Chapter 1.2 - Visualizing Data

> In God we trust. All others must bring data. - W. Edwards Deming

Read through the following passage from an incredible author and communicator, Charles Wheelan. I encourage you to choose one of the problems he cites in the passage then try to look into how we can solve that problem with statistics. Remember: *Google is free.*

Wheelan, C. (2013). Naked Statistics. Norton.

Page 4.

> Hal Varian, chief economist at Google, told the *New York Times* that being a statistician will be "the sexy job" over the next decade. I'll be the first to concede that economists sometimes have a warped definition of "sexy". Still, consider the following disparate questions:
>
> How can we catch schools that are cheating on their standardized tests?
>
> How does Netflix know what kind of movies you like?
>
> How can we figure out what substances or behaviors cause cancer, given that we cannot conduct cancer-causing experiments on humans?
>
> Does praying for surgical patients improve their outcomes?
>
> Is there really an economic benefit to getting a degree from a highly selective college or university?
>
> What is causing the rising incidence of autism?
>
> Statistics can help answer these questions (or, we hope, can soon).
>
> The world is producing more and more data, ever faster and faster, Yet, as the *New York Times* has noted, "Data is merely the raw material of knowledge." Statistics is the most powerful tool we have for using information to some meaningful end, whether that is identifying underrated baseball players or paying teachers more fairly...

Data

Many people have a decent idea of what data looks like in their minds, despite the fact that most can't provide a clear definition of it.

Some of us think of tables filled with numbers and characters:

N	Age Class	Weight	Sex	Location
1	0.5	30.8	M	B
2	0.5	21.8	M	B
3	2.5	47.6	M	A
4	0.5	29.0	F	B
5	2.5	65.8	M	A

Others may think of an email from a group project partner with a link to a poorly formatted Google Sheet. Tech savvy individuals might consider a USB or SSD filled with pictures, programs, and documents. Overworked laboratory assistants might imagine a mass of papers filled with plate counts or NMR results.

The reality is that the definition is vague and flexible so all of these things are, in fact, data. But what makes them all data? It might help to have a formal definition:

- **Data:** Information that has been collected
- **Individual:** *Something* the information has been collected on (People/Places/Things/etc.)
- **Variables:** Characteristics about the *individuals* we collected information (*data*) from

Let's look at some data from one of the examples we'll revisit many times. Below is a snapshot from a dataset on white-tailed deer:

N	Age Class	Weight	Sex	Location
1	0.5	30.8	M	B
2	0.5	21.8	M	B
3	2.5	47.6	M	A
4	0.5	29.0	F	B
5	2.5	65.8	M	A

- We collected *information* on deer.
- The *variables* are age class, weight, sex, and species.
- The *values* of those variables are called *data*.

Variables are something you'll (hopefully) come to be exceedingly familiar with. They're the defining characteristic we use to develop statistical models and tests. More often then not, we design entire experiments and studies around what variables we can reasonably measure.

Variables

A lot of how we do statistics depends on what data we have.

Consider the following data set from an experiment on wheat growth relative to soil nitrogen content:

N	Treatment	% Nitrogen	Replicate	Stage	Plot Location
1	E	1.29	4	P3	East
2	C	2.16	3	P2	West
3	C	2.33	2	P2	West
4	B	1.46	3	P3	East
5	D	2.42	4	P1	West
6	A	2.14	1	P2	East

Take a moment and think of an answer to this question: *What's the difference between column 2 of this table and column 3?*

Hold onto that answer as you keep reading and ask yourself if your definition matches what we'll formally define.

Qualitative (Categorical) variable: The value of the variable represents a *descriptive categories*

- These tend to be identifying labels or names
- The problem is we can't really do math with a label or name
- But we can "code" these into numbers to fix that
- i.e., Cat-owners: 0, Dog-owners: 1, Both: 2

Quantitative variable: The value of the variable represents a *meaningful number*

- The height of a person, the number of sales of a product
- These are simple, we can inherently do math with these

N	Treatment	% Nitrogen	Replicate	Stage	Plot Location
1	E	1.29	4	P3	East
2	C	2.16	3	P2	West
3	C	2.33	2	P2	West
4	B	1.46	3	P3	East
5	D	2.42	4	P1	West
6	A	2.14	1	P2	East

How would you organize Column 2?

What about Column 6?

