Chapter 2 Summarizing Data

Instructor: Joyce Fu

University of California, Riverside
materials adapted from and kindly shared by Lauren Cappiello

January 9, 2020

Chapter 2 Summarizing Data

Chapter 2 is all about summarizing data through summary statistics and graphs. We can get a lot of information out of these things!

These concepts are also important foundations for the rest of the course.

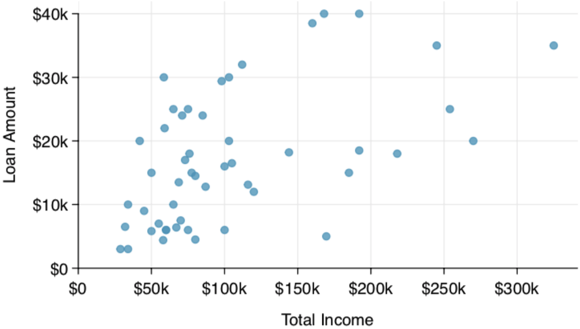
Numerical Data

Let's start by thinking of a simple numeric variable: how many hours of sleep everyone had last night. How could you describe of summarize

this data?

2.1.1 Scatterplots

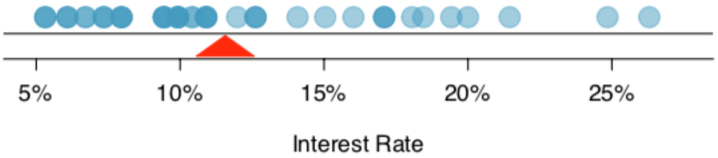
A **scatterplot** shows a case-by-case view of two numerical variables.



What can we learn from the scatterplot?

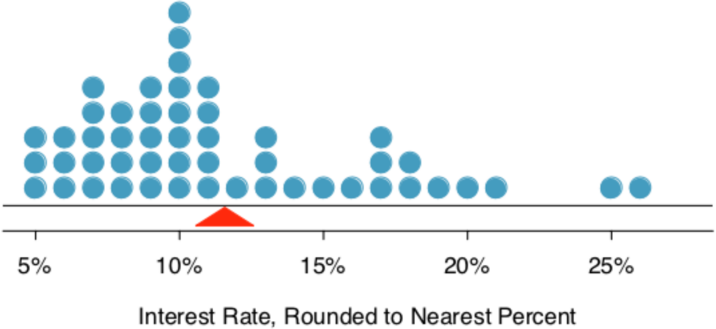
2.1.2 Dot Plots

A **dot plot** is like a scatterplot with only one variable. It shows how a single, *continuous* numerical variable falls on a number line.



2.1.2 Dot Plots

A **stacked dot plot** shows the same information for a *discrete* numerical variable.

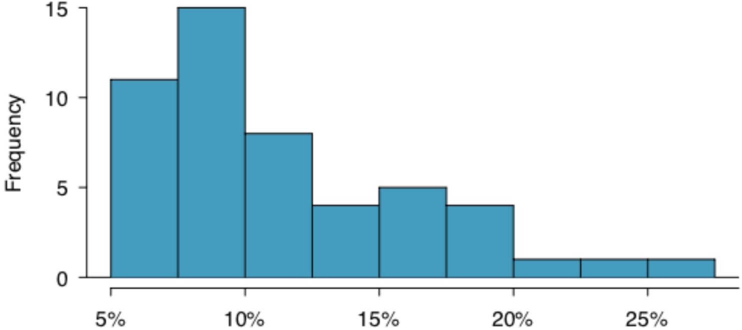


2.1.3 Histograms

A **histogram** is similar to a dot plot, but instead of showing the exact value for each observation, values are put into **bins**.

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27
Count	11	15	8	4	···	1

Figure 2.5: Counts for the binned `interest_rate` data.



2.1.3 The Mean

For example, if we had a variable called `hours of sleep` with the values 9, 5, 6, 7, 6, and 3, the mean would be

$$\frac{\text{sum of values}}{\text{total \# of observations}} = \frac{9+5+6+7+6+3}{6}.$$

We denote the mean by $\boldsymbol{\bar{x}}$. In this case, $\bar{x} = 6$

2.1.3 The Mean

In math notation, the formula for the mean looks like this:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

In our example, $n = 6$ observations and each x_i is one of the hours of sleep.

