find that 350 of them watched the game. What can we conclude at a significance level of  $\alpha = 0.05$ ?

First build the hypothesis statements.

$H_0$ : 32 % of Americans watched the Super Bowl,  $p = 0.32$

$H_a$ : The proportion of Americans who watched the Super Bowl was not 32 % ,  $p \neq 0.32$



We already calculated that the standard error of the proportion is  $\sigma_{\hat{p}} \approx 0.0148$ , the sample proportion is  $\hat{p} = 0.35$ , and the  $z$ -value of the test-statistic is  $z \approx 2.03$ .

Because we're doing a two-tailed test,  $\alpha = 0.05$  needs to be split as 0.025 in the lower tail and 0.025 in the upper tail. Which means we're looking for the value in the  $z$ -table that corresponds to  $1 - 0.025 = 0.9750$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

So  $z = \pm 1.96$  will be the critical values. Our  $z$ -value exceeds  $z = 1.96$  and therefore still falls in the region of rejection (even though we've switched from a one-tailed test to a two-tailed test), which means we'll again reject the null hypothesis and conclude that the proportion of Americans who watch the Super Bowl is some proportion other than 32 %.

# Confidence interval for the difference of means

So far, we've been looking at hypothesis testing for one population, but the goal of many statistical studies is to compare two populations, often in terms of their means.

## The difference of means

For example, maybe we want to compare the effectiveness of two different diet plans. We could define two populations; one group that follows the first diet plan, and a second group that follows the second diet plan.

Then the population means are the mean weight lost on the first diet plan  $\mu_1$ , and the mean weight lost on the second diet plan  $\mu_2$ . We could take a sample from each population, in which case the point estimators are the mean of each sample,  $\bar{x}_1$  and  $\bar{x}_2$ , respectively.

Now let's say that what we're actually interested in is the difference of means,  $\mu_1 - \mu_2$ . In other words, we want to know how much more or less weight we can expect to lose if we follow the first diet plan instead of the second diet plan.

In this lesson, we'll look at how to build a confidence interval around the difference of sample means. By the end of this lesson, we'll want to be able to make a statement like

*"95 % of the confidence intervals I construct around the sample statistic  $\bar{x}_1 - \bar{x}_2$  will contain the population parameter  $\mu_1 - \mu_2$ ,"*



or in simpler terms

*"I'm 95 % that the difference in population means  $\mu_1 - \mu_2$  will fall within the confidence interval  $(\bar{x}_1 - \bar{x}_2) \pm \text{margin of error.}$ "*

Just like before when we were investigating the mean of only one population (by taking only one sample), when we build a confidence interval for the difference of means, we'll use one confidence interval formula when population standard deviations are known, and a different confidence interval formula when population standard deviations are unknown and/or the sample sizes are small  $n_1, n_2 < 30$ .

## With known standard deviations

This won't usually be the case, but let's assume that we know the standard deviation of both populations,  $\sigma_1$  and  $\sigma_2$ . We'll take a sample of size  $n_1$  from the first population, and a sample of size  $n_2$  from the second population.

Then we'll calculate the mean of each sample to find  $\bar{x}_1$  and  $\bar{x}_2$ .

As long as both of the original populations were normally distributed, and/ we take large enough samples  $n_1, n_2 \geq 30$  (so that the Central Limit Theorem kicks in), then the sampling distributions of  $\bar{x}_1$  and  $\bar{x}_2$  will be normally distributed, and therefore the **sampling distribution of the difference of means**  $\bar{x}_1 - \bar{x}_2$  will be normally distributed as well. The mean of the sampling distribution of the difference of means will be

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$$



In other words, the mean of the sampling distribution of the difference of means is the difference of the means of the sampling distributions of the sample means.

The standard error of the sampling distribution of the difference of means will be

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Then the formula for the confidence interval is

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{\bar{x}_1 - \bar{x}_2}$$

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where  $\bar{x}_1 - \bar{x}_2$  is the difference of sample means,  $z_{\alpha/2}$  is the critical  $z$ -value,  $\sigma_1$  and  $\sigma_2$  are the population standard deviations, and  $n_1$  and  $n_2$  are the sample sizes.

Let's work through an example where we calculate the confidence interval around a difference of means.

### Example

A research team is testing the effect of a low-carb diet on people with type 2 diabetes. 400 people are assigned to group 1 and put on a low-carb diet, while another 400 people are assigned to group 2 and put on a standard diet. Given the sample means and population standard



deviations below, estimate a 95 % confidence interval for the difference between the mean drop in blood sugar levels.

**Group 1 (Low carb)**

$$n_1 = 400$$

$$\bar{x}_1 = 9.5 \text{ mg/dL drop}$$

$$\sigma_1 = 0.35 \text{ mg/dL}$$

**Group 2 (Standard)**

$$n_2 = 400$$

$$\bar{x}_2 = 3.2 \text{ mg/dL drop}$$

$$\sigma_2 = 0.28 \text{ mg/dL}$$

At a confidence level of 95 %, we know  $\alpha = 0.05$  and  $\alpha/2 = 0.025$ . Using a  $z$ -table, we need to find the  $z$ -score that corresponds to  $1 - 0.025 = 0.9750$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817

So  $z_{\alpha/2} = 1.96$ , and we can substitute everything we know into the confidence interval formula.

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(a, b) = (9.5 - 3.2) \pm 1.96 \sqrt{\frac{0.35^2}{400} + \frac{0.28^2}{400}}$$

$$(a, b) = 6.3 \pm 1.96 \sqrt{\frac{0.1225}{400} + \frac{0.0784}{400}}$$

$$(a, b) = 6.3 \pm 1.96 \cdot \frac{\sqrt{0.2009}}{20}$$

$$(a, b) \approx 6.3 \pm 0.044$$

$$(a, b) \approx (6.256, 6.344)$$

So we can say with 95 % confidence that the mean drop in blood sugar levels is somewhere between 6.256 mg/dL and 6.344 mg/dL higher in the low-carb diet group than the drop seen in the standard diet group.

---

## With unknown standard deviations and/or small samples

When our population standard deviations  $\sigma_1$  and  $\sigma_2$  are unknown, we can use the sample standard deviations  $s_1$  and  $s_2$  in their place. When we do, we have to consider two possible scenarios.

### Unequal population variances

If the sample variances are significantly unequal, then we assume that the population variances are unequal, and our confidence interval formula will be

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



$$\text{with } df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left( \frac{s_2^2}{n_2} \right)^2}$$

Note: round down when  $df$  is not an integer, so that the estimate is more conservative

### Equal (or almost equal) population variances

If the sample variances are equal or almost equal, then we assume that the population variances are approximately equal as well, and we calculate a pooled variance by combining the two sample variances into one. The formula for **pooled variance** is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and therefore **pooled standard deviation** is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

As a rule of thumb, we can use pooled variance when the two samples were taken from the same population, or when neither sample variance is more than twice the other.

In other words, if we take our samples from the same population, and we have no reason to believe that their variances or standard deviations will be different, then we can use pooled variance and pooled standard



deviation. But even if we take our samples from different populations, if their variances and standard deviations turn out to be close enough in value (neither is more than twice the other), then we can use the pooled formulas.

The standard error of the sampling distribution of the difference of means will change to

$$\sigma_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and our confidence interval formula will be

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

with  $df = n_1 + n_2 - 2$

Let's rework the same example from before, but this time we'll use smaller samples, such that we'll be forced to use one of these  $t$ -score confidence interval formulas.

### Example

A research team is testing the effect of a low-carb diet on people with type 2 diabetes. 25 people are assigned to group 1 and put on a low-carb diet, while another 25 people are assigned to group 2 and put on a standard diet. Given the sample means and sample standard deviations below, estimate a 95 % confidence interval for the difference between the mean drop in blood sugar levels.



**Group 1 (Low carb)**

$$n_1 = 25$$

$$\bar{x}_1 = 9.5 \text{ mg/dL drop}$$

$$s_1 = 0.35 \text{ mg/dL}$$

**Group 2 (Standard)**

$$n_2 = 25$$

$$\bar{x}_2 = 3.2 \text{ mg/dL drop}$$

$$s_2 = 0.28 \text{ mg/dL}$$

Because  $s_1 = 0.35$ , the sample variance of group 1 is  $s_1^2 = 0.35^2 = 0.1225$ ; and because  $s_2 = 0.28$ , the sample variance of group 2 is  $s_2^2 = 0.0784$ . So neither sample variance is more than twice the other, and we can say that the sample variances are approximately equal, and therefore that the population variances are approximately equal. Therefore, we can use the pooled standard deviation formula.

$$s_p = \sqrt{\frac{(25 - 1)0.35^2 + (25 - 1)0.28^2}{25 + 25 - 2}}$$

$$s_p = \sqrt{\frac{24(0.1225) + 24(0.0784)}{48}}$$

$$s_p = \sqrt{\frac{0.1225 + 0.0784}{2}}$$

$$s_p = \sqrt{\frac{0.2009}{2}}$$

$$s_p \approx 0.317$$

The number of degrees of freedom is given by



$$df = 25 + 25 - 2 = 48$$

When we look up these degrees of freedom for a 95 % confidence level in the student's  $t$ -table, we find 2.011. Now we can substitute what we know into the confidence interval formula.

$$(a, b) = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(a, b) \approx (9.5 - 3.2) \pm 2.011 \times 0.317 \sqrt{\frac{1}{25} + \frac{1}{25}}$$

$$(a, b) \approx 6.3 \pm 2.011 \times 0.317 \sqrt{\frac{1}{25} + \frac{1}{25}}$$

$$(a, b) \approx 6.3 \pm 0.637363 \cdot \frac{\sqrt{2}}{5}$$

$$(a, b) \approx 6.3 \pm 0.180$$

$$(a, b) \approx (6.3 - 0.180, 6.3 + 0.180)$$

$$(a, b) \approx (6.12, 6.48)$$

So we can say with 95 % confidence that the mean drop in blood sugar levels is higher in the group that was on the low-carb diet than in the group that was on a standard diet, and that the difference between means will fall between 6.12 and 6.48 mg/dL.



# Hypothesis testing for the difference of means

Now that we know how to build a confidence interval around the difference of means, let's work through the entire hypothesis testing procedure when we want to use the difference of sample means to make an inference about the difference of population means.

## Building hypothesis statements

The null and alternative hypotheses will always be formulated in terms of the difference between the two population means,  $\mu_1 - \mu_2$ , and we can have three different scenarios.

In a two-tailed test, the null hypothesis will state that the means don't differ, whereas the alternative hypothesis states that there *is* a difference between means. So we write the hypothesis statements for a two-tailed test as

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

or

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

In an upper-tailed test, the alternative hypothesis states that the difference in means is positive, so we write



$$H_0 : \mu_1 - \mu_2 \leq 0$$

$$H_a : \mu_1 - \mu_2 > 0$$

or

$$H_0 : \mu_1 \leq \mu_2$$

$$H_a : \mu_1 > \mu_2$$

In a lower-tailed test, the alternative hypothesis states that the difference in means is negative, so we write

$$H_0 : \mu_1 - \mu_2 \geq 0$$

$$H_a : \mu_1 - \mu_2 < 0$$

or

$$H_0 : \mu_1 \geq \mu_2$$

$$H_a : \mu_1 < \mu_2$$

## Calculating the test statistic

### Large samples, unequal population variances

If the independent random samples we take from each population are both large enough,  $n_1, n_2 \geq 30$ , and the population variances are unequal, then the test statistic formula we'll use is



$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When our hypothesis statements only test for a difference of means, then  $\mu_1 - \mu_2 = 0$ , and the test statistic formula simplifies to

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

### Large samples, equal population variances

On the other hand, if the independent random samples we take from each population are both large enough,  $n_1, n_2 \geq 30$ , but our population variances are reasonably equal, then we use the formula for pooled variance,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

and the formula for the test statistic becomes

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

or when  $\mu_1 - \mu_2 = 0$ , the formula simplifies to

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



## Small sample(s), unequal population variances

If one or both of our independent random samples is/are small,  $n_1 < 30$  and/or  $n_2 < 30$ , and the population variances are unequal, then the test statistic formula we'll use is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

or when  $\mu_1 - \mu_2 = 0$ , the formula simplifies to

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When we calculate degrees of freedom, we'll use

$$df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left( \frac{s_2^2}{n_2} \right)^2}$$

If we find a non-integer value for degrees of freedom, we should always round down to the next lowest integer so that the estimate is more conservative.

## Small sample(s), equal population variances

On the other hand, if one or both of the independent random samples we take from each population is/are small,  $n_1 < 30$  and/or  $n_2 < 30$ , and the



population variances are reasonably equal, then we use the formula for pooled variance,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

and the formula for the test statistic becomes

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

or when  $\mu_1 - \mu_2 = 0$ , the formula simplifies to

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

When we calculate degrees of freedom, we'll use  $df = n_1 + n_2 - 2$ .

## Making a conclusion

Once we've calculated a test statistic, we'll use the significance level  $\alpha$ , along with either the  $p$ -value approach or the critical value approach, to determine whether or not we can reject the null hypothesis.

Let's work through an example so that we can see how this works.

### Example



Suppose we want to determine whether the mean height of men is significantly higher than the mean height of women in a certain city, so we randomly sample 100 men and 100 women. Given the mean and standard deviation of both samples below, use the critical value approach to say whether men are significantly taller than women at a 1% level of significance.

**Men**

$$n_1 = 100$$

$$\bar{x}_1 = 69.5 \text{ inches}$$

$$s_1 = 1.25 \text{ inches}$$

**Women**

$$n_2 = 100$$

$$\bar{x}_2 = 67.8 \text{ inches}$$

$$s_2 = 1.12 \text{ inches}$$

Since we want to test the claim the the mean height of men is higher than the mean height of women, our hypothesis statements will be

$$H_0 : \mu_M - \mu_W \leq 0$$

$$H_a : \mu_M - \mu_W > 0$$

Because the sample standard deviations are  $s_1 = 1.25$  and  $s_2 = 1.12$ , the sample variances are  $s_1^2 = 1.25^2 = 1.5625$  and  $s_2^2 = 1.12^2 = 1.2544$ . The sample variance 1.5625 isn't more than twice the sample variance 1.2544, which means we can assume that the sample variances are reasonably equal, and therefore that the population variances are reasonably equal, so we'll use pooled variance.



$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$s_p = \sqrt{\frac{(100 - 1)1.25^2 + (100 - 1)1.12^2}{100 + 100 - 2}}$$

$$s_p = \sqrt{\frac{99(1.5625) + 99(1.2544)}{198}}$$

$$s_p = \sqrt{\frac{1.5625 + 1.2544}{2}}$$

$$s_p = \sqrt{\frac{2.8169}{2}}$$

$$s_p \approx 1.187$$

Now calculate the  $z$ -test statistic.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$z = \frac{69.5 - 67.8}{1.187 \sqrt{\frac{1}{100} + \frac{1}{100}}}$$

$$z = \frac{1.7}{1.187 \cdot \frac{\sqrt{2}}{10}}$$

$$z \approx 10.13$$

Because we want to test at a significance level of 1%, our confidence level is  $1 - \alpha = 1 - 0.01 = 0.99$ . We're using a right-tailed test, so we need to use the  $z$ -table to find the  $z$ -score that corresponds to the probability 0.99. In a  $z$ -table, we find  $z = 2.33$ .

$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	<b>.9901</b>	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

Therefore, if the  $z$ -test statistic that we've found is larger than 2.33, we would reject the null hypothesis. Since  $10.13 > 2.33$ , we can reject the null hypothesis at  $\alpha = 0.01$  and conclude that the mean height of men is greater than the mean height of women.

Let's use another example to illustrate the use of a  $t$ -test for the difference of means when the variances are unequal.

### Example

A company believes its new light bulb will last at least 30 days longer than its old bulb. They take random samples of 20 old bulbs and 20 new bulbs, and find the following:

#### New bulb

$$n_1 = 20$$

$$\bar{x}_1 = 254 \text{ days}$$

#### Old bulb

$$n_2 = 20$$

$$\bar{x}_2 = 205 \text{ days}$$

$$s_1 = 5 \text{ days}$$

$$s_2 = 13 \text{ days}$$

At a 0.05 level of significance, test the claim that the new bulb lasts at least 30 days longer than the old bulb.