Qualitative variables can be **ordinal** or **nominal**

- **Ordinal variables**: Categories/values of the variable have a natural ordering
  - Letter grade: A, B, C, D
  - Clothing size: S, M, L
- **Nominal variable**: Categories/values of the variable cannot be ordered *naturally*
  - State of residence
  - Degree program

Quantitative variables can be **discrete** or **continuous**

- **Discrete variable**: A countable number of values (0, 1, 2, 3, 4, . . .)
  - Number of students in a classroom
  - Population size of fish in a pond
  - How many times a coin flip was successfully called
- **Continuous variable**: A continuous range of numbers (0, 0.1, 0.11, 0.111, . . .)
  - Temperature
  - Volume of liquid in a glass
  - Height/Weight

Quantitative variables can be categorized by *level of measurement* used for obtaining data values:

- **Interval** level
  - Differences between values make sense
  - Ratios don't make sense because *zero has no meaning*
  - Temperature in Celsius/Fahrenheit (Does 0 mean there's no heat?)
  - Dates (Is there a meaningful ratio you can make out of 1997 and 2020?)
- **Ratio** level
  - Differences between values make sense
  - Ratios *also make sense*
  - Zero *has meaning*, it represents absence of the quantity
  - Height (If you're 0 inches tall, do you have height? Is there a meaningful percentage difference in height between 64 and 67 inches?)

It's okay to be overwhelmed or confused on the differences between variable *sub*-types, as we'll call them. I myself struggled to make the differentiation through most of my early graduate schooling. Whatever method makes the most sense to you is the one you should use but I found that the idea of "bounded by zero" to be the most effective.

Communicating with Data

I've sat in on an excess of presentations featuring a slide deck filled with grotesque tables of data. The idea feels reasonable, why wouldn't everyone want to see your data set? You needed to marry yourself to every cell in that precious Excel sheet to build your presentation, everyone else needs to do the same if they want a chance at understanding the problem!

Except that we haven't addressed the more egregious problem, one that statistics elegantly solves.

Raw data isn't always useful, let alone attractive to look at.

age	**sex**	**cp**	**trestbps**	**chol**	**fbs**	**restecg**	**thalach**	**exang**	**oldpeak**	**slope**	**ca**	**thal**	**target**
41	1	1	110	235	0	1	153	0	0.0	2	0	2	1
62	1	2	130	231	0	1	146	0	1.8	1	3	3	1
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
52	0	2	136	196	0	0	169	0	0.1	1	0	2	1
40	1	0	110	167	0	0	114	1	2.0	1	0	3	0
64	1	0	128	263	0	1	105	1	0.2	1	1	3	1
52	1	0	128	204	1	1	156	1	1.0	1	0	0	0
61	0	0	145	307	0	0	146	1	1.0	1	0	3	0
65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
44	1	2	120	226	0	1	169	0	0.0	2	0	2	1

It's a story as old as time that a company retains an analyst at a lower level than they should simply because they're highly competent with the data the company regularly produces. In fact, many hobbies that involve some form of quantitative review or analysis are held up by communities of people who look at raw data and make rough inferences from them.

This is where an education in statistics can simplify things significantly.

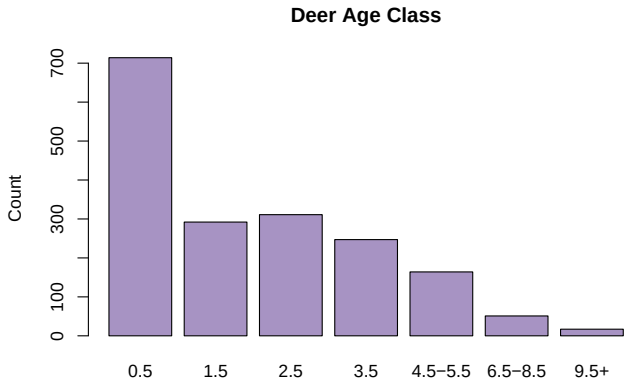
> Statistics is really good at *summarizing* and *visualizing* data

Graphics are the best choice for "downward communication", communicating information to those who know very little about our data or industry. Choosing the "best" graph for displaying our data depends on our data. As usual, we should ask ourselves some questions before building our graphics.