2.1.3 Measures of Center

The mean is a common way to measure the center (middle) of the **distribution** of the data.

You can think of the distribution as the way that the data is *distributed* from left to right on a histogram.

2.1.3 Measures of Center

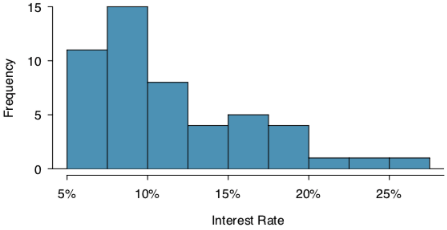
The mean of a variable is denoted by $\bar{x}$. This is what we refer to as the **sample mean**.

The mean of the entire population is typically something that we don't have exact data on (we usually don't have data for every single member of a population). Instead, we estimate the population mean using a sample mean.

The **population mean** is denoted by $\boldsymbol{\mu}$. This is the Greek letter *mu*.

2.1.3 Shape

Remember our histogram?



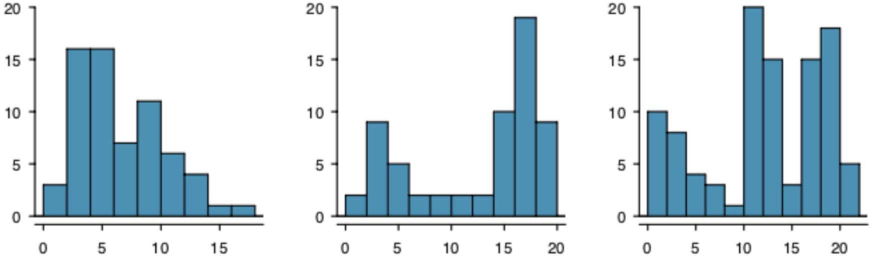
- When data tail off to the right, we say that the shape is **right skewed**.
- When data with long, thin tail on the left, so we say that the shape is **left skewed**.
- If the data have roughly equal tails, we say the distribution is **symmetric**.

2.1.3 Mode

A **mode** is any prominent peak in the distribution. These can be found in a histogram!

- A distribution with one prominent peak is called **unimodal**.
- Distributions with two prominent peaks are **bimodal**.
- Distributions with three or more promiment peaks are **multimodal**.

2.1.3 Modes



How many modes are there in each distribution?
Remember that we only count *prominent* peaks.

2.1.3 Modes

Bin widths, our particular sample, and differing opinions can all impact where we see a ”prominent” mode.

...but that is okay! The goal of examining the shape of our data is simply to better understand the nature of our data. This allows us to make more informed technical decisions down the line.

2.1.4 Which mutual fund you want to invest in ?

Suppose we want to select a mutual fund to invest in. The annual returns in the past 6 years are both averaged to be 4 (%)

mutual fund 1:	3.8, 3.9, 4.1, 4.1, 3.9, 4.2
mutual fund 2:	4.2, 1, 7, 2.5, 5.5, 3.8

In both cases, we get a sample average of $\bar{x} = 4$. Do you have any preference between these two funds?

2.1.4 Variability

Deviation from the mean:

The distance between an observation and its mean is called the **deviation**. From mutual fund 1 (3.8, 3.9, 4.1, 4.1, 3.9, 4.2), the deviations for each observations are

$$\begin{aligned} x_1 - \bar{x} = 3.8 - 4 = -0.2, x_4 - \bar{x} &= 4.1 - 4 = 0.1 \\ x_2 - \bar{x} = 3.9 - 4 = -0.1, x_5 - \bar{x} &= 3.9 - 4 = -0.1 \\ x_3 - \bar{x} = 4.1 - 4 = 0.1, x_6 - \bar{x} &= 4.2 - 4 = 0.2 \end{aligned}$$

If we add up all of the deviations for a sample, what will happen?

2.1.4 Variability

There are two simple ways to get rid of the signs to focus on distance (without direction).

1. Take the absolute value of the number.
2. Square the number.

It turns out that there are a whole lot of mathematical reasons why it's easier to work with squares than with absolute values!