Given  $\mu_1$  as the life expectancy of the new bulb, and  $\mu_2$  as the life expectancy of the old bulb, our hypothesis statements for the right-tailed test will be

$$H_0 : \mu_1 - \mu_2 \leq 30$$

$$H_a : \mu_1 - \mu_2 > 30$$

Because the sample standard deviations are  $s_1 = 5$  and  $s_2 = 13$ , the sample variances are  $s_1^2 = 5^2 = 25$  and  $s_2^2 = 13^2 = 169$ . We can see that  $s_2^2$  is more than double  $s_1^2$ , so we can assume the sample variances are unequal, and therefore that the population variances are unequal. Because of this, and the fact that we have small samples,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where  $\mu_1 - \mu_2 = 30$ . Substitute what we know to calculate the  $t$ -test statistic.

$$t = \frac{(254 - 205) - 30}{\sqrt{\frac{5^2}{20} + \frac{13^2}{20}}}$$



$$t = \frac{19}{\sqrt{\frac{25}{20} + \frac{169}{20}}}$$

$$t = \frac{19}{\sqrt{\frac{194}{20}}}$$

$$t = 38\sqrt{\frac{5}{194}}$$

$$t \approx 6.101$$

The degrees of freedom will be

$$\text{df} = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left( \frac{s_2^2}{n_2} \right)^2}$$

$$\text{df} = \frac{\left( \frac{5^2}{20} + \frac{13^2}{20} \right)^2}{\frac{1}{20 - 1} \left( \frac{5^2}{20} \right)^2 + \frac{1}{20 - 1} \left( \frac{13^2}{20} \right)^2}$$

$$\text{df} = \frac{\left( \frac{25}{20} + \frac{169}{20} \right)^2}{\frac{1}{19} \left( \frac{25}{20} \right)^2 + \frac{1}{19} \left( \frac{169}{20} \right)^2}$$

$$\text{df} = \frac{\frac{37,636}{400}}{\frac{625}{7,600} + \frac{28,561}{7,600}}$$



$$df = \frac{37,636}{400} \left( \frac{7,600}{29,186} \right)$$

$$df = \frac{715,084}{29,186}$$

$$df \approx 24.501$$

Always round down for a more conservative estimate.

$$df \approx 24$$

Now find the critical  $t$ -value from the  $t$ -table using  $1 - \alpha = 0.95$  and  $df = 24$ . Because we're running an upper-tailed test, the whole region of rejection is consolidated into the upper tail, which means we're looking at upper-tail probability of 0.05.

df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

The critical  $t$ -value is 1.711. Using the critical value approach, we compare 6.101 to 1.711. Because  $6.101 > 1.711$ , we reject the null hypothesis and conclude that the new bulb lasts at least 30 days longer than the old bulb.

# Matched-pair hypothesis testing

We've recently been looking at hypothesis testing for the difference of means when we take two independent samples from one or two populations. Technically, we say that we have **independent samples** when there's no relationship between the observations we find for each sample.

But sometimes we'll want to run a hypothesis test on the difference of means between **dependent samples**, which are samples for which the observations from one sample are related to an observation from the other sample.

## Matched-pair tests

When we do hypothesis testing with dependent samples, we often call it a **matched-pair test**, because each subject in the second sample matches with a particular subject in the first sample.

It's common to run a matched-pair test that compares some new technique or method to an old one, or looks at a before-and-after change.

For instance, a weight-loss study could define Population 1 as the set of starting weights for each participant, and Population 2 as the set of ending weights for each participant. Each participant's starting and ending weights (from Populations 1 and 2, respectively) form a matched-pair for that individual.

In this example, there's an advantage to using a matched-pair test, instead of a difference of means test with independent samples. If we took the



independent samples approach, sample 1 could be taken from the population before the weight loss study begins, and sample 2 could be taken from the population after the weight loss study ends. This approach introduces extra variability unnecessarily because we'll get different people in both samples.

But if we take the matched-pair approach, we keep the people the same across both samples, creating a matched-pair of each person's starting and ending weights.

In general, hypothesis testing with dependent samples will follow a really similar process as the one we've used for the difference of means with independent samples, except that we'll create one variable as the difference between the two samples, and we'll perform the hypothesis test with just this one variable, instead of with two variables.

Let's work through an example so that we can see how to use dependent samples in a matched-pair hypothesis test.

### Example

A fast food restaurant is implementing new workplace policies with the goal of increasing employee satisfaction by 2 points on a scale of 1 to 10. The restaurant surveys 10 employees, asking them both before and after the policies are enacted to rate their workplace satisfaction on the 1 – 10 scale, and records the results in the table below.



Employee	1	2	3	4	5	6	7	8	9	10
Before $x_1$	3	3	5	7	1	0	2	6	6	5
After $x_2$	3	6	9	7	3	5	5	5	9	9
Difference, $d$	0	3	4	0	2	5	3	-1	3	4
$d^2$	0	9	16	0	4	25	9	1	9	16

Can the restaurant say at 5% significance that the policies increased employee satisfaction by 2 points?

The restaurant will define the “before” responses as Population 1, and the “after” responses as Population 2. The samples are dependent because it’s reasonable to see how an employee’s “after” response could be affected by their “before” response.

Then their null and alternative hypotheses will be

$$H_0 : \mu_2 - \mu_1 \leq 2$$

$$H_a : \mu_2 - \mu_1 > 2$$

where  $\mu_1$  is the mean employee satisfaction before the new workplace policies are implemented, and  $\mu_2$  is the mean employee satisfaction after the new workplace policies are implemented. And because  $\mu_2 - \mu_1$  is the difference in employee ratings, the hypothesis statements could also be written as

$$H_0 : \mu_d \leq 2$$

$$H_a : \mu_d > 2$$

where  $\mu_d$  is the mean difference between the two populations.

To find the mean difference, we'll sum the differences and divide by the number of matched-pairs in our sample,  $n = 10$ .

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{0 + 3 + 4 + 0 + 2 + 5 + 3 + (-1) + 3 + 4}{10} = \frac{23}{10} = 2.3$$

So the sample mean tells us that employee satisfaction increases by about 2.3 on a scale of 1 to 10. Then the sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

To calculate this, we'll first find

$$\sum_{i=1}^n (d_i - \bar{d})^2$$

$$(0 - 2.3)^2 + (3 - 2.3)^2 + (4 - 2.3)^2 + (0 - 2.3)^2 + (2 - 2.3)^2$$

$$+(5 - 2.3)^2 + (3 - 2.3)^2 + (-1 - 2.3)^2 + (3 - 2.3)^2 + (4 - 2.3)^2$$

$$(-2.3)^2 + 0.7^2 + 1.7^2 + (-2.3)^2 + (-0.3)^2 + 2.7^2 + 0.7^2 + (-3.3)^2 + 0.7^2 + 1.7^2$$

$$5.29 + 0.49 + 2.89 + 5.29 + 0.09 + 7.29 + 0.49 + 10.89 + 0.49 + 2.89$$

$$36.1$$

Then the sample standard deviation is



$$s_d = \sqrt{\frac{36.1}{9}}$$

$$s_d \approx \sqrt{4.011}$$

$$s_d \approx 2.003$$

Because the population standard deviations are unknown, and/or because both sample sizes are small,  $n_1, n_2 < 30$ , the test statistic will be

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

$$t \approx \frac{2.3 - 2}{\frac{2.003}{\sqrt{10}}}$$

$$t \approx 0.3 \cdot \frac{\sqrt{10}}{2.003}$$

$$t \approx 0.474$$

and the degrees of freedom are

$$df = n - 1 = 10 - 1 = 9$$

At a significance level of 5% (a confidence level of 95%) for an upper-tail test, and  $df = 9$ , the  $t$ -table gives 1.833.



df	Upper-tail probability p									
	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

The restaurant's  $t$ -test statistic  $t \approx 0.474$  doesn't meet the threshold  $t = 1.833$ , so the critical value approach tells them that they can't reject the null hypothesis, and therefore can't conclude that the new workplace policies increased employee satisfaction by 2 points.

## Confidence intervals for matched-pair tests

If the restaurant from the previous example had known the population standard deviation  $\sigma_d$ , they could have calculated a confidence interval around the difference  $\bar{d}$  using

$$(a, b) = \bar{d} \pm z_{\alpha/2} \sigma_{\bar{d}}$$

$$(a, b) = \bar{d} \pm z_{\alpha/2} \frac{\sigma_d}{\sqrt{n}}$$

If, instead, the restaurant had an unknown population standard deviation  $\sigma_d$  and/or a small sample  $n < 30$ , to find a confidence interval around the difference  $\bar{d}$  they would have used

$$(a, b) = \bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}} \text{ with df} = n - 1$$

Let's continue with the previous example in order to calculate the confidence interval.

---

### Example (cont'd)

Find a 95% confidence interval around  $\bar{d}$  using the information in the previous example.

From the previous example, we see that population standard deviation  $\sigma_d$  is unknown, and we have a small sample  $n = 10 < 30$ , so we'll calculate the confidence interval as

$$(a, b) = \bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

$$(a, b) \approx 2.3 \pm 2.262 \cdot \frac{2.003}{\sqrt{10}}$$

$$(a, b) \approx 2.3 \pm 1.433$$

So the margin of error is 1.433 and the confidence interval is

$$(a, b) \approx (2.3 - 1.433, 2.3 + 1.433)$$

$$(a, b) \approx (0.867, 3.733)$$



Therefore, there's a 95 % chance that the change in employee satisfaction changes between 0.867 points and 3.733 points.

---



# Confidence interval for the difference of proportions

In the same way that we can run a hypothesis test on the difference of means, we can do hypothesis testing on the difference of proportions. Before we work through the entire hypothesis test though, let's start with simply building a confidence interval around the difference of proportions.

## The point estimator and standard error

Imagine that a department store wants to know whether the proportion of walk-in customers who complete a purchase at its New York store is different than the same proportion at its San Francisco location.

They could define Population 1 as all of their New York customers and Population 2 as all of their San Francisco customers. They're trying to estimate the difference between  $p_1$  and  $p_2$  (the proportion of walk in customers who complete a purchase in New York and San Francisco, respectively), which means they're looking for  $p_1 - p_2$ . To do so, they can take a sample of size  $n_1$  in New York and a sample of size  $n_2$  in San Francisco, and find

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}$$

where  $x_1$  and  $x_2$  are the number of “successes” in samples  $n_1$  and  $n_2$ , respectively. Then the point estimator of  $p_1 - p_2$  is  $\hat{p}_1 - \hat{p}_2$ . Then the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  will have a mean of  $p_1 - p_2$  and a standard error of



$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

and will be normal as long as  $n_1 p_1 \geq 5$ ,  $n_1(1-p_1) \geq 5$ ,  $n_2 p_2 \geq 5$ , and  $n_2(1-p_2) \geq 5$ .

## Confidence interval around the difference of proportions

The confidence interval around the point estimator  $\hat{p}_1 - \hat{p}_2$  will be given by  $(\hat{p}_1 - \hat{p}_2) \pm \text{margin of error}$ , where the margin of error is

$$z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

So the confidence interval formula is

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Let's work through an example so that make sure we know how to calculate the confidence interval for the difference of proportions.

### Example

A team of scientists wants to determine whether a new cholesterol lowering drug is more effective than a previous version. They take two random samples of 250 people. For three months, the first group gets the new drug while the second group gets the old drug. 155 people from the



first group and 107 people from the second group show decreased cholesterol levels. Estimate a 99 % confidence interval for the difference of proportions.

We know  $n_1 = 250$  and  $n_2 = 250$ , that  $z_{\alpha/2} = 2.58$  for a 99 % confidence level, and that the sample proportions are

$$\hat{p}_1 = \frac{155}{250} = 0.620$$

$$\hat{p}_2 = \frac{107}{250} = 0.428$$

Substituting all these values into the confidence interval formula gives

$$(a, b) = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$(a, b) = (0.620 - 0.428) \pm 2.58 \sqrt{\frac{0.620(1 - 0.620)}{250} + \frac{0.428(1 - 0.428)}{250}}$$

$$(a, b) = 0.192 \pm 2.58 \sqrt{\frac{0.620(0.38)}{250} + \frac{0.428(0.572)}{250}}$$

$$(a, b) = 0.192 \pm 2.58 \sqrt{\frac{0.2356}{250} + \frac{0.244816}{250}}$$

$$(a, b) = 0.192 \pm 2.58 \sqrt{\frac{0.480416}{250}}$$

Simplify to find the confidence interval.

$$(a, b) \approx 0.192 \pm 0.113$$

$$(a, b) \approx (0.192 - 0.113, 0.192 + 0.113)$$

$$(a, b) \approx (0.079, 0.305)$$

$$(a, b) \approx (0.08, 0.31)$$

With 99 % confidence, we can say that the new drug was associated with a greater decrease in cholesterol than the old drug, and that the difference of proportions lies between 0.08 and 0.31.

---

## When the confidence interval contains 0

When the confidence interval we calculate contains 0, such that the lower end of the confidence interval is negative and the upper end of the confidence interval is positive, there's likely no difference in proportions.

On the other hand, when the confidence interval doesn't contain 0, like in this last example where  $(a, b) \approx (0.08, 0.31)$ , then it's likely that there *is* a difference in proportions.



# Hypothesis testing for the difference of proportions

Now that we know how to find a confidence interval around the difference of proportions, let's look at how to conduct a hypothesis test with the difference of proportions, when we want to use the difference of sample proportions to make an inference about the difference of population proportions.

## Building hypothesis statements

The null and alternative hypotheses will always be formulated in terms of the difference between the two population proportions,  $p_1 - p_2$ , and we can have three different scenarios.

In a two-tailed test, the null hypothesis will state that the proportions don't differ, whereas the alternative hypothesis states that there *is* a difference between proportions. So we write the hypothesis statements for a two-tailed test as

$$H_0 : p_1 - p_2 = 0$$

$$H_a : p_1 - p_2 \neq 0$$

or

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$



In an upper-tailed test, the alternative hypothesis states that the difference in proportions is positive, so we write

$$H_0 : p_1 - p_2 \leq 0$$

$$H_a : p_1 - p_2 > 0$$

or

$$H_0 : p_1 \leq p_2$$

$$H_a : p_1 > p_2$$

In a lower-tailed test, the alternative hypothesis states that the difference in means is negative, so we write

$$H_0 : p_1 - p_2 \geq 0$$

$$H_a : p_1 - p_2 < 0$$

or

$$H_0 : p_1 \geq p_2$$

$$H_a : p_1 < p_2$$

## Calculating the test statistic

As long as we take independent random samples from each population, and  $n_1\hat{p}_1 \geq 5$ ,  $n_1(1 - \hat{p}_1) \geq 5$ ,  $n_2\hat{p}_2 \geq 5$ , and  $n_2(1 - \hat{p}_2) \geq 5$ , then the test statistic formula we'll use is



$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where  $\hat{p}_1$  and  $\hat{p}_2$  are the sample proportions,  $p_1$  and  $p_2$  are the population proportions,  $n_1$  and  $n_2$  are the sample sizes, and  $\hat{p}$  is the proportion of the combined sample, given by

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

which we can also write as

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

where  $x_1$  and  $x_2$  are the number of “successes” in each sample. We say that the null hypothesis always states a zero difference between population proportions, such that  $p_1 - p_2 = 0$ , so the test statistic formula actually simplifies to

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Let's rework the same example from the previous section.

### Example

A team of scientists claims that a new cholesterol lowering drug is more effective than an older version. The team takes two random samples of 250



people, and for 3 months administer the new drug to the first group and the old drug to the second group. 120 people in the first group and 107 people in the second group show decreased cholesterol levels. Can the team conclude at a 99 % confidence level that the new drug is more effective than the old drug at lowering cholesterol?

If  $p_1$  is the proportion of population 1 (the population that takes the new drug) whose cholesterol decreases, and  $p_2$  is the proportion of population 2 (the population that takes the old drug) whose cholesterol decreases, then the null and alternative hypotheses are

$$H_0 : p_1 - p_2 \leq 0$$

$$H_a : p_1 - p_2 > 0$$

The pooled proportion is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\hat{p} = \frac{120 + 107}{250 + 250}$$

$$\hat{p} = \frac{227}{500}$$

$$\hat{p} = 0.454$$

and the sample proportions are

$$\hat{p}_1 = \frac{120}{250} = 0.480$$

$$\hat{p}_2 = \frac{107}{250} = 0.428$$

So the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$z = \frac{0.480 - 0.428}{\sqrt{0.454(1 - 0.454)\left(\frac{1}{250} + \frac{1}{250}\right)}}$$

$$z = \frac{0.052}{\sqrt{0.454(0.546)\left(\frac{1}{125}\right)}}$$

$$z = \frac{0.052}{\sqrt{\frac{0.247884}{125}}}$$

$$z = 0.052\sqrt{\frac{125}{0.247884}}$$

$$z \approx 1.17$$

Now we need to determine the critical  $z$ -value. Our level of significance is  $\alpha = 0.01$ , and for a right-tailed test the corresponding  $z$ -value is  $z = 2.33$ .