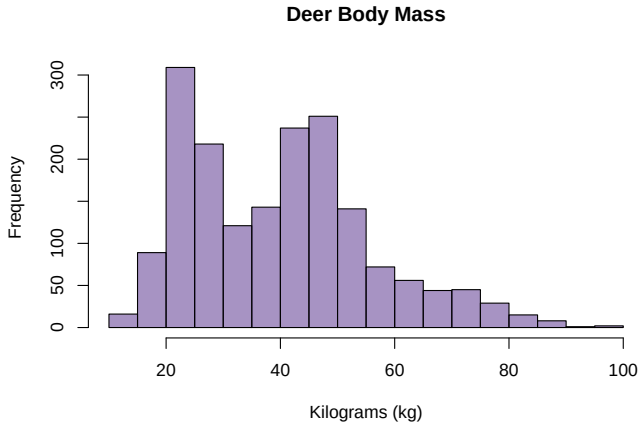
- What kind of data do we have? Is it *categorical* or *numerical*? (Qualitative or Quantitative)
- What are we trying to do? Describe our sample? Look at the distribution of our data? See how two or more variables are related?

Let's walk through a few ways we can "*present*" our data to make it a little more digestible and attractive.

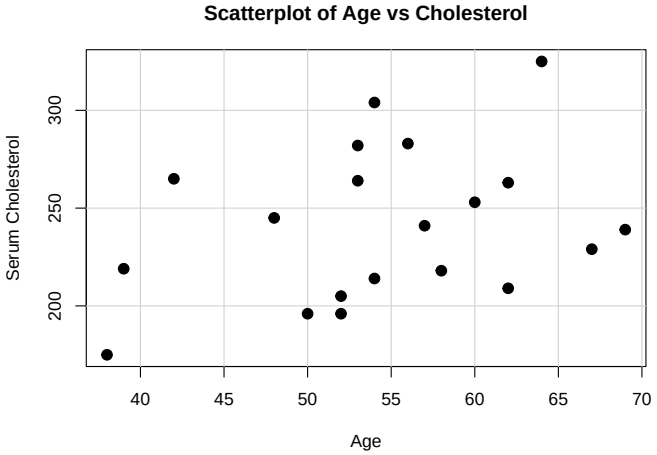
Bar graph: One or more *categorical* variable



Histogram: One *numerical* variable



Scatterplot: *More than one* numerical variable



Summarizing Data

Even when clean, data is messy

Interpreting information is how we make decisions— and every decision we make is data driven, even when it's "emotional data".

Think of the last time you had to make a choice between two possible options. You may have briefly considered the "pros and cons" of each option. Maybe you resolved that the options were neglibly different and resolved to let a coin flip decide. *Even then, you still informed yourself with data.*

Statistics gives us tools to summarize and interpret data rapidly. As with any tool you might use to hang a painting on a wall or build a house, it's best to use the one most suited for the task. That said, it's also important that you use the tool you *best understand*.

There's no point in using a table saw if you're cutting up firewood, but it's even more pointless if you don't even know how to use a table saw.

Frequency Distribution

Let's consider a simple problem. We have some White-tailed deer that come from one of two possible parks labeled "Boyer" and "Desoto":

Location of harvest	Date of harvest	Sex	Age class	Body mass in kg
Boyer	2005-10-15	Female	3.5	34.0
Desoto	2004-12-12	Male	3.5	71.2
Desoto	2009-10-17	Male	0.5	21.8
Desoto	2010-01-02	Male	0.5	19.5
Desoto	2005-12-11	Female	3.5	45.4

It's only 5 data points, but we can still simplify it by putting the locations into their own table:

Boyer	Desoto	Desoto	Desoto	Desoto

We now have a *somewhat useful* visual tool for our single variable of interest, but how could we simplify it *even further*?

But the solution feels intuitive— we're going to count each recurring variable. Why not skip the table and just say that there were 4 deer from Desoto in our random sample?

That's actually exactly what we're going to do.

When we group these variables into a new table that describes the *frequency* each value occurs in the data, we refer to it as a **frequency distribution**.

Frequency distribution:

- Groups data into categories
- Records the number of observations that fall into each category
- "How *frequently* do these variables occur in my sample?"

Another quick trick we can use to simplify larger or messier samples is to consider the proportion of each recurring value *relative* to the total sample. This is referred to as a **relative frequency distribution**.

Relative frequency distribution

- Divide the number in each category by the total number of observations
- This gives us the *proportion* of units in each category
- "What *percentage* of my sample is represented by this variable?"