2.1.4 Variance

Let's return to our mutual fund example, we calculate take a square of each deviation, then get the average.
(in class calculation)

2.1.4 Variance

Here comes the formula of variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note: Technically, we divide by $n - 1$ instead of by n. We may talk more about this later, but in the meantime just know that there's some mathematical nuance that makes the variance formula a little bit more complicated.

2.1.4 Standard Deviation

The variance can be described as the average squared distance from the mean. That probably doesn't sound like a very intuitive way to measure variability.

However, the **standard deviation** is easier to conceptualize than the variance: it gets at our original goal of estimating how far a typical observation is from the mean.

2.1.4 Standard Deviation

Fortunately for us, the standard deviation doesn't require any additional mathematical nuance! In order to calculate the standard deviation, we simply take the square root of the variance.

Returning again to our example, mutual fund 1 has standard deviation

$$s = \sqrt{s^2} = \sqrt{0.024} \approx 0.155$$

mutual fund 2 has standard deviation

$$s = \sqrt{s^2} = \sqrt{4.516} \approx 2.125$$

2.1.4 Population Variability

Like the mean, the **sample variance** and **sample standard deviation** also have population counterparts.

- The **population variance** is denoted $\boldsymbol{\sigma^2}$.
- The **population standard deviation** is denoted $\boldsymbol{\sigma}$.

σ is the Greek letter *sigma*. (We often use Greek letters to denote values from our population.)

2.1.4 Symbols to Remember

Let's update our list with variance and standard deviation.

- n: number of observations/cases
- $\bar{x}$: sample mean
- μ: population mean
- s^2: sample variance
- s: sample standard deviation
- σ^2: population variance
- σ: population standard deviation

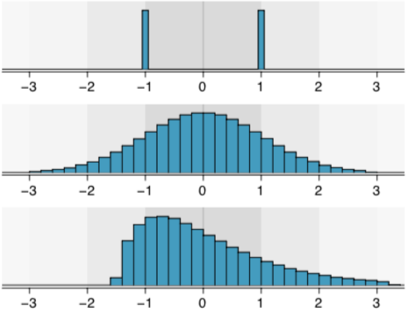
2.1.4 Mean, Standard Deviation, and Shape

Mean, standard deviation, and shape together give us a good description of our distribution.

- If any one of these is missing, we miss crucial information.
- Without the mean, we lack information about the center of the distribution.
- Without the standard deviation, we are unable to capture how spread out the data are.

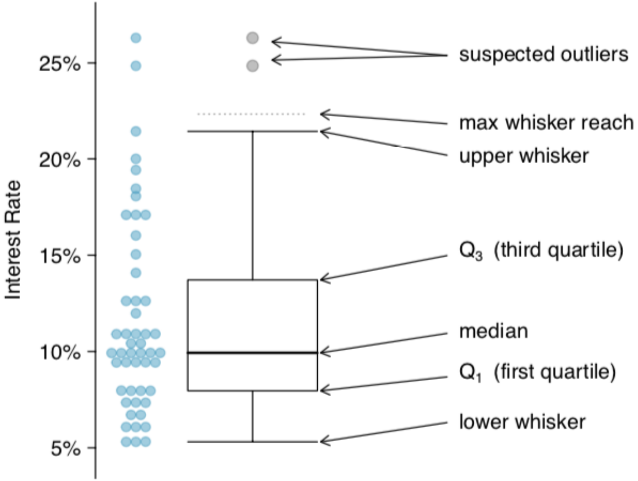
2.1.4 Why Shape?

These three distributions have the same mean ($\bar{x} = 0$) and standard deviation ($s = 1$)!



A good description of shape should include modality and skewness (or symmetry). To give an even clearer picture, we can report where the modes are and the sharpness of the peaks.

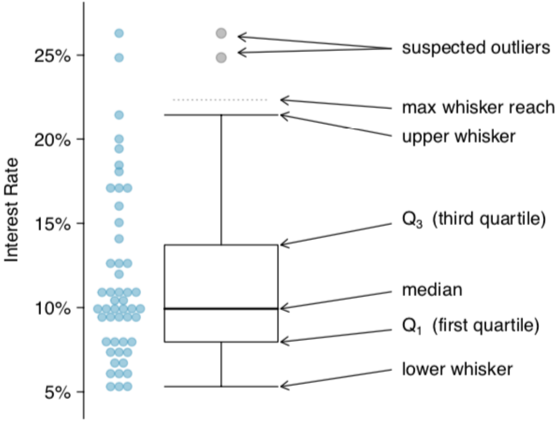
2.1.5 Box Plots



A stacked dot plot next to a vertical box plot.