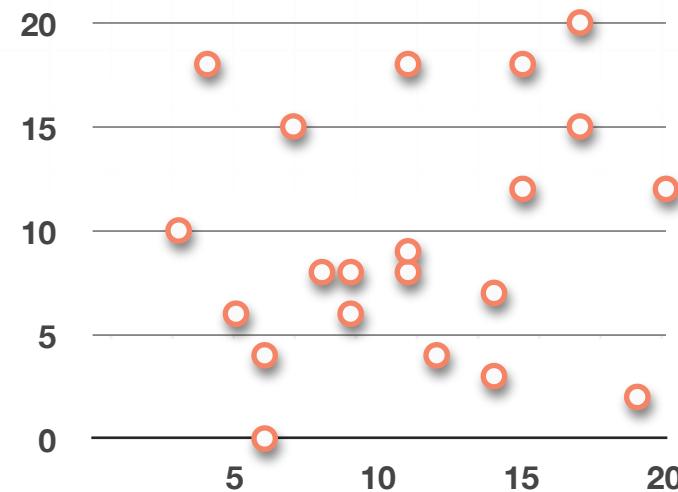
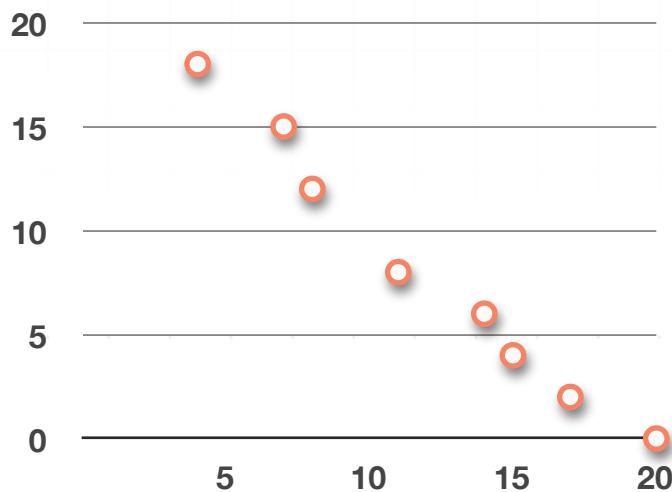
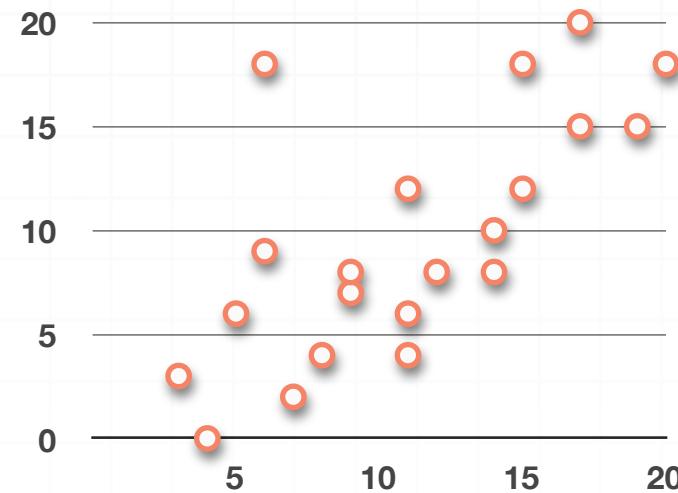
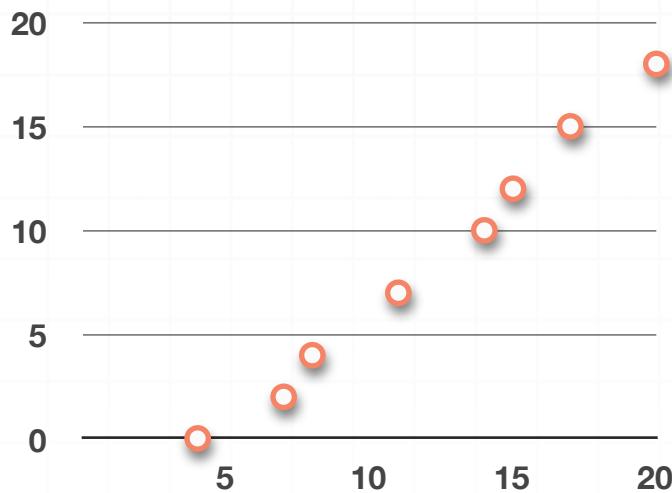


Using the critical-value approach, we'll therefore reject  $H_0$  if  $z \geq 2.33$ . Since  $1.17 \not\geq 2.33$ , the team can't reject the null hypothesis. Therefore, they're unable to provide support for the hypothesis that the new drug is more affective than the old drug.

---

# Scatterplots and regression

A **scatterplot**, also called a **scattergraph** or **scatter diagram**, is a plot of the data points in a set. It plots data that takes two variables into account at the same time. Here are some examples of scatterplots:



Even though scatterplots can look like a mess, sometimes we're able to see trends in the data. For example, the two graphs on the left definitely seem to be roughly following a line: the one on top looks like it follows a line with a positive slope; the bottom one looks like it follows a line with a negative slope.

The graph in the upper right looks like it might be following a positively-sloped line, but if it is, the trend is not as clear as either of the graphs on the left.

And the graph in the lower right doesn't look like it's following any trend at all.

When we say that the data in a scatterplot appears to follow a trend, what we're really saying is that it appears to follow some line, or maybe some other kind of curve, like for example an exponential curve or sinusoidal curve. No matter the shape of the curve that the data follows, we call it the **approximating curve**, and the process of finding the equation of the approximating curve is called **curve fitting**.

The regression line is one of the most important approximating curves we'll talk about, so let's take a look at that now.

## Regression line

It was intuitive for us to start looking for trends in the scatterplots as soon as we saw the plotted points. And, in fact, spotting trends is probably what we spend most of our time doing when we work with scatterplots. The plot alone isn't super helpful, but if we can use the plot to observe some kind of a trend in the data, then we might be able to use that trend to draw conclusions or make predictions about the data.

The most common way that we'll do this is with a **regression line**. It's the line that best shows the trend in the data given in a scatterplot. A



regression line is also called the **best-fit line**, **line of best fit**, or **least-squares line**.

The regression line is a trend line we use to model a linear trend that we see in a scatterplot, but realize that some data will show a relationship that isn't necessarily linear. For example, the relationship might follow the curve of a parabola, in which case the regression curve would be parabolic in nature. For the rest of this lesson we'll focus mostly on linear regression.

## Equation of the regression line

There are a few ways to calculate the equation of the regression line. The equation for a regression line is most often given in slope-intercept form,  $\hat{y} = a + bx$ , where  $x$  is the independent, or explanatory variable,  $b$  is the slope, and  $a$  is the intercept when  $x = 0$ , or the  $y$ -intercept. Sometimes we'll see the regression equation written as  $\hat{y} = \beta_0 + \beta_1x$ , but this is exactly the same thing since  $\beta_0 + \beta_1$  is the same as  $a + b$ . The regression line formula then calculates the slope  $b$  and the  $y$ -intercept  $a$  using

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n}$$

In the formulas for the slope  $b$  and  $y$ -intercept  $a$ ,

$n$  is the number of data points in the set,

$\sum xy$  is the sum of all the products of the  $x$  and  $y$ ,

$\sum x$  is the sum of all the  $x$ -values,

$\sum y$  is the sum of all the  $y$ -values,

$\sum x^2$  is the sum of all the squared  $x$ -values, and

$(\sum x)^2$  is the square of the sum of all the  $x$ -values.

Once we find the equation of the regression line, we denote it with  $\hat{y}$ , (pronounced “y-hat”), to indicate that it’s a regression line, and remind us that it’s an approximation for the data set. So the equation of the regression line is

$$\hat{y} = a + bx$$

Let’s work through an example of how to find the equation of the regression line.

### Example

Find the least-squares line for the data set.

x	y
0	0.8
2	1.0
4	0.2
6	0.2
8	2.0
10	0.8
12	0.6

We'll start by calculating the slope,  $m$ . There are 7 data points in this set, so  $n = 7$ . It can be helpful to calculate  $xy$  and  $x^2$  for each data point, plus find the sum of the  $x$ -values and the sum of the  $y$ -values, and add all of these into the data table, since we'll be using them in our calculations. Our new table that includes this extra information will be

	x	y	xy	$x^2$
	0	0.8	0	0
	2	1.0	2	4
	4	0.2	0.8	16
	6	0.2	1.2	36
	8	2.0	16	64
	10	0.8	8	100
	12	0.6	7.2	144
<b>Sum:</b>	42	5.6	35.2	364

Let's plug what we've found into the formula for slope.

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{7(35.2) - (42)(5.6)}{7(364) - (42)^2}$$

$$b = \frac{246.4 - 235.2}{2,548 - 1,764}$$

$$b = \frac{11.2}{784}$$

$$b \approx 0.0143$$

Now let's plug what we've found into the formula for the  $y$ -intercept.

$$a = \frac{\sum y - b \sum x}{n}$$

$$a = \frac{5.6 - \frac{11.2}{784}(42)}{7}$$

$$a = \frac{5.6 - 0.6}{7}$$

$$a = \frac{5}{7}$$

$$a \approx 0.7143$$

In statistics, we usually write the slope and  $y$ -intercept to the ten-thousandths place (four decimal places) to prevent severe rounding errors

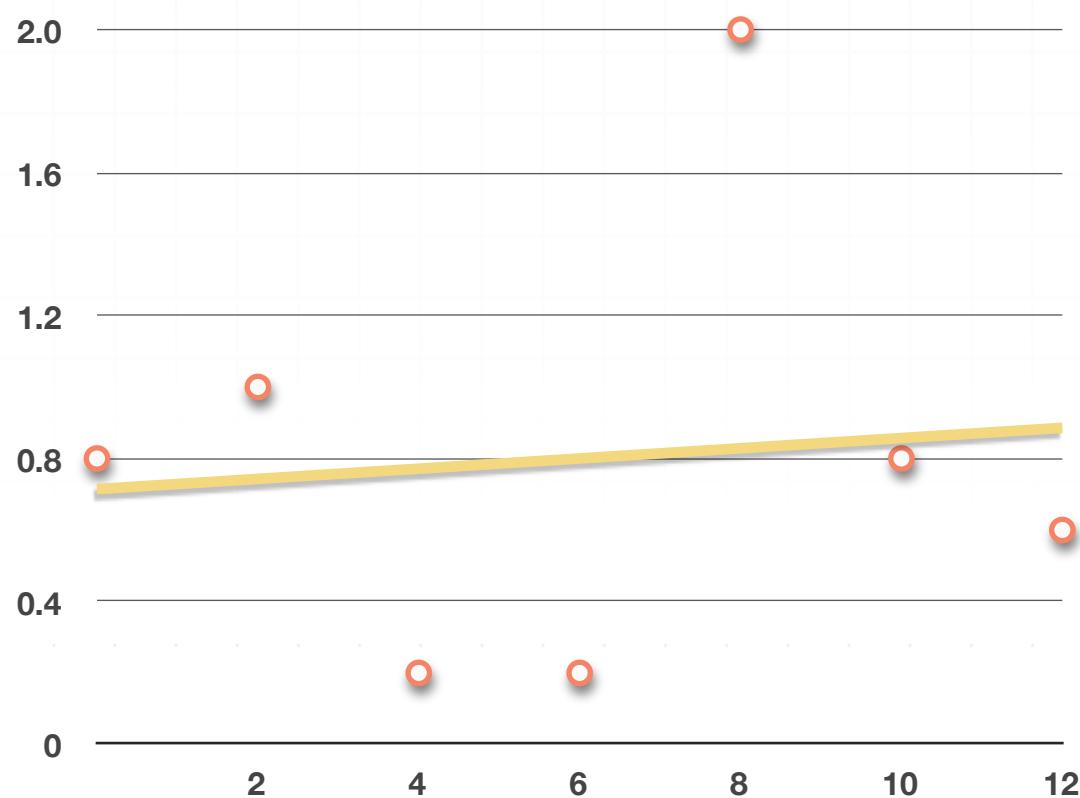


from occurring in the estimated values, and giving us an inaccurate regression line. Therefore, we can say that the regression line is given approximately by

$$\hat{y} = a + bx$$

$$\hat{y} = 0.0143x + 0.7143$$

Let's plot the data points on a scatterplot and then add in the regression line we found to double-check ourselves.



The regression line looks like it runs roughly through the data, indicating the trend.

---

With this last example, we might notice that the data actually wasn't super linear. If we look at the scatterplot we made, we might even say it has

more of a sinusoidal shape, and we can see that the point around  $x = 8$  looks like an outlier.

So the next question we need to answer starts to become obvious, and that is “Is the regression line even a good estimate of the data?” Luckily, there are ways to measure how good of a fit this line is to the data points, and we’ll look at those techniques a little bit later on.

## Describing the trend

Whenever we describe a relationship in the data, we should describe

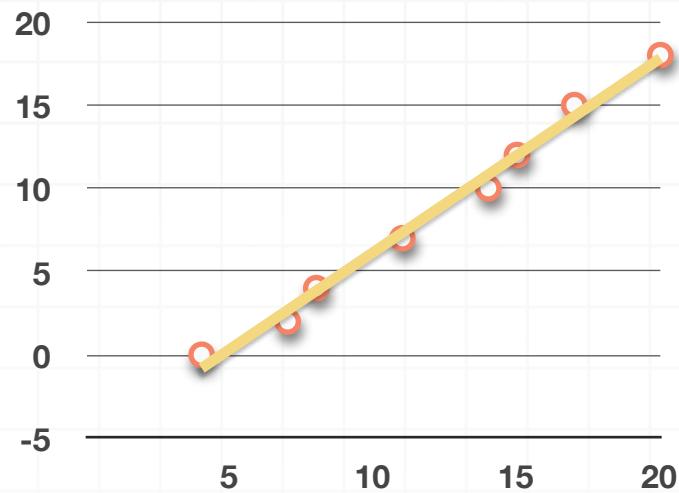
- the form (linear, parabolic, sinusoidal, etc.),
- the direction (positive, negative),
- the strength (strong, weak), and
- the outliers (outliers, no outliers).

### Form

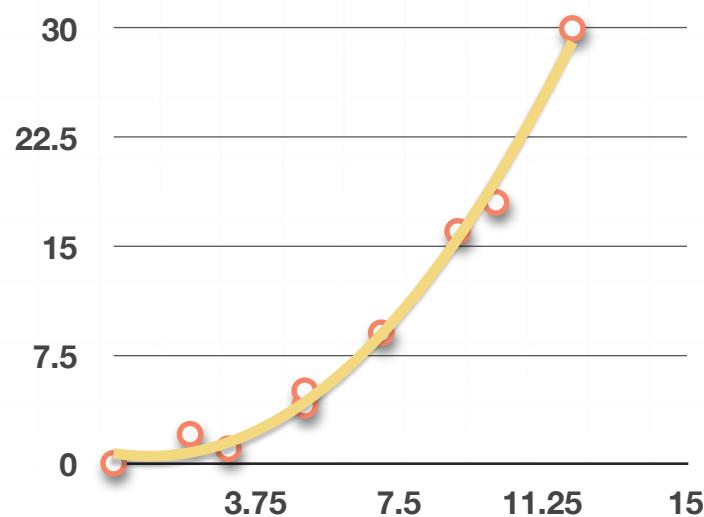
If the data roughly follows a linear trend line, we can say the relationship is linear. If the data more closely follows a parabolic curve, we would say the relationship is parabolic. If the scatterplot just looks like one big blob, and we can't really see any relationship in the data, then we would say there's no relationship or correlation at all.