Location	Frequency	Relative Frequency
Boyer	1	$1/5 = 0.20$
Desoto	4	$4/5 = 0.80$
Total	5	$5/5 = 1.00$

Count up how many times each variable occurs in the sample. For each variable, divide the *occurrences* of the variable by the *sample total*:

- 4 deer from Desoto
- 5 deer total in the sample
- $\frac{4}{5} = 0.80$
- $0.80 * 100\% = 80\%$

Bar Graphs

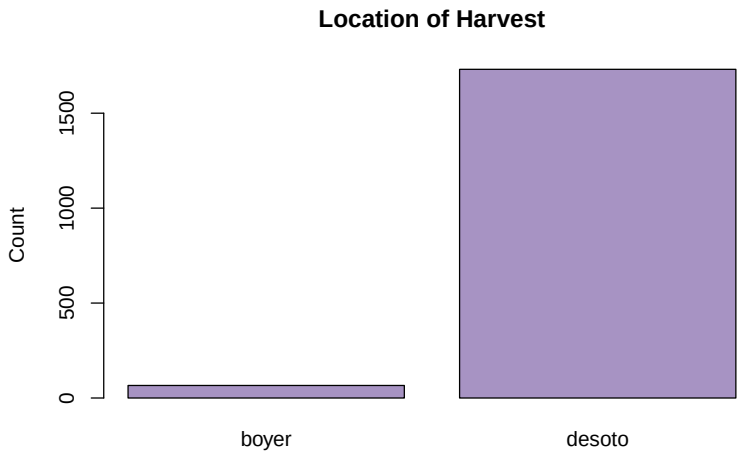
A key component of communication is aesthetics. While we generally won't admit it we tend to listen with our eyes as much as our ears. The same concept applies to *scientific* communication— which also applies to statistics since we've established it as a *science*.

Presenting raw data at a conference is the equivalent of showing up to a job interview in sweatpants. If we want people to believe we're the expert then we dress like the expert. If we want our data to be seen as high quality we should dress it up as such.

The best way to achieve that is with *graphs*.

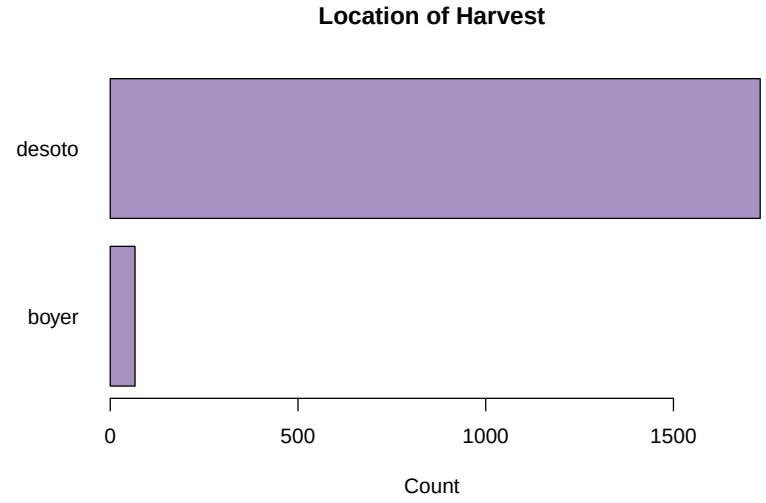
Let's revisit the White-tailed deer, but with the *full dataset* of 1797 deer. We want to understand where these deer were sourced from but 1797 deer is a lot to handle.

We know we're dealing with a single categorical variable, so we should use a *bar graph*:



This bar graph achieves the same result as the *frequency distribution*, with the added benefit of being significantly nicer to look at.

We can also just flip this graph horizontal:

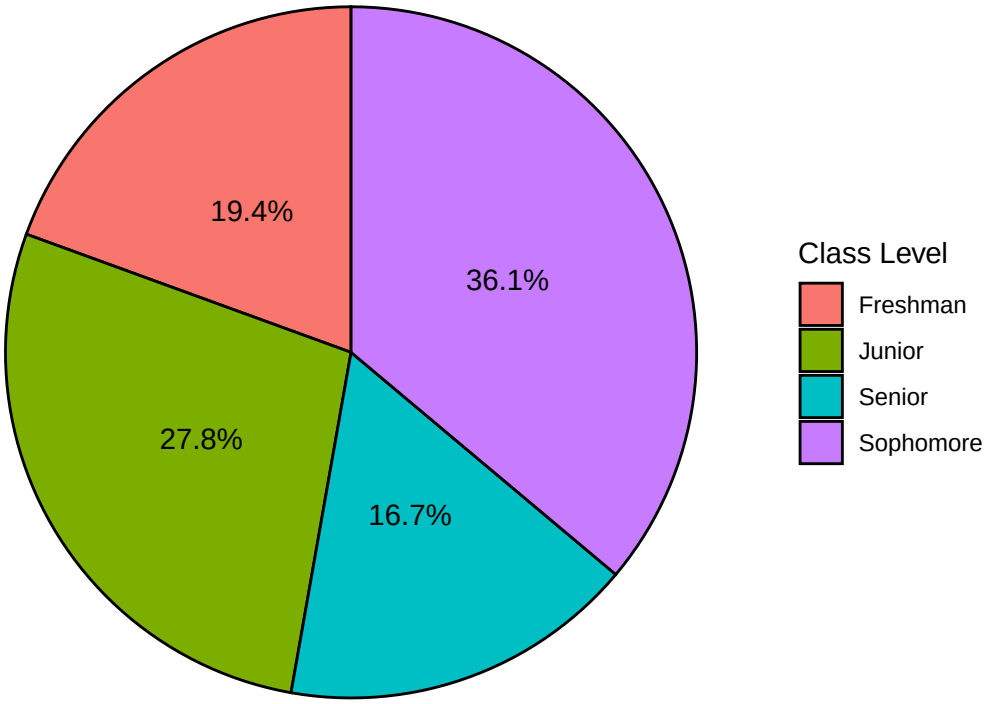


While this flip accomplished nothing in this case, it's useful for when the variable names are *especially long*. Small font sizes won't save us all the time.

Pie Charts

Pie charts get a lot of attention in media and business consulting despite their somewhat niché existence everywhere else.

Generally a pie chart will show relative frequency, which means that any *bar graph* with a *complete relative frequency of* 1.00 can be converted into a pie chart.



As I said before, they're popular in media and business. Almost exclusively because they're very pretty and very simple. But they're not particularly interpretable.

As we stack more and more possible values onto a pie chart they can become problematically clustered. Try to consider what would happen if we instead decided to present the day of the month each student in our sample was born. What would that pie chart look like?

Interpretability is everything, no matter how beautiful your figure is it still has to *communicate something.*

Visualizing Quantitative Data

We've looked at some *qualitative* (categorical) visualizations. These tend to be the easiest to present (although quite a pain to work with). But what about **quantitative** (numerical) visualizations?

Most statistics textbooks discuss a "variety" of method for visualizing *one quantitative* variable: **Histograms**, **Steam-and-leaf plots**, and **Dotplots** are generally those "options. The reality is that the only option used in practice is histograms. We'll look into the reasoning for this later on.

With *two* quantitative variables we generally use a **scatterplot**. We'll cover these at length during our discussion on correlation, so don't worry too much about them. However I encourage you to think on this point:

We can use more than two quantitative variables in a scatterplot, but we generally don't. Why might that be?

Quantitative Frequency Distributions

Quantitative data tends to be the type analysts are most concerned with simply because it's the most *difficult* to understand and communicate.

Location of harvest	Date of harvest	Sex	Age class	Body mass in kg
Desoto	2004-10-16	Female	2.5	45.8
Desoto	2004-12-12	Male	2.5	65.8
Desoto	2007-01-06	Female	4.5–5.5	44.5
Desoto	2005-12-11	Male	3.5	71.2
Desoto	2005-12-11	Female	4.5–5.5	42.2
Desoto	2005-01-09	Male	3.5	68.9
Desoto	2004-12-11	Male	2.5	61.7
Desoto	2010-01-02	Male	0.5	19.5
Desoto	2004-12-11	Male	4.5–5.5	70.8
Desoto	2007-01-06	Female	2.5	41.3

Statisticians aren't very creative and mathematics as a whole strives to be as lazy as possible. Hence, we reuse our methods whenever we can. So naturally, we can summarize *quantitative variables* with a frequency distribution. We have to define interval(s) for the data (referred to as **classes**/**class**), then record the number of observations that fall into each class.

Looking at the total data set of deer body weights would be cumbersome, so a frequency distribution helps us get a more concise understanding of that data.

Class	Frequency	Relative Frequency
0-20	90	0.05
20-40	806	0.45
40-60	701	0.39
60+	200	0.11
Total	1797	1.00

There's no "one" right way to choose the number of classes or the width for a frequency distribution. There are, however, best practices. Remember our golden rule of interpretability and try to keep your classes to a readable level. Otherwise, we select the class number and width differently based on the data we have.