Linear correlation:

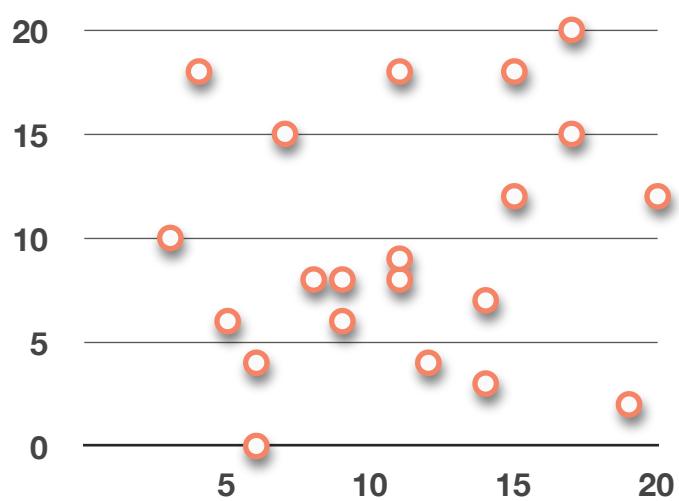




Parabolic correlation:



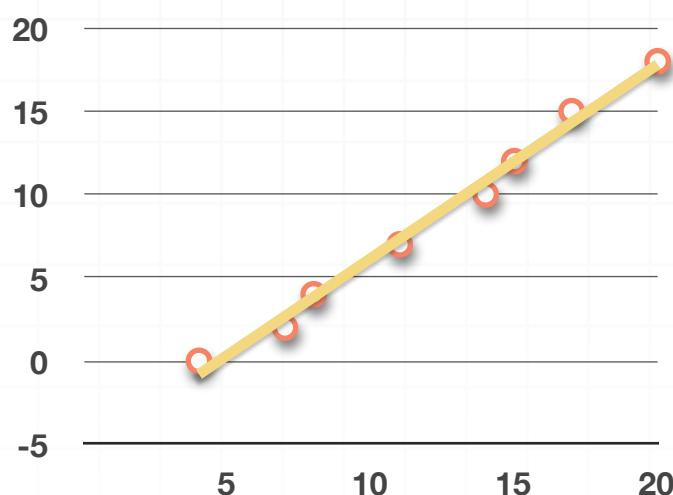
No correlation:



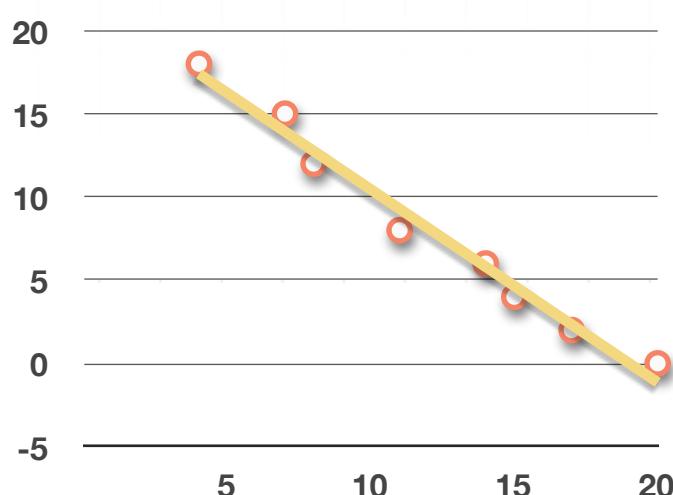
Direction

If the regression line has a positive slope, the data has a **positive linear relationship**; if the regression line of the data has a negative slope, the data has a **negative linear relationship**.

Positive linear relationship:



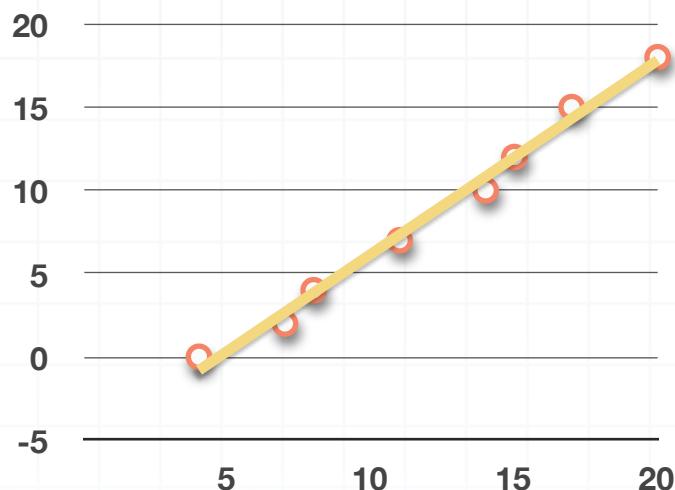
Negative linear relationship:



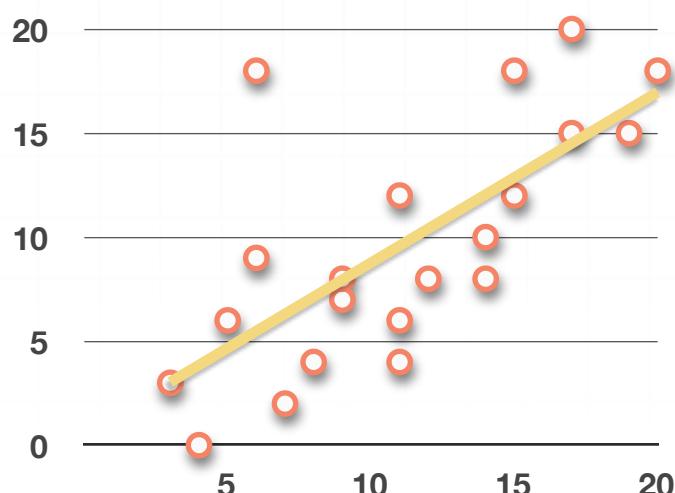
## Strength

If the data is clustered tightly around its regression line, we might say it shows a **strong linear relationship**. If the data is loosely clustered, we might say it shows a **moderate linear relationship**. A **weak linear relationship** would be data that is spread out but still noticeably in the form of a trend line or curve.

## Strong linear relationship:

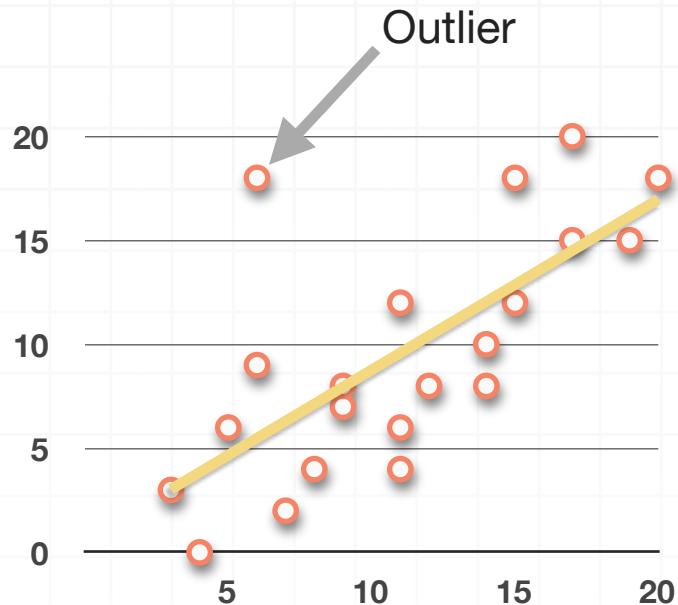


## Moderate linear relationship:



## Outliers

Whether the data has a strong or weak relationship of any kind can also be affected by the existence of outliers, or lack thereof. Remember that an **outlier** is a data point that lies far away from the trend line.

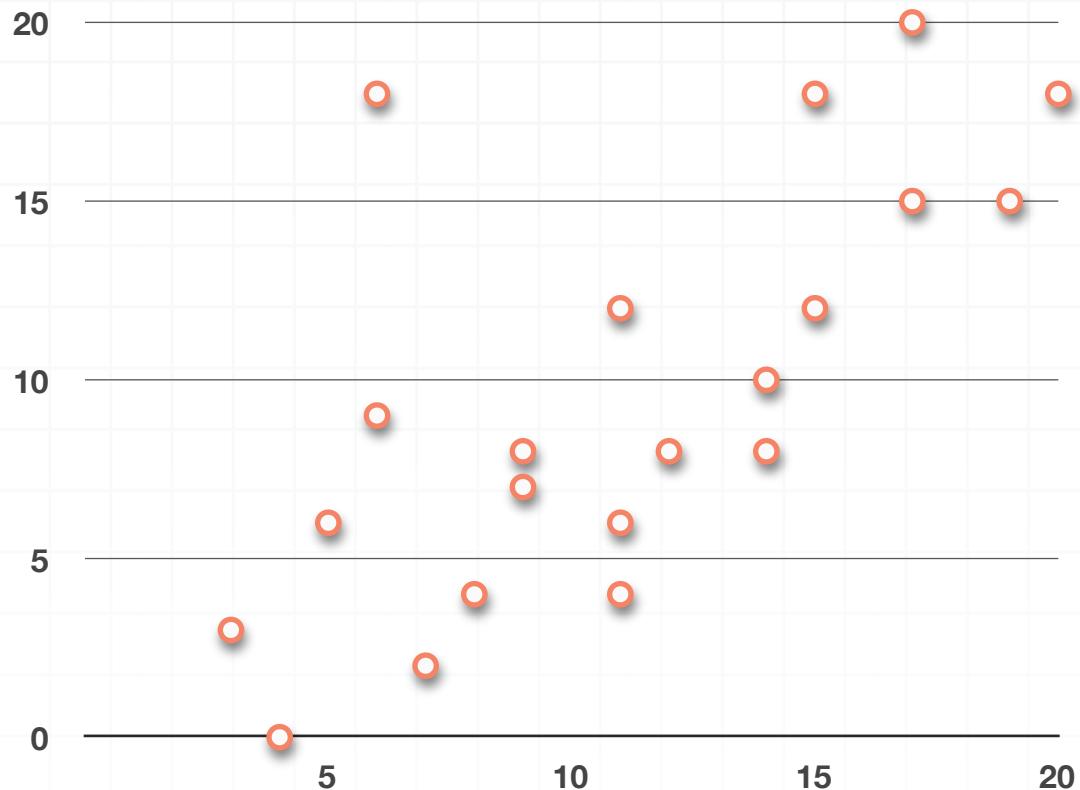


If all of the data points are very tightly clustered, then there are no outliers, which means the data shows a strong relationship. But if there are some or many outliers away from the majority, then the data shows a moderate relationship.

The more outliers there are, and the further away they are, the weaker the relationship. The fewer outliers there are, and the more tightly clustered the data, the stronger the relationship.

### Example

Describe any trend in the data, in terms of form, direction, strength, and outliers.



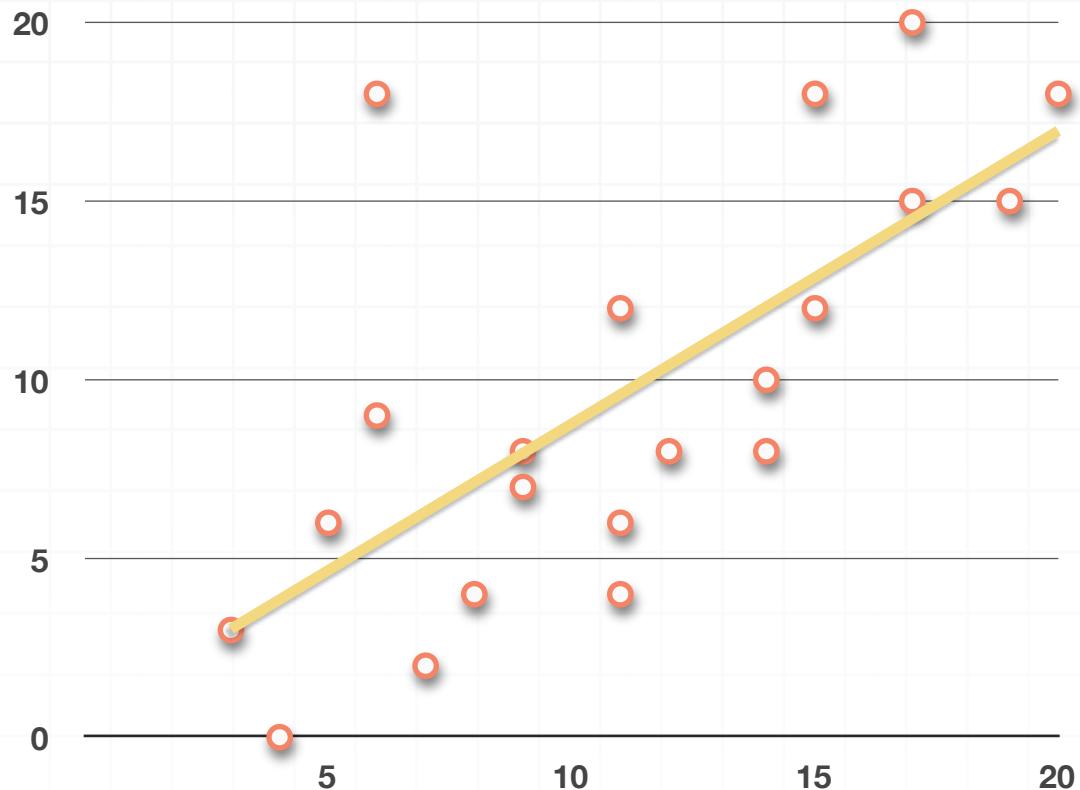
Let's look at each part one at a time: form, direction, strength, and outliers.

**Form:** The scatterplot appears to have a roughly linear relationship, as opposed to a parabolic, or other relationship.

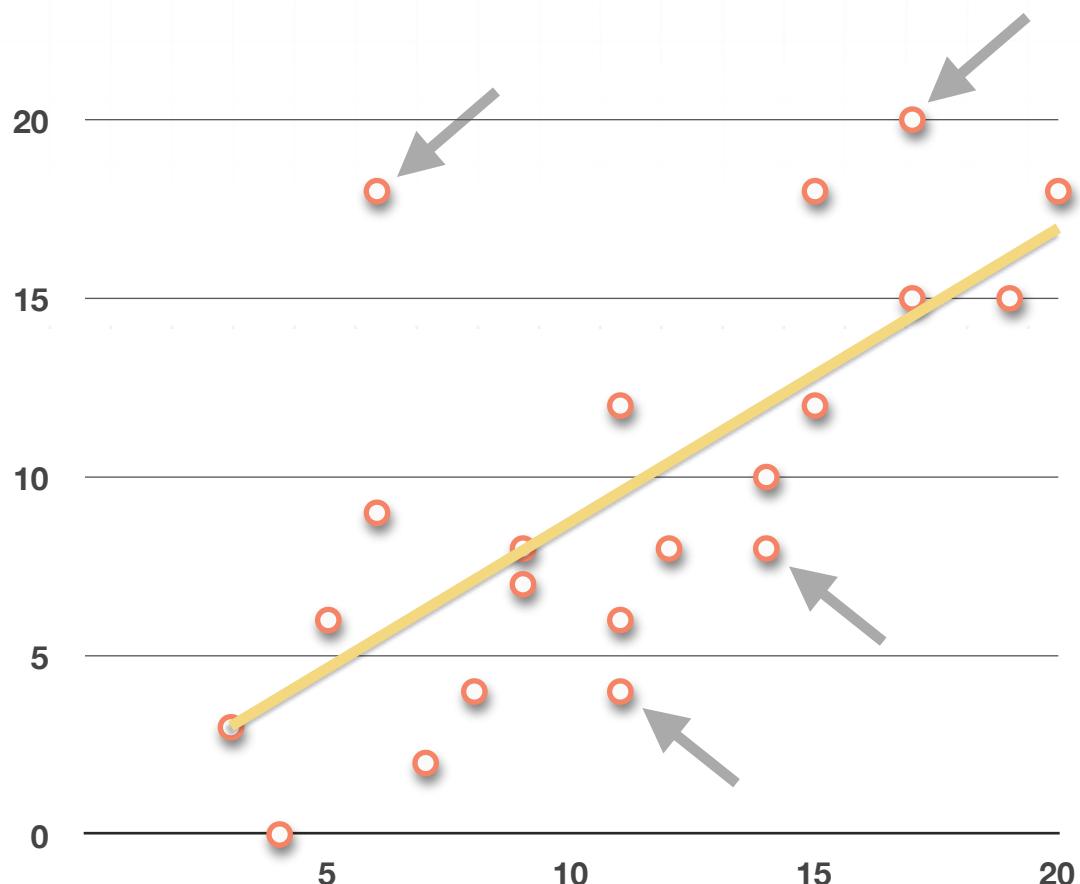
**Direction:** It certainly has a positive relationship, because the data moves up and to the right.

**Strength and outliers:** The strength of the relationship is moderate, which is due to how spread out the data points are, and the existence of outliers in the data set, like (6,18).