Histograms

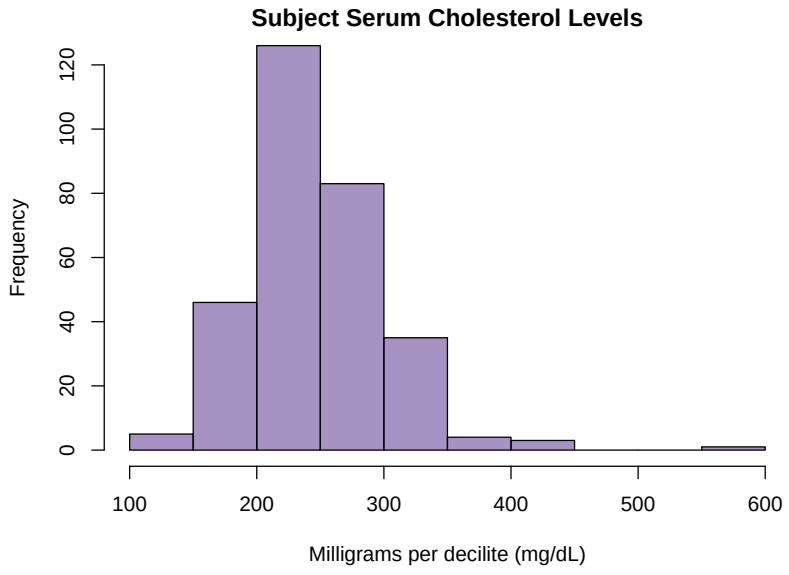
Statistics is an interesting field, in that as you learn more about it you'll become more inclined to use simpler techniques from elementary classes as opposed to the complex and "boutique" methods you may learn in an advanced course.

Histograms are an excellent example of that. The most complex fields of study within statistics still default to visualizing quantitative data through histograms. So it's important to familiarize yourself with this graphic, not just because of its ubiquity but because of it's usefulness.

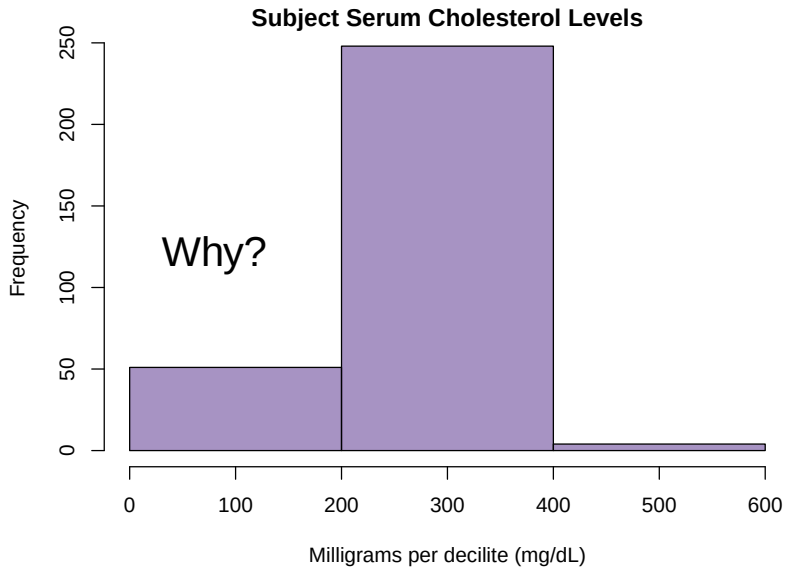
Histogram: visual representation of a frequency distribution

It's important to note that a histogram is *not a bar graph.*

- Bar *height* (y-axis) represents **class frequency**
- Bar *width* (x-axis) represents **class width**



This histogram has 10 classes. You can choose a different number of classes, you can choose different widths, free will exists, there are no rules.



Above is *objectively* a histogram (despite it being quite horrifying to look at), is this interpretable though?

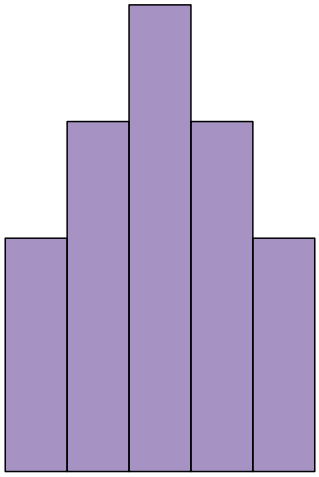
We care about the *shape* of our data, this is the primary purpose of a histogram. So we want to *not* fail at that task. We care about the shape of our data because it can help us observe the **distribution** of our data. Below are some examples of possible data shapes you may encounter— please note that this is not an exhaustive list but rather the most common ones to encounter in the wild.

Shape	**Description**
Symmetric	Mirror image on both sides of it's center
Unimodal	One peak/hump
Bimodal	Two peaks/humps
Positively-skewed	Long, narrow tail to the right
Negatively-skewed	Long, narrow tail to the left
Uniform	A flat box

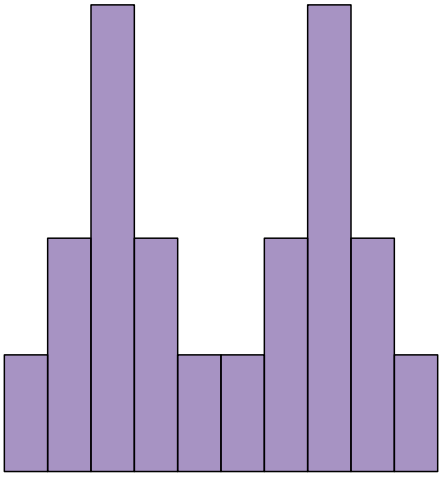
Inevitably we'll learn why these shape descriptors are sufficient to handle the majority of data we might encounter. And just as we typically learn how to summarize complex thoughts and feelings into short phrases, we'll learn some better ways to describe unique types of data besides smashing these adjectives together.

For now though, it's important to know how to recognize these shapes. We want to be able to refer back to something simple and foundational when we eventually encounter something foreign.

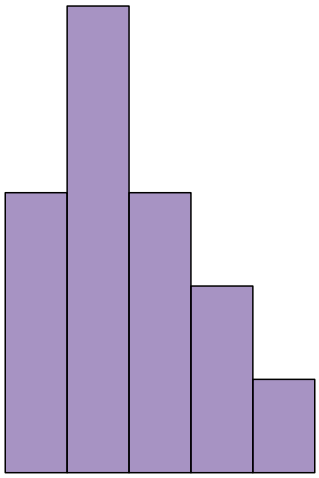
Symmetric, Unimodal



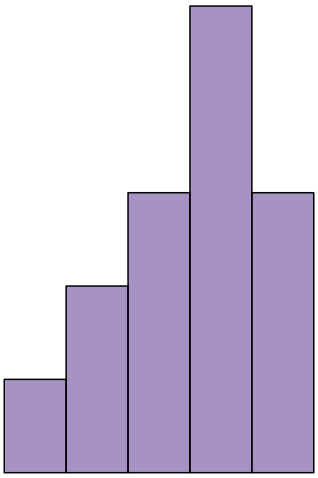
Symmetric, Bimodal

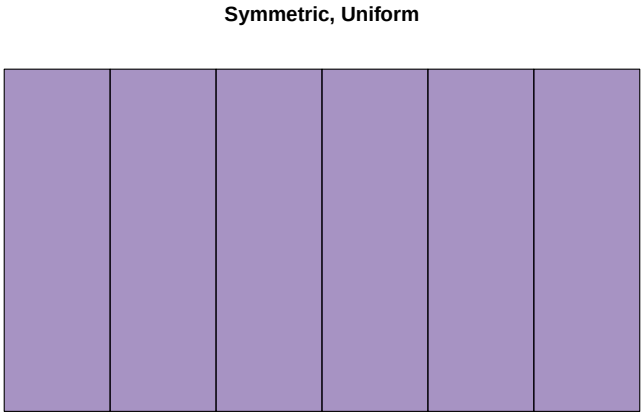


Positively Skewed, Unimodal



Negatively Skewed, Unimodal





Histograms, as we've previously established, are the primary method for visualizing single quantitative variables. It's important to know the history of any science, thus we'll address the other methods for visualizing single quantitative variables despite them being somewhat useless in modern science.

Stem-and-leaf plots

Consider that each observation has *at least two* digits, where single digit values have a leading zero (i.e., $9 = 09$.