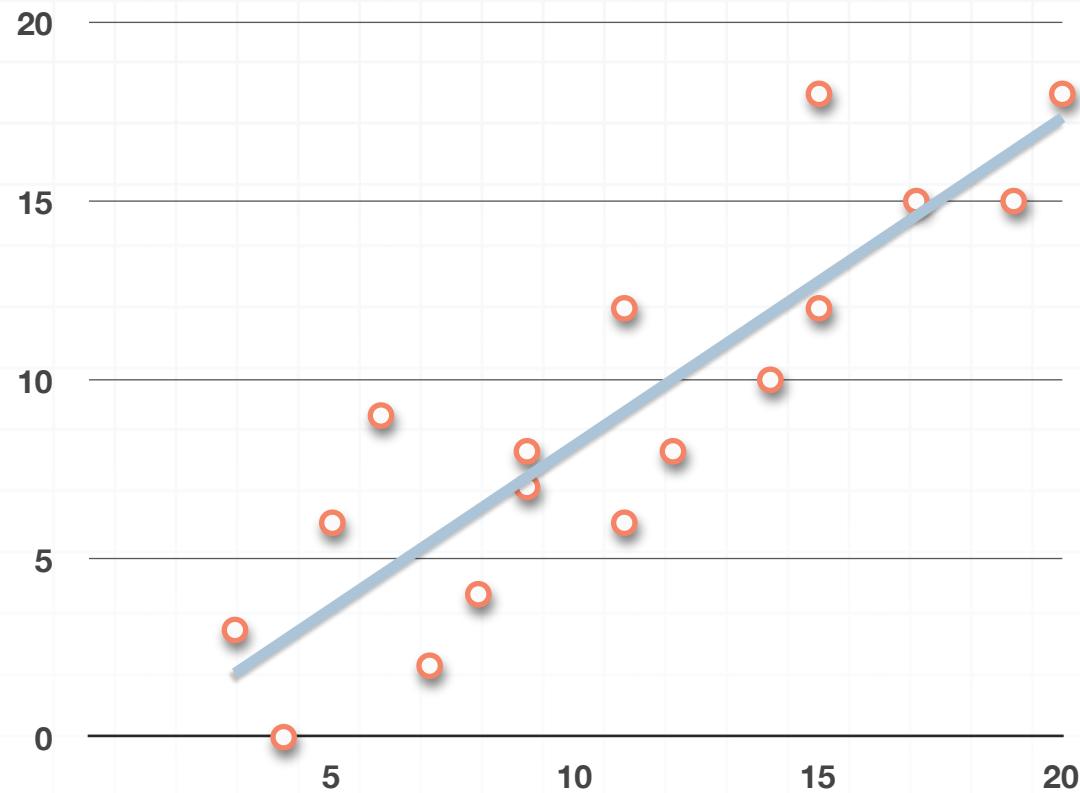
If we plot the trend line, we get



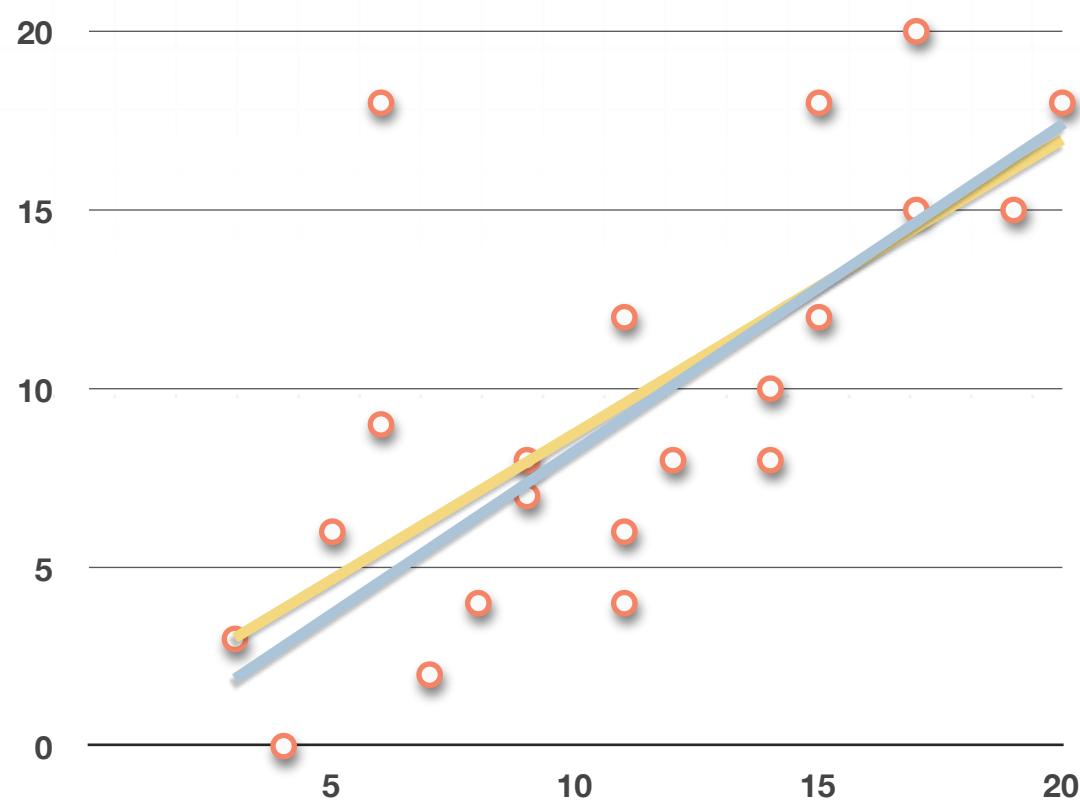
If we take out a few points that are further from the regression line, like (6,18), (11,4), (14,8), and (17,20),



we can see that the new, adjusted regression line fits the remaining data a little bit better:



If we plot both lines on the original scatterplot, we get



and we can see the effect that some of these outliers have on the regression line.

## The purpose of regression

So what's the purpose of curve fitting in general, or finding the equation of the regression line specifically? Well, the main purpose for finding the approximating curve, whether it's a regression line or a regression curve with some other shape, is to come up with an equation that we can use to make predictions.

In the first example from this section, we were given a data table:

x	y
0	0.8
2	1.0
4	0.2
6	0.2
8	2.0
10	0.8
12	0.6

If we were asked to give an approximate value of  $y$  for  $x = 9$ , or for  $x = 100$ , based on this data set, it'd be awfully hard to do using just the data points in the table. After all, the table doesn't give a value for  $x = 9$ , and it *certainly* doesn't give a value for  $x = 100$ .

But if we calculate the equation of the regression line that approximates the data, then we can plug  $x = 9$ ,  $x = 100$ , or any other value into the equation, and we'll get back an estimated value of  $y$ .

And that's the purpose of regression. Technically, **regression** is just the process of estimating the value of the dependent variable from a given value of the independent variable.



# Correlation coefficient and the residual

In the last section we talked about the regression line, and how it was the line that best represented the data in a scatterplot. In this section, we're going to get technical about different measurements related to the regression line.

## Correlation coefficient, $r$

The **correlation coefficient**, denoted with  $r$ , tells us how strong the relationship is between  $x$  and  $y$ . It's given by

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

Notice that in this formula for correlation coefficient, we have the values  $(x_i - \bar{x})/s_x$  and  $(y_i - \bar{y})/s_y$ , where  $s_x$  and  $s_y$  are the standard deviations with respect to  $x$  and  $y$ , and  $\bar{x}$  and  $\bar{y}$  are the means of  $x$  and  $y$ . Therefore  $(x_i - \bar{x})/s_x$  and  $(y_i - \bar{y})/s_y$  are the  $z$ -scores for  $x$  and  $y$ , which means we could also write the correlation coefficient as

$$r = \frac{1}{n-1} \sum (z_{x_i})(z_{y_i})$$

The value of the correlation coefficient will always fall within the interval  $[-1, 1]$ . If  $r = -1$ , it indicates that a regression line with a negative slope will perfectly describe the data. If  $r = 1$ , it indicates that a regression line with a positive slope will perfectly describe the data. If  $r = 0$ , then we can say that



a line doesn't describe the data well at all. In other words, the data may just be a big blob (no association), or sharply parabolic. In other words, the relationship is nonlinear.

Don't confuse a value of  $r = -1$  with a slope of  $-1$ . A correlation coefficient of  $r = -1$  does not mean the slope of the regression line is  $-1$ . It simply means that some line with a negative slope (we're not sure what the slope is, we just know it's negative) perfectly describes the data.

"Perfectly describes the data" means that all of the data points lie exactly on the regression line. In other words, the closer  $r$  is to  $-1$  or  $1$  (or the further it is away from  $0$ , in either direction), the stronger the linear relationship. If  $r$  is close to  $0$ , it means the data shows a weaker linear relationship.

### Example

Using the data set from the last section, find the correlation coefficient.

x	y
0	0.8
2	1.0
4	0.2
6	0.2
8	2.0
10	0.8
12	0.6

First, we need to find both means,  $\bar{x}$  and  $\bar{y}$ ,

$$\bar{x} = \frac{0 + 2 + 4 + 6 + 8 + 10 + 12}{7} = \frac{42}{7} = 6$$

$$\bar{y} = \frac{0.8 + 1.0 + 0.2 + 0.2 + 2.0 + 0.8 + 0.6}{7} = \frac{5.6}{7} = 0.8$$

and both standard deviations  $s_x$  and  $s_y$ .

$$s_x = \sqrt{\frac{\sum_{i=1}^7 (x_i - \bar{x})^2}{7 - 1}} = \sqrt{\frac{36 + 16 + 4 + 0 + 4 + 16 + 36}{6}} = \sqrt{16} \approx 4.3205$$

$$s_y = \sqrt{\frac{\sum_{i=1}^7 (y_i - \bar{y})^2}{7 - 1}} = \sqrt{\frac{0 + 0.04 + 0.36 + 0.36 + 1.44 + 0 + 0.04}{6}} \approx 0.6110$$

Then if we plug these values for  $\bar{x}$ ,  $\bar{y}$ ,  $s_x$ , and  $s_y$ , plus the points from the data set, into the formula for the correlation coefficient, we get

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

$$r = \frac{1}{7 - 1} \left[ \left( \frac{0 - 6}{4.3205} \right) \left( \frac{0.8 - 0.8}{0.6110} \right) + \left( \frac{2 - 6}{4.3205} \right) \left( \frac{1.0 - 0.8}{0.6110} \right) \right]$$

$$+ \left( \frac{4 - 6}{4.3205} \right) \left( \frac{0.2 - 0.8}{0.6110} \right) + \left( \frac{6 - 6}{4.3205} \right) \left( \frac{0.2 - 0.8}{0.6110} \right) + \left( \frac{8 - 6}{4.3205} \right) \left( \frac{2.0 - 0.8}{0.6110} \right)$$

$$+ \left( \frac{10 - 6}{4.3205} \right) \left( \frac{0.8 - 0.8}{0.6110} \right) + \left( \frac{12 - 6}{4.3205} \right) \left( \frac{0.6 - 0.8}{0.6110} \right) \Big]$$

$$r = \frac{1}{6} \left[ \left( -\frac{6}{4.3205} \right) \left( \frac{0}{0.6110} \right) + \left( -\frac{4}{4.3205} \right) \left( \frac{0.2}{0.6110} \right) \right.$$

$$\left. + \left( -\frac{2}{4.3205} \right) \left( -\frac{0.6}{0.6110} \right) + \left( \frac{0}{4.3205} \right) \left( -\frac{0.6}{0.6110} \right) + \left( \frac{2}{4.3205} \right) \left( \frac{1.2}{0.6110} \right) \right]$$

$$\left. + \left( \frac{4}{4.3205} \right) \left( \frac{0}{0.6110} \right) + \left( \frac{6}{4.3205} \right) \left( -\frac{0.2}{0.6110} \right) \right]$$

$$r = \frac{1}{6} \left[ -\frac{4}{4.3205} \left( \frac{0.2}{0.6110} \right) + \frac{2}{4.3205} \left( \frac{0.6}{0.6110} \right) \right.$$

$$\left. + \frac{2}{4.3205} \left( \frac{1.2}{0.6110} \right) - \frac{6}{4.3205} \left( \frac{0.2}{0.6110} \right) \right]$$

$$r = \frac{1}{6} \left( -\frac{0.8}{2.6398} + \frac{1.2}{2.6398} + \frac{2.4}{2.6398} - \frac{1.2}{2.6398} \right)$$

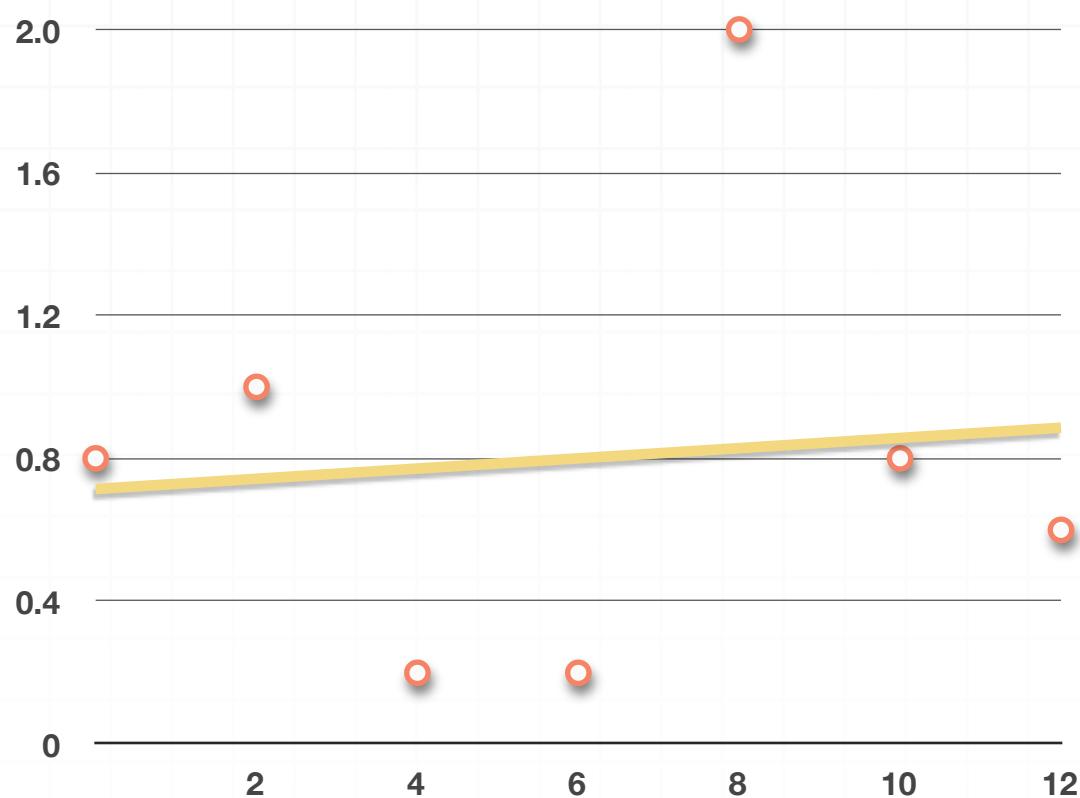
$$r = \frac{1}{6} \left( \frac{1.6}{2.6398} \right)$$

$$r = \frac{1.6}{15.8390}$$

$$r \approx 0.1010$$

This positive correlation coefficient tells us that the regression line will have a positive slope. The fact that the positive value is much closer to 0

than it is to 1 tells us that the data is very loosely correlated, or that it has a weak linear relationship. And if we look at a scatterplot of the data that includes the regression line, we can see how this is true.



In this graph, the regression line has a positive slope, but the data is scattered far from the regression line, with several outliers, such that the relationship is weak.

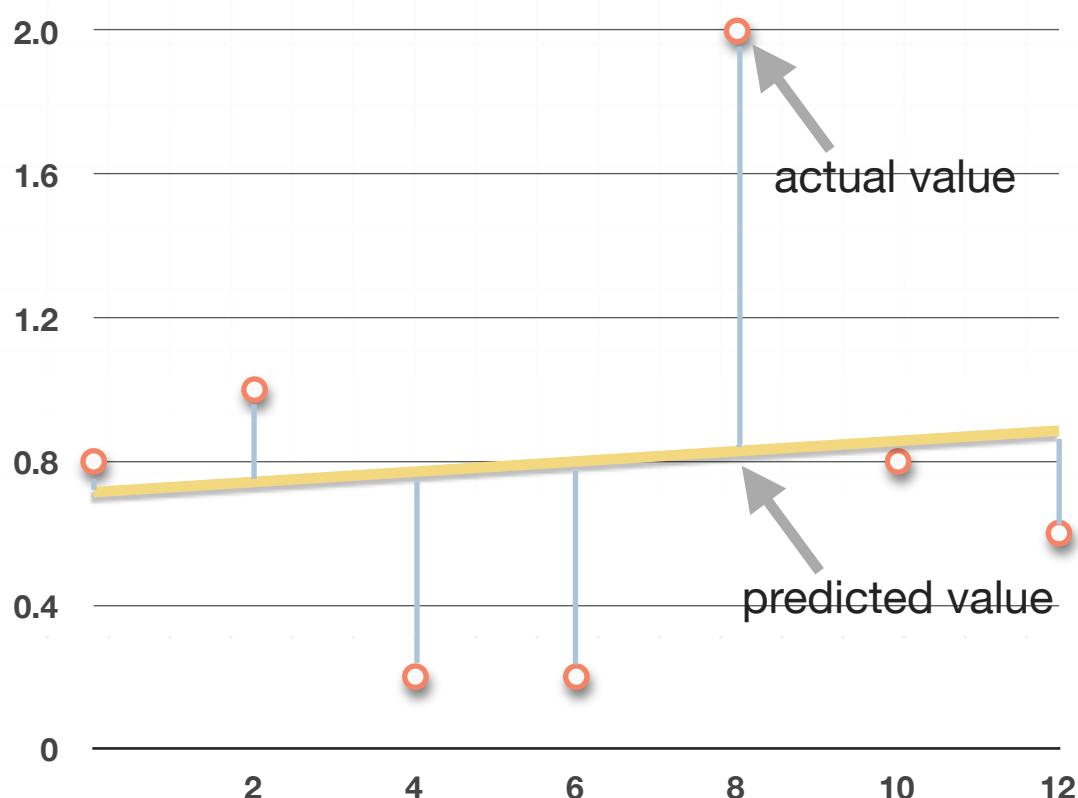
In general, the data set has a

- strong negative correlation when  $-1 < r < -0.7$
- moderate negative correlation when  $-0.7 < r < -0.3$
- weak negative correlation when  $-0.3 < r < 0$
- weak positive correlation when  $0 < r < 0.3$

- moderate positive correlation when  $0.3 < r < 0.7$
- strong positive correlation when  $0.7 < r < 1$

## Residual, $e$

The **residual** for any data point is the difference between the **actual value** of the data point and the **predicted value** of the same data point that we would have gotten from the regression line.



The blue lines in the chart represent the residual for each point. Notice that the absolute value of the residual is the distance from the predicted value on the line to the actual value of the point. The point (8,2) will have a large residual because it's far from the regression line, and the point (10,0.8) will have a small residual because it's close to the regression line.

If the data point is below the line, the residual will be negative; if the data point is above the line, the residual will be positive. In other words, to find the residual, we use the formula

$$\text{residual} = \text{actual} - \text{predicted}$$

The residual then is the vertical distance between the actual data point and the predicted value. Many times we use the variable  $e$  to represent the residual (because we also call the residual the **error**), and we already know that we represent the regression line with  $\hat{y}$ , which means we can also state the residual formula as

$$e = y - \hat{y}$$

Now that we know about the residual, we can characterize the regression line in a slightly different way than we have so far.

For any regression line, the sum of the residuals is always 0,

$$\sum e = 0$$

and the mean of the residuals is also always 0.

$$\bar{e} = 0$$

If we have the equation of the regression line, we can do a simple linear regression analysis by creating a chart that includes the actual values, the predicted values, and the residuals. We do this by charting the given  $x$  and  $y$  values, then we can evaluate the regression line at each  $x$ -value to get the predicted value  $\hat{y}$  ("y-hat"), and find the difference between  $y$  and  $\hat{y}$  to get the residual,  $e$ .



If we use the same data set we've been working with, then the equation of the regression line is

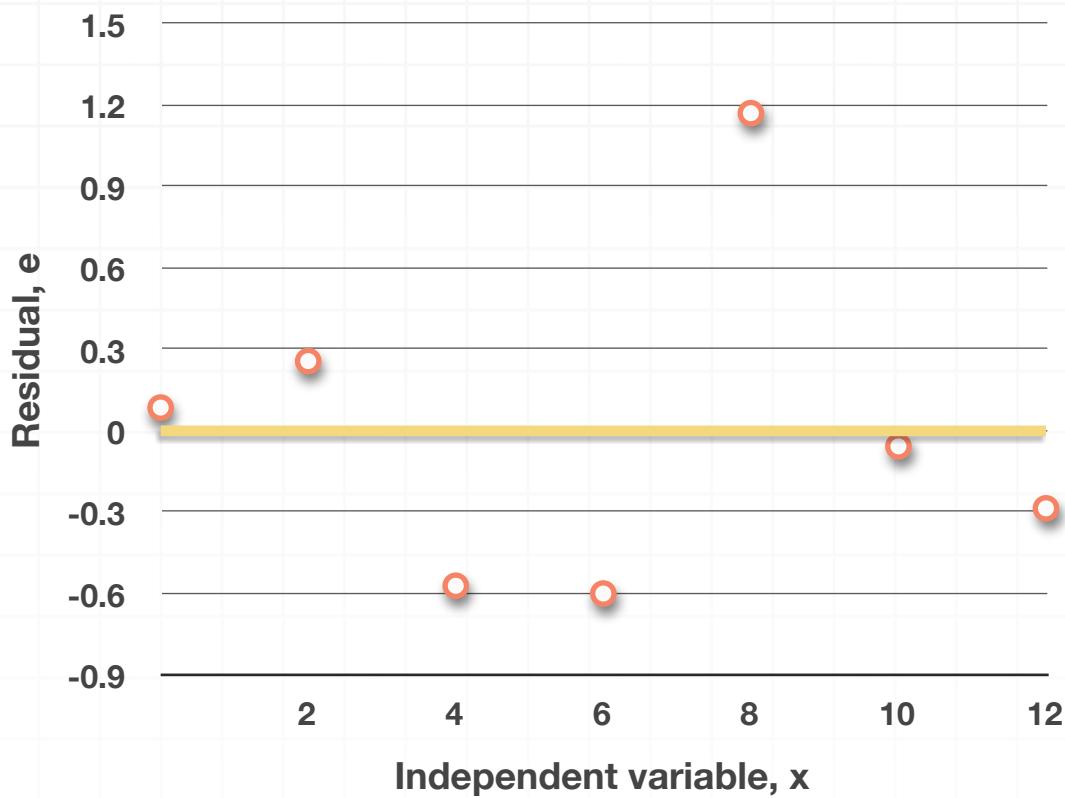
$$y = 0.0143x + 0.7143$$

and we can do the simple linear regression analysis by filling in the chart.

x	Actual	Predicted	e
0	0.8	0.7143	0.0857
2	1.0	0.7429	0.2571
4	0.2	0.7715	-0.5715
6	0.2	0.8001	-0.6001
8	2.0	0.8287	1.1713
10	0.8	0.8573	-0.0573
12	0.6	0.8859	-0.2859

Notice how, if we compare the chart to the scatterplot with the regression line, the negative residuals correspond to points below the regression line, and the positive residuals correspond to points above the regression line.

If we make a new scatterplot, with the independent variable along the horizontal axis, and the residuals along the vertical axis, notice what happens to the regression line.



This should make sense, since we said that the sum and mean of the residuals are both always 0. Whenever this graph produces a random pattern of points that are spread out below 0 and above 0, that tells us that a linear regression model will be a good fit for the data.

On the other hand, if the pattern of points in this plot is non-random, for instance, if it follows a u-shaped parabolic pattern, then a linear regression model will not be a good fit for the data.

## Minimizing residuals

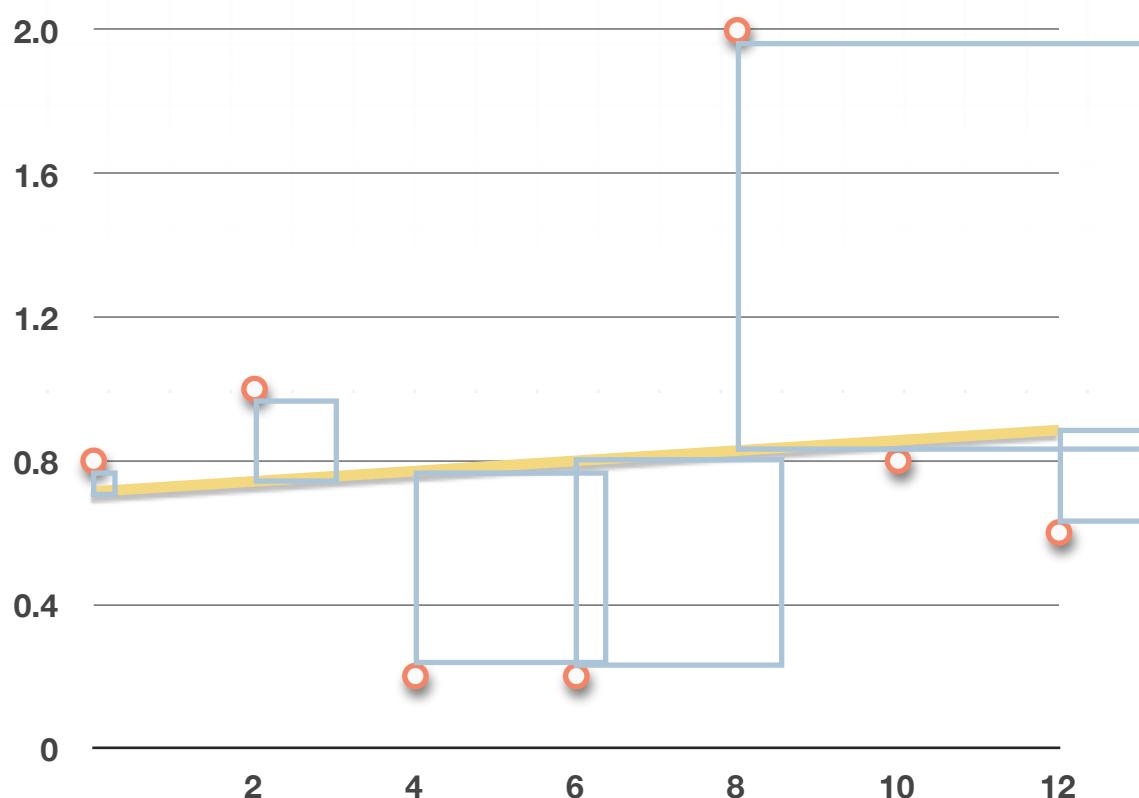
To find the very best-fitting line that shows the trend in the data (the regression line), it makes sense that we want to minimize all the residual values, because doing so would minimize all the distances, as a group, of each data point from the line-of-best-fit.

In order to minimize the residual, which would mean to find the equation of the very best-fitting line, we actually want to minimize

$$\sum (e_n)^2$$

where  $e_n$  is the residual for each of the given data points.

We square the residuals so that the positive and negative values of the residuals do not equal a value close to 0 when they're summed together, which can happen in some data sets when we have residuals evenly spaced both above and below the line of best fit. Squaring them takes out the negative values and keeps them from canceling each other out so that all the residuals can be minimized.



This process of trying to minimize residuals by minimizing the squares of the residuals, is where we get the names **least-squares-line**, **line of least squares**, and **least-squares regression**. We're trying to minimize the area of the squares.

# Coefficient of determination and RMSE

At the end of the last section, we said that, in order to find the equation of the line-of-best-fit, we actually want to minimize

$$\sum (e_n)^2$$

where  $e_n$  is the residual for each data point.

## Coefficient of determination

If  $e_n$  is the residual, the next thing we want to talk about is  $r^2$ , the coefficient of determination. But let's take a step back for a moment.

We've been talking about finding the regression line that best approximates the trend in the data. But we could have simply found the average of all the  $y$ -values in the set and then drawn a horizontal line through the data at that point instead.

For instance, given the data set

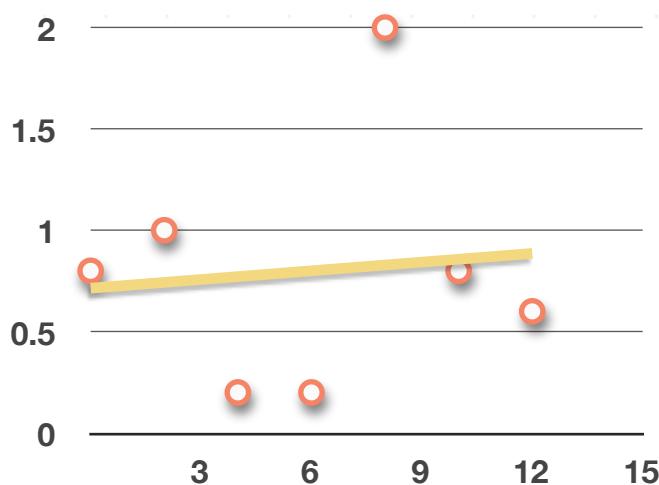
x	y
0	0.8
2	1.0
4	0.2
6	0.2
8	2.0
10	0.8
12	0.6

the mean of the  $y$ -values is

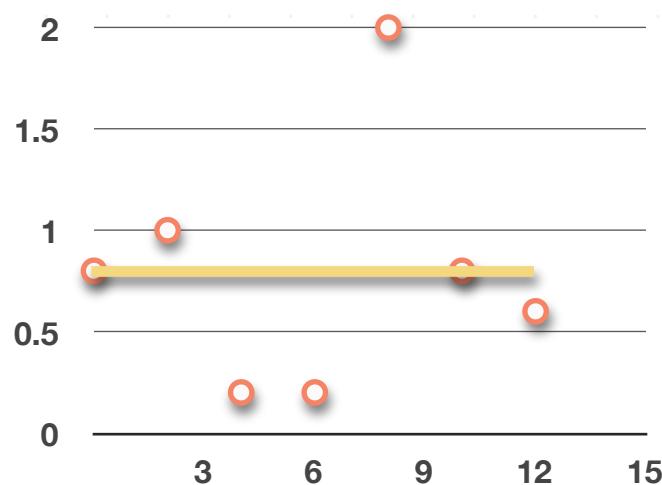
$$\bar{y} = 0.8$$

So, instead of sketching in the regression line, we could have taken a simpler path and sketched in the horizontal line at  $\bar{y} = 0.8$ .

Regression line:



Simple horizontal line:



The horizontal line doesn't fit the data as well as the regression line, but it's much faster to find. What we want to talk about now is how much error we eliminate by using the regression line, instead of the horizontal line.

If we eliminate a large amount of error, then we know that the regression line is a much better approximation than the simpler horizontal line. But if we only eliminate a small amount of error, then we know the horizontal line was actually a pretty good trend line for the data, and that the regression line doesn't do much better.

So how do we find the amount of error eliminated by using the regression line instead of the horizontal line? We do it with the **coefficient of determination**,  $r^2$ , which measures the percentage of error we eliminated by using least-squares regression instead of just  $\bar{y}$ .

We already know that the residual  $e$  of a data point is the distance between the point's actual  $y$ -value and its predicted  $\hat{y}$ -value from the regression line.

For the horizontal line through the data, if we were to take the residual of each data point and square it, and then add up the area inside all of those actual squares, we'd get a "sum of squares" that, in a way, measures the error of just drawing the horizontal line.

If instead we find the regression line that more accurately fits the data, and then we go through the same procedure of finding the residual for each data point compared to the regression line, and take the sum of squares again, what we'll find is that we significantly reduce the sum of squares, and therefore reduce the error. In other words,  $r^2$  tells us how well the regression line approximates the data.

For this reason, the coefficient of determination is often written as a percent, where 100 % would describe a line that's a perfect fit to the data. The higher the value of  $r^2$ , the more data points the line passes through. If



$r^2$  is very small, it means the regression line doesn't pass through many of the data points.

The coefficient of determination is the square of the correlation coefficient, which is why we use  $r$  for correlation coefficient and  $r^2$  for coefficient of determination.

## Root-mean-square error

**Root-mean-square error (RMSE)**, also called **root-mean-square deviation (RMSD)**, we can think of as the standard deviation of the residuals.

In the same way that we talked about the standard deviation of normally distributed data, and how many data points fall within one, two, or three standard deviations from the mean, we can think about RMSE as the standard deviation of the data away from the least-squares line.

Once we find the least-squares line, and we've sketched that through the data, we could draw parallel lines on either side of the least-squares line that represent standard deviations away from the regression line. If the standard deviation is very large, and these lines are far from the least-squares line, it tells us that the least-squares line doesn't fit the data very well. But if the standard deviation is very small, and these lines are close to the least-squares line, it tells us that the least-squares line does a very good job showing the trend in the data.

To find RMSE, we'll find the residual for each data point, then square it. We'll add up all of those square residuals, and then divide by  $n$ . Then we'll



take the square root of that result, and we'll get the standard deviation of the residuals.