- The digit furthest to the *right* is the "leaf"
- The digits to the *left* form the "stem"

This below table is our toy data set:

87	7	95	76	32	28	84	98	93	88
78	100	68	76	55	65	42	57	77	96

And as a stem and leaf plot:

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

We describe the shape of the data the exact same way we do with a histogram because this is essentially a histogram turned on it's side.

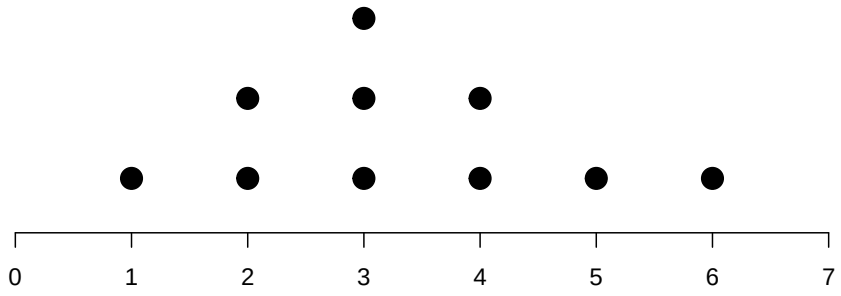
Dotplots

Dotplots and stem-and-leaf plots are a product of needing to visualize smaller datasets, which is a relic of the past when the entire science was done (essentially) by hand. Dotplots are just histograms that visualize small, granular data sets.

Given a toy data set:

3	6	2	5	1
2	3	4	3	4

Our dotplot visualization looks strikingly familiar:

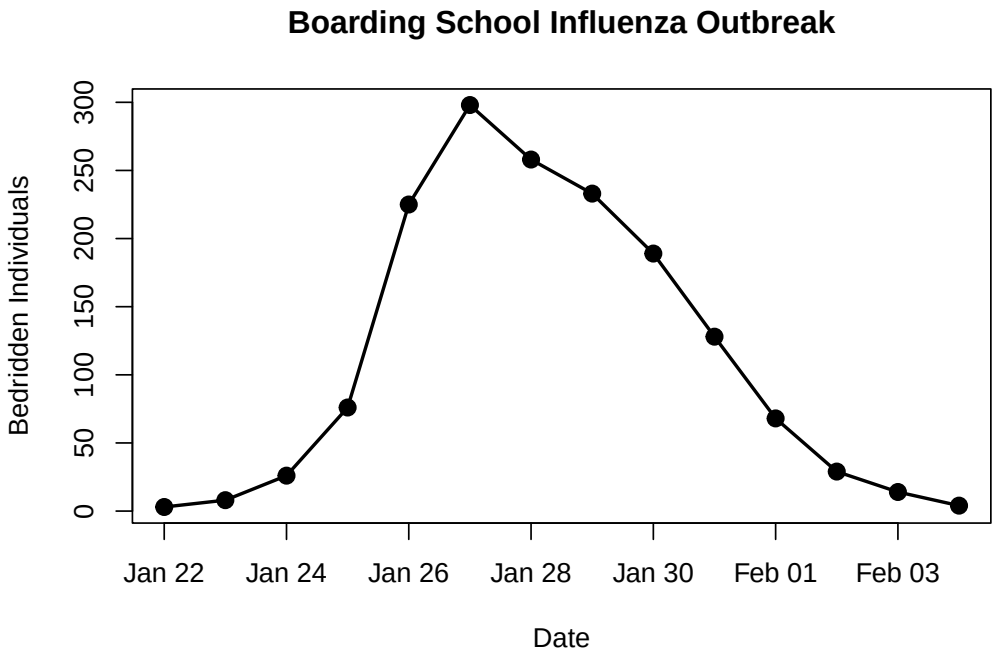


We can see how this is easily developed into a histogram. As such, we shouldn't spend much time worrying about this or stem-and-leaf plots. We'd rather spend our efforts defining smaller classes for histograms.

Time Plots

Statisticians refer to working with time as "Temporal Statisticians". Temporality is fairly important considering *most things* happen over a period of time, (this is of particular interest with disease modeling since the variables most capable of predicting disease spread are location and time).

We can represent events or changes over a period of time with a scatterplot using some small adjustments.



As always, there aren't rules so much as best practices:

- Time should always be on the horizontal scale
- Your measured variable should be on the vertical scale
- Generally, you want to include points
  - There should typically be a line connecting points

The nice thing about time plots is that we can easily ask questions of the data *in reference to* time of occurrence:

i.e. How many students were bedridden on Jan 30?