If we continue on with the data set we've been working with,

x	y	"y-hat"	e
0	0.8	0.7143	0.0857
2	1.0	0.7429	0.2571
4	0.2	0.7715	-0.5715
6	0.2	0.8001	-0.6001
8	2.0	0.8287	1.1713
10	0.8	0.8573	-0.0573
12	0.6	0.8859	-0.2859

we want to start by squaring the residuals.

x	y	"y-hat"	e	$e^2$
0	0.8	0.7143	0.0857	0.0073
2	1.0	0.7429	0.2571	0.0661
4	0.2	0.7715	-0.5715	0.3266
6	0.2	0.8001	-0.6001	0.3601
8	2.0	0.8287	1.1713	1.3719
10	0.8	0.8573	-0.0573	0.0033
12	0.6	0.8859	-0.2859	0.0817

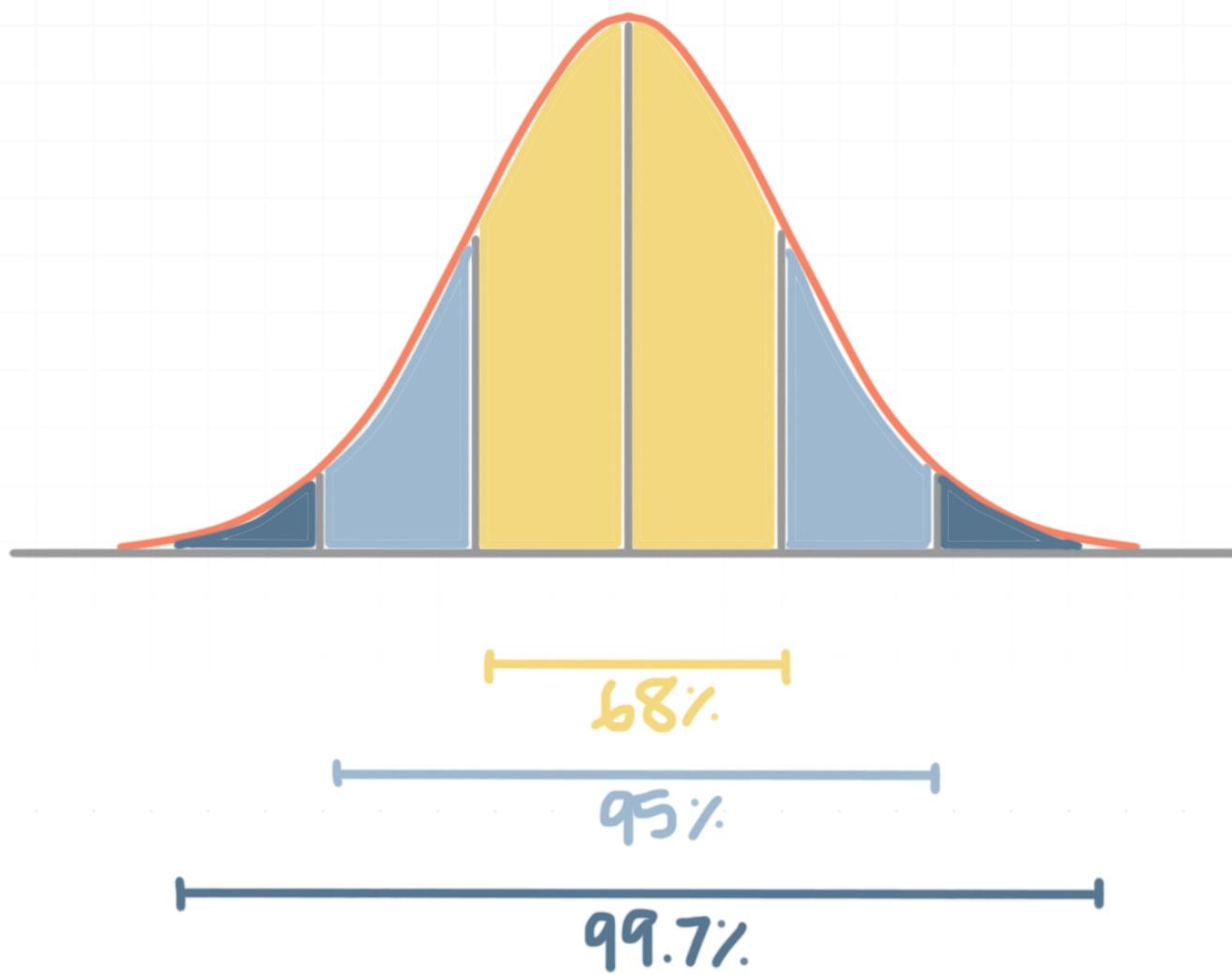
Then we sum the residuals, divide that sum by  $n$ , and take the square root of that result. Since we have  $n = 7$  data points, we get

$$RMSE \approx \sqrt{\frac{0.0073 + 0.0661 + 0.3266 + 0.3601 + 1.3719 + 0.0033 + 0.0817}{7}}$$

$$RMSE \approx \sqrt{\frac{2.2170}{7}}$$

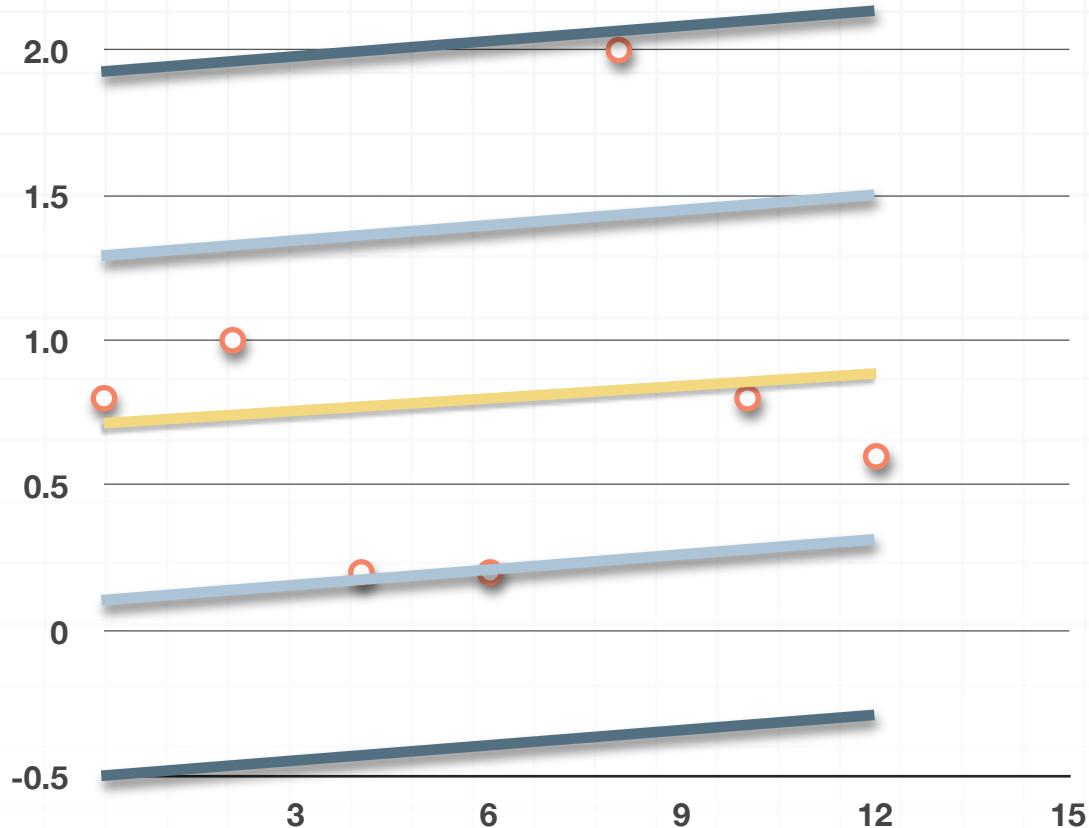
$$RMSE \approx 0.5628$$

This value is the standard deviation of the residuals, which means that, based on the normal curve from earlier in the course,



68 % of the data points will fall within  $\pm 0.5628$  (one standard deviation) of the regression line, that 95 % of the data points will fall within  $\pm 2(0.5628)$  (two standard deviations) of the regression line, and that 99 % of the data points will fall within  $\pm 3(0.5628)$  (three standard deviations) of the regression line.

We could roughly sketch these standard deviation boundaries into the graph.



68 % of the data points will fall inside the light blue lines, and 95 % of the data will fall inside the dark blue lines.

The larger the RMSE (standard deviation),

- the further apart these lines will be,
- the more scattered the data points are, and
- the weaker the correlation is in the data

The smaller the RMSE (standard deviation),

- the closer together these lines will be,
- the more tightly clustered the data points are, and
- the stronger the correlation is in the data

# Chi-square tests

There are three kinds of  $\chi^2$ -tests (chi-squared) we want to look at.

1. A  $\chi^2$ -test for homogeneity
2. A  $\chi^2$ -test for association/independence
3. A  $\chi^2$  goodness-of-fit test

In general,  $\chi^2$ -tests let us investigate the relationship between categorical variables. For instance, the  $\chi^2$ -test for homogeneity is a test we can use to determine whether the probability distributions for two different groups are homogeneous with respect to some characteristic. Whereas the  $\chi^2$ -test for association lets us determine whether two variables are related in the same group. And the  $\chi^2$  goodness-of-fit test is used to determine whether data fits a specified distribution.

We'll cover all three of these  $\chi^2$ -tests in much more detail in this section. But first, we need to remember our conditions for sampling, since we'll often be using samples to represent populations in  $\chi^2$  problems.

## Conditions for inference

Sometimes we'll have a  $\chi^2$  problem where someone's taking a sample of a population. When sampling occurs, in order for us to be able to use a  $\chi^2$ -test, we need to be able to meet typical sampling conditions:

1. **Random:** Any sample we're using needs to be taken randomly.



2. **Large counts:** Each expected value (more on “expected value” later) that we calculate needs to be 5 or greater.
3. **Independent:** We should be sampling with replacement, but if we’re not, then the sample we take shouldn’t be larger than 10 % of the total population.
4. **Categorical:** The variables that we study are categorical.

If any of these conditions aren’t met, we can’t use a  $\chi^2$  testing method. But assuming we can meet these three conditions, then each of these three  $\chi^2$ -tests looks pretty much the same.

## $\chi^2$ -test for homogeneity

When we use a  $\chi^2$ -test for homogeneity, we’re trying to determine whether the distributions for two variables are similar, or whether they differ from each other.

Let’s use an example. Let’s say that we take a sample of male students and a sample of female students and ask all of them whether they prefer cats, dogs, or some other animal as pets.

	Cats	Dogs	Other	Totals
Male	19	34	22	75
Female	26	28	6	60
Totals	45	62	28	135

This is a  $\chi^2$ -test for homogeneity because we're sampling from two different groups (males and females) and comparing their probability distributions.

So in this study we have two variables, gender and pet, and we want to know whether gender affects pet preference, or if pet preference is not affected at all by gender. If pet preference isn't affected by gender, then the distribution of pet preference for males should be the same or similar to the distribution of pet preference for females. If gender does affect pet preference, then the distributions will be different (they won't be homogeneous).

To test this, we want to state the null hypothesis, which will be that gender doesn't affect pet preference.

$H_0$ : gender doesn't affect pet preference

$H_a$ : pet preference is affected by gender

The totals in the table margins let us determine the overall probability of being male or female in this study, regardless of pet preference, and the overall probability of preferring cats, dogs, or some other pet, regardless of gender.

If the distributions for pet preference for males and females are homogeneous, then we should be able to use these marginal probabilities to predict the expected number of students who should be in each cell. If the actual value is very different than the predicted value based on the marginal probabilities, that difference tells us that pet preference might be



affected by gender, and therefore that the distributions of pet preference for males and females might not be homogeneous.

In this survey, the overall probability of being male is  $75/135 \approx 0.56$ , and the overall probability of liking cats is  $45/135 \approx 0.33$ . Which means that if the distributions for males and females are homogeneous, then 33% of 56% of the participants should be male and prefer cats. The value that *should* be in each cell if the distributions are homogeneous is called the **expected value** for that cell.

A simple way to find the expected value for each cell is to multiply the row and column totals and then divide by the overall total.

$$\text{Expected Male-Cats: } (75 \cdot 45)/135 = 25$$

$$\text{Expected Male-Dogs: } (75 \cdot 62)/135 \approx 34.4$$

$$\text{Expected Male-Other: } (75 \cdot 28)/135 \approx 15.6$$

$$\text{Expected Female-Cats: } (60 \cdot 45)/135 = 20$$

$$\text{Expected Female-Dogs: } (60 \cdot 62)/135 \approx 27.6$$

$$\text{Expected Female-Other: } (60 \cdot 28)/135 \approx 12.4$$

Then we can put each **actual value**, which is the count we got for each category when we surveyed the students, next to the (expected value) in an updated table.



	Cats	Dogs	Other	Totals
Male	19 (25.0)	34 (34.4)	22 (15.6)	75
Female	26 (20.0)	28 (27.6)	6 (12.4)	60
Totals	45	62	28	135

If we've done this correctly, the expected values should still sum to the same row and column totals.

## The $\chi^2$ distribution and degrees of freedom

This is where the  $\chi^2$  part comes in. We use the  $\chi^2$  formula to compare the actual and expected value in each cell, and then we sum all those values together to get  $\chi^2$ .

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \frac{(19 - 25.0)^2}{25} + \frac{(34 - 34.4)^2}{34.4} + \frac{(22 - 15.6)^2}{15.6}$$

$$+ \frac{(26 - 20.0)^2}{20.0} + \frac{(28 - 27.6)^2}{27.6} + \frac{(6 - 12.4)^2}{12.4}$$

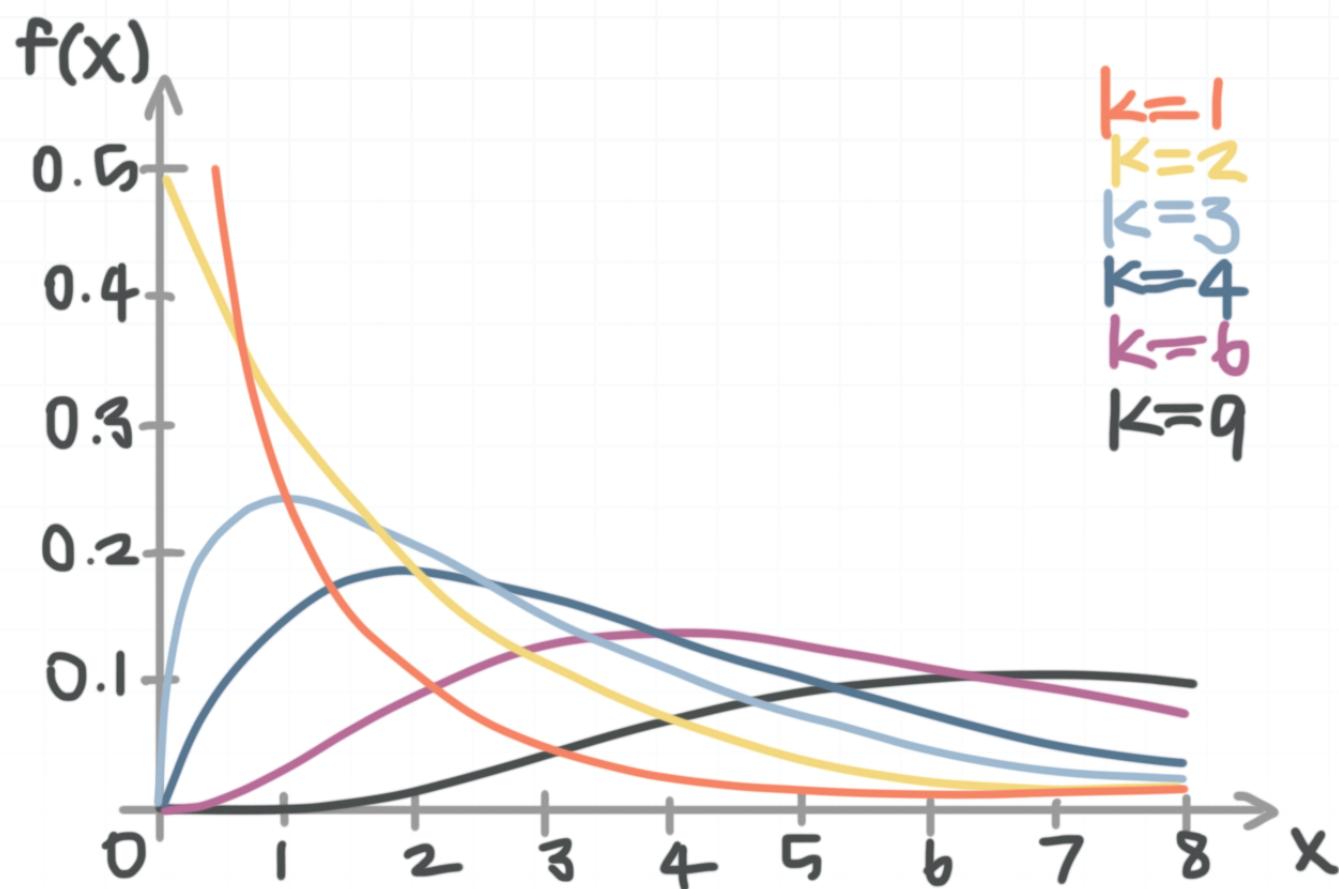
$$\chi^2 = 1.44 + 0.00 + 2.63 + 1.80 + 0.01 + 3.30$$

$$\chi^2 = 9.18$$

In general, the larger the value of  $\chi^2$ , the more likely it is that the variables affect each other (not homogeneous).



But to determine whether or not we can reject the null hypothesis, we need to look up the  $\chi^2$  value and the degrees of freedom in the  $\chi^2$ -table. Similar to the normal- and  $t$ -distributions,  $\chi^2$  has its own probability distribution that looks like this:



There's a distinct  $\chi^2$ -distribution for each degree of freedom. So we can see the  $\chi^2$ -distribution for  $df = 1$  in red, the  $\chi^2$ -distribution for  $df = 2$  in yellow, the  $\chi^2$ -distribution for  $df = 3$  in light blue, etc.

Of course, like the  $t$ -table for  $t$ -distributions, the  $\chi^2$ -table for  $\chi^2$ -distributions includes a degrees of freedom value.

The degrees of freedom is always the minimum number of data points we'd need to have all of the information in the table. So for example in a  $\chi^2$ -test for homogeneity, degrees of freedom is given by

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

When we calculate degrees of freedom, we need to make sure to only include rows and columns from the body. We should never include the total column or the total row. So for the table in the example we've been working with, we get

$$df = (2 - 1)(3 - 1)$$

$$df = (1)(2)$$

$$df = 2$$

It makes sense that  $df = 2$  for this example, because if we have any two pieces of information from the table, we can always figure out the rest of the information.

For instance, if we only know Male-Cats and Male-Dogs, we can find the rest of the missing values.

	Cats	Dogs	Other	Totals
Male	19	34		75
Female				60
Totals	45	62	28	135

Taking Cats-Total minus Male-Cats gives Female-Cats; taking Dogs-Total minus Male-Dogs gives Female-Dogs; taking Male-Total minus Male-Cats minus Male-Dogs gives Male-Other, and then once we have Male-Other, taking Other-Total minus Male-Other gives Female-Other.

But back to the example, looking up  $\chi^2 = 9.18$  and  $df = 2$  in the  $\chi^2$ -table,

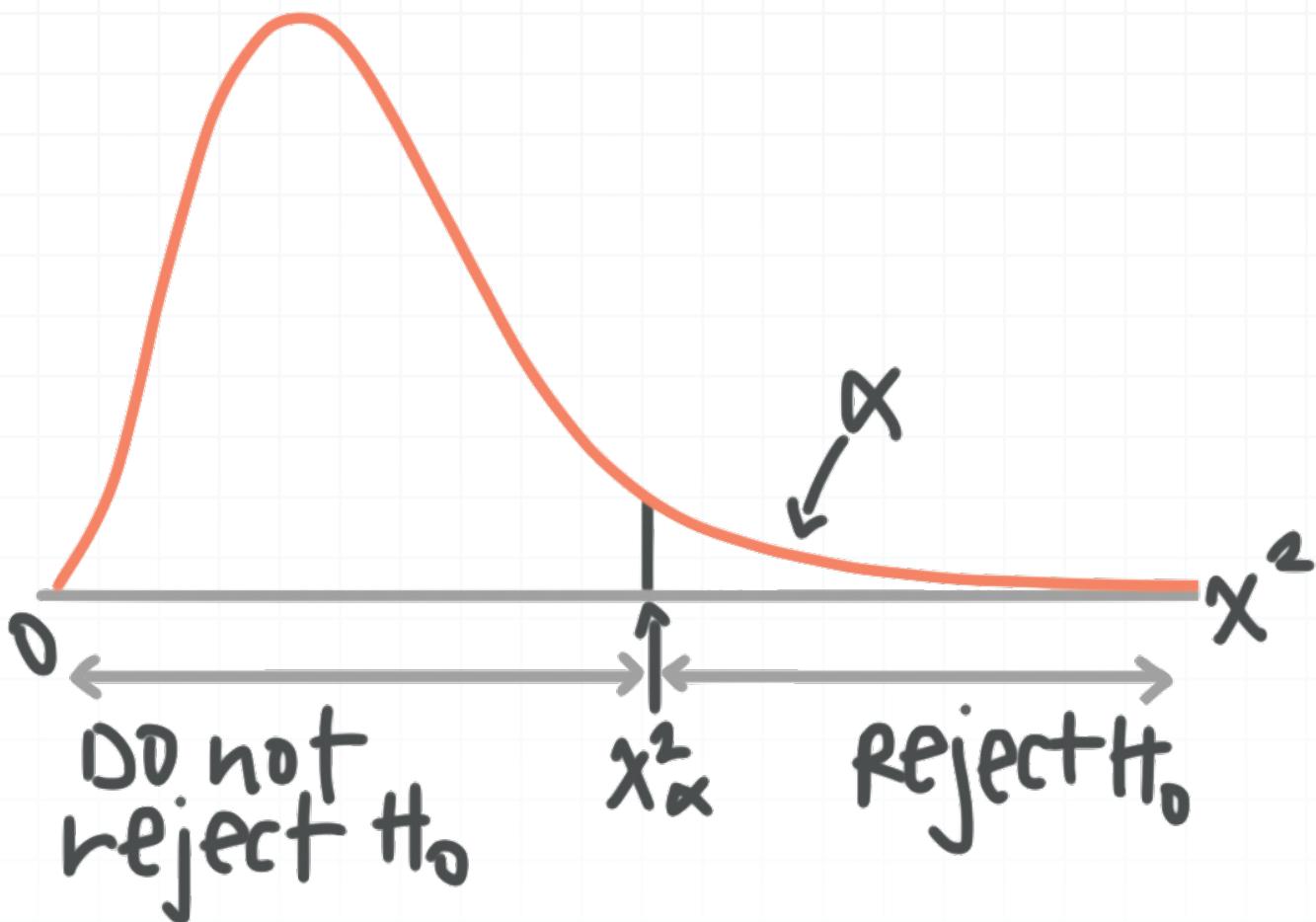


df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.81	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73

puts us between a  $p$ -level of 0.02 and 0.01, closer to 0.01. So we could have picked an alpha value as high as  $\alpha = 0.02$  and still rejected the null hypothesis, concluding that we've found support for the alternative hypothesis that gender affects pet preference (or that pet preference is affected by gender). But if we had picked an alpha value of  $\alpha = 0.01$ , we'd be unable to reject the null hypothesis.

So for this particular example, we can reject the null hypothesis and conclude that there's a relationship between gender and pet preference (at  $\alpha = 0.02$ ).

In general, we usually look at the  $\chi^2$ -table to determine whether the  $\chi^2$ -value falls into the region of rejection. We set  $\chi_{\alpha}^2$  as the critical value that corresponds to our level of significance  $\alpha$ , but the value of  $\chi_{\alpha}^2$  will be determined not only by  $\alpha$ , but also by the degrees of freedom. Once we've calculated our  $\chi^2$ -value, we compare it to  $\chi_{\alpha}^2$ . If  $\chi^2 > \chi_{\alpha}^2$ , we reject the null hypothesis, and if  $\chi^2 < \chi_{\alpha}^2$ , we fail to reject the null hypothesis.



For example, if we'd chosen  $\alpha = 0.02$ , our critical value  $\chi_{\alpha}^2$  at  $df = 2$  would be  $\chi_{\alpha}^2 = 7.82$ , and because  $9.18 > 7.82$ , we would've rejected the null hypothesis.

## $\chi^2$ -test for association/independence

In a  $\chi^2$ -test for association or independence, we're sampling from one group (instead of from two, like we did with the homogeneity test), but we're thinking about two different categorical variables for that same group.

In the example that follows, we're taking a sample of employees at our company (sampling from one group), and looking at their eye color and handedness (two variables about them), so see whether or not eye color and handedness are associated.

## Example

We want to know whether eye color and handedness are associated in employees of our 15,000-person company, so we take a random sample of company employees and ask them their eye color, and whether they're left- or right-handed.

	Brown	Blue	Green	Hazel	Totals
Left-handed	72	36	20	12	140
Right-handed	460	215	130	55	860
Totals	532	251	150	67	1,000

Use a  $\chi^2$ -test for independence with  $\alpha = 0.05$  to say whether or not eye color and handedness are associated.

Start by computing expected values.

$$\text{Expected Left-Brown: } (140 \cdot 532)/1,000 = 74.48$$

$$\text{Expected Left-Blue: } (140 \cdot 251)/1,000 = 35.14$$

$$\text{Expected Left-Green: } (140 \cdot 150)/1,000 = 21.00$$

$$\text{Expected Left-Hazel: } (140 \cdot 67)/1,000 = 9.38$$

$$\text{Expected Right-Brown: } (860 \cdot 532)/1,000 = 457.52$$

$$\text{Expected Right-Blue: } (860 \cdot 251)/1,000 = 215.86$$



Expected Right-Green:  $(860 \cdot 150)/1,000 = 129.00$

Expected Right-Hazel:  $(860 \cdot 67)/1,000 = 57.62$

Then fill in the table.

	Brown	Blue	Green	Hazel	Totals
Left observed	72	36	20	12	140
Left expected	74.48	35.14	21.00	9.38	140
Right observed	460	215	130	55	860
Right expected	457.52	215.86	129.00	57.62	860
Totals	532	251	150	67	1,000

Now we'll check our sampling conditions. The problem told us that we took a random sample, and all of our expected values are at least 5 (Left-Hazel has the smallest expected value at 9.38), so we've met the random sampling and large counts conditions.

And even though we're sampling without replacement, there are 15,000 employees in our company and we're sampling less than 10% of them (1,000 is less than 10% of 15,000), so we've met the independence condition as well.

We'll state the null hypothesis.

$H_0$ : eye color and handedness are independent (not associated)

$H_a$ : eye color and handedness aren't independent (they're associated)

Calculate  $\chi^2$ .

$$\begin{aligned}\chi^2 = & \frac{(72 - 74.48)^2}{74.48} + \frac{(36 - 35.14)^2}{35.14} + \frac{(20 - 21.00)^2}{21.00} + \frac{(12 - 9.38)^2}{9.38} \\ & + \frac{(460 - 457.52)^2}{457.52} + \frac{(215 - 215.86)^2}{215.86} + \frac{(130 - 129.00)^2}{129.00} + \frac{(55 - 57.62)^2}{57.62}\end{aligned}$$

$$\chi^2 = 0.082578 + 0.021047 + 0.047619 + 0.731812$$

$$+ 0.013443 + 0.003426 + 0.007752 + 0.119132$$

$$\chi^2 \approx 1.03$$

The degrees of freedom are

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

$$df = (2 - 1)(4 - 1)$$

$$df = (1)(3)$$

$$df = 3$$

With  $df = 3$  and  $\chi^2 \approx 1.03$ , the  $\chi^2$ -table gives

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

It's clear that  $\chi^2 \approx 1.03 < \chi_{\alpha}^2 = 7.81$ . Therefore, we fail to reject the null hypothesis, which means we can't see an association between eye color

and handedness, and we therefore assume that those two variables are not associated, which means they're independent.

---

## $\chi^2$ goodness-of-fit test

In a  $\chi^2$ -test for homogeneity, we sample from two different groups and compare whether their probability distributions are similar, and in a  $\chi^2$ -test for independence, we sample from one group and try to determine the association between two variables about this one group.

But in a  $\chi^2$  goodness-of-fit test, we have a data table that gives some value or values, contingent upon some other condition, and we test to see whether the sample results are consistent with the values in the data table we've been given.

For instance, in the example that follows, we're working with a contingency table where we're given cupcake sales by month. Which means we're looking at cupcake sales, contingent upon the month of the year.

We're not sampling from two different groups so it doesn't make sense to use a  $\chi^2$ -test for homogeneity, and we're not sampling from one group to look at the relationship between multiple variables of theirs, so it doesn't make sense to use a  $\chi^2$ -test for association, either.

Instead, we'll use a  $\chi^2$  goodness-of-fit test.



## Example

A small cupcake company wants to know if their sales are affected by month. They've recorded actual sales over the past year and collected the data in a table. Use a  $\chi^2$ -test with  $\alpha = 0.05$  to say whether the company's cupcake sales are affected by month.

Month	J	F	M	A	M	J	J	A	S	O	N	D	Total
Sales	60	80	65	70	80	100	140	120	90	90	60	65	1,020

The null hypothesis would be that sales aren't affected by month.

$H_0$ : sales aren't affected by month (the month doesn't affect sales)

$H_a$ : sales are affected by month (the month does affect sales)

If sales aren't affected by month, then they should be evenly distributed over each month, which means sales each month should be

$$\text{expected monthly sales} = \frac{1,020 \text{ total sales}}{12 \text{ months}} = 85 \text{ sales per month}$$

Which means we can expand the table to show observed versus expected sales.

Month	J	F	M	A	M	J	J	A	S	O	N	D	Total
Observed	60	80	65	70	80	100	140	120	90	90	60	65	1,020
Expected	85	85	85	85	85	85	85	85	85	85	85	85	1,020



Now that we have actual and expected values, we can calculate  $\chi^2$ .

$$\begin{aligned}\chi^2 &= \frac{(60 - 85)^2}{85} + \frac{(80 - 85)^2}{85} + \frac{(65 - 85)^2}{85} + \frac{(70 - 85)^2}{85} \\ &\quad + \frac{(80 - 85)^2}{85} + \frac{(100 - 85)^2}{85} + \frac{(140 - 85)^2}{85} + \frac{(120 - 85)^2}{85} \\ &\quad + \frac{(90 - 85)^2}{85} + \frac{(90 - 85)^2}{85} + \frac{(60 - 85)^2}{85} + \frac{(65 - 85)^2}{85} \\ \chi^2 &= \frac{625}{85} + \frac{25}{85} + \frac{400}{85} + \frac{225}{85} \\ &\quad + \frac{25}{85} + \frac{225}{85} + \frac{3,025}{85} + \frac{1,225}{85} \\ &\quad + \frac{25}{85} + \frac{25}{85} + \frac{625}{85} + \frac{400}{85} \\ \chi^2 &= \frac{6,850}{85} \approx 80.59\end{aligned}$$

In a problem like this one, degrees of freedom is simply given by  $n - 1$ . Because looking at the body of the table, there are 12 pieces of information in the body (one for each month), and we would need 11 of these values in order to be able to figure out the 12th. So we can only be missing one pieces of data, and degrees of freedom is therefore  $n - 1 = 12 - 1 = 11$ . Look up  $\chi^2 \approx 80.59$  and  $df = 11$  in the  $\chi^2$ -table.



df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.24	24.05	26.22	28.30	30.32	32.91	34.82

It's clear that  $\chi^2 \approx 80.59 > \chi^2_{\alpha} = 19.68$ , and there's a massive difference between the actual and expected values, which tells us that cupcake sales are very likely affected by month of the year. Which means we can reject the null hypothesis that sales aren't affected by month of the year, and conclude that cupcake sales are affected by month.

Let's walk through another example with a contingency table that looks a little different.

### Example

Marla owns a dog walking service in which she employs 5 of her friends as dog-walkers. She created a table to record the number of friends she believes are walking dogs for her at any given time.

Number of walkers working	1	2	3	4	5
Percent of the time	15%	25%	30%	20%	10%

To test her belief, she took a random sample of 100 times and recorded the number of dog walkers working at that time.

Number of walkers working	1	2	3	4	5
Times	15	16	37	24	8

Say whether a  $\chi^2$ -test can be used to say whether her findings disagree with her belief. If a  $\chi^2$ -test is valid, say whether her findings support her belief with 95 % confidence.

Since Marla is sampling, we need to meet three conditions if she's going to use a  $\chi^2$ -test.

First, the sample needs to be random, and we were told in the problem that she took a random sample.

Second, every expected value needs to be at least 5. To find the expected values, we multiply her expected percentages by the total number of samples, 100.

Walkers working	1	2	3	4	5
Observed	15	16	37	24	8
Expected	$15\% * 100 = 15$	$25\% * 100 = 25$	$30\% * 100 = 30$	$20\% * 100 = 20$	$10\% * 100 = 10$

The smallest of these expected values is 10, which is greater than 5, so we've met the large counts condition.

Third, Marla isn't sampling with replacement, so the sample can't be more than 10 % of the total population. It's safe to assume that Marla could continue taking an infinite number of samples at any given time, in theory



gathering hundreds or thousands of samples as her business continues, so 100 samples shouldn't violate the independence condition.

Therefore, it's appropriate for her to use a  $\chi^2$ -test. She'll state the null hypothesis that her model for the number of friends walking dogs at any given time is correct (her model matches the actual counts that she collected).

To compute  $\chi^2$ , she'll use the actual and expected values.

$$\chi^2 = \frac{(15 - 15)^2}{15} + \frac{(16 - 25)^2}{25} + \frac{(37 - 30)^2}{30} + \frac{(24 - 20)^2}{20} + \frac{(8 - 10)^2}{10}$$

$$\chi^2 = \frac{0}{15} + \frac{81}{25} + \frac{49}{30} + \frac{16}{20} + \frac{4}{10}$$

$$\chi^2 = \frac{911}{150} \approx 6.07$$

With 5 possibilities for number of walkers, there are 4 degrees of freedom, so look up  $df = 4$  with  $\chi^2 \approx 6.07$ .

In order for Marla to reject the null hypothesis at 95 % confidence, she would have needed to surpass a value of 9.49 to be above the 5 % threshold.

df	Upper-tail probability p											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11

However,  $\chi^2 \approx 6.07 < \chi_{\alpha}^2 = 9.49$ . Therefore, Marla cannot reject the null hypothesis, which means she can't conclude that her model is incorrect. In other words, her findings are consistent enough with her model that she can continue to use it.

---