2.1.5 The Median

The first step in constructing a box plot is to draw a line at the median.



2.1.5 The Median

- The **median** takes the data and splits it in half.
- The median is also called the **50th percentile** because 50% of the data is below this value.
- The median is another **measure of center**.
- To find the median, we sort our numerical variable and then find the halfway point.

2.1.5 The Median

If we have an odd number of observations, say,

$$1, 2, 3, 4, 5$$

we take the observation in the middle (the $\frac{n+1}{2}$th observation).

In this case,

$$1, 2, \mathbf{3}, 4, 5$$

3 is the median.

2.1.5 The Median

- If we have an even number of observations

  $$1, 2, 3, 4, 5, 6$$

  we cut the data exactly in half

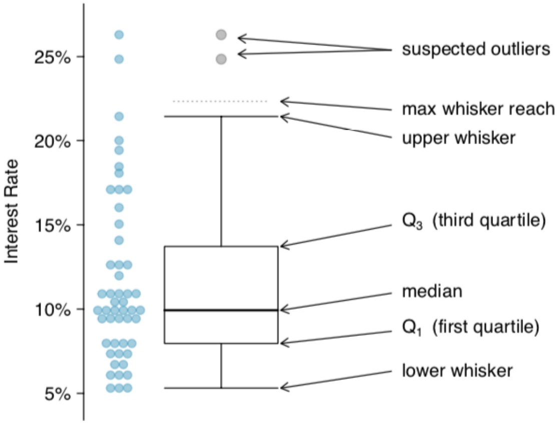
  $$1, 2, 3 \quad | \quad 4, 5, 6$$

  and the median is the average of the two observations closest to the halfway point

  $$\frac{3+4}{2} = 3.5$$

2.1.5 Quartiles

The next step in our box plot is to draw a box connecting the first and third quartiles.



2.1.5 Quartiles

Quartiles split our data into *quarters*.

- 25% of the data falls below the **first quartile** (Q1).
  - This is the 25th percentile.
- 50% of the data falls below the median.
- 75% of the data falls below the third quartile (Q3).
  - This is the 75th percentile.

What percent of the data falls between Q1 and the median? What percent between Q1 and Q3?

2.1.5 Finding Quartiles

1. Find the median.
2. Take all of the data that falls *below* the median and find the middle of that data using the same steps we used to find the median. This is the first quartile.
3. Repeat with the data that falls *above* the median. This is the third quartile.

2.1.5 Interquartile Range

- The distance between the first and third quartiles is referred to as the **interquartile range** (or IQR).
- This value is easy to calculate!

$$IQR = Q3 - Q1$$

- The IQR is another measure of variability.

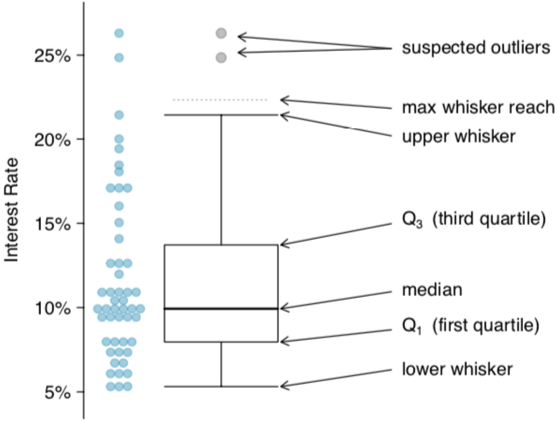
2.1.5 Whiskers

Now we need to find the whiskers.

Image from BBC Wildlife
www.discoverwildlife.com/animal-facts/mammals/how-do-whiskers-work/

2.1.5 Whiskers

Now we need to find the whiskers.



2.1.5 Whiskers

The **whiskers** capture (most of) the rest of the data.

- Each whisker is no longer than

$$1.5 \times IQR.$$

  and stops at the point closest to, but still within, this range.

2.1.5 Whiskers

- The upper whisker goes no farther than

  $$Q3 + 1.5 \times IQR$$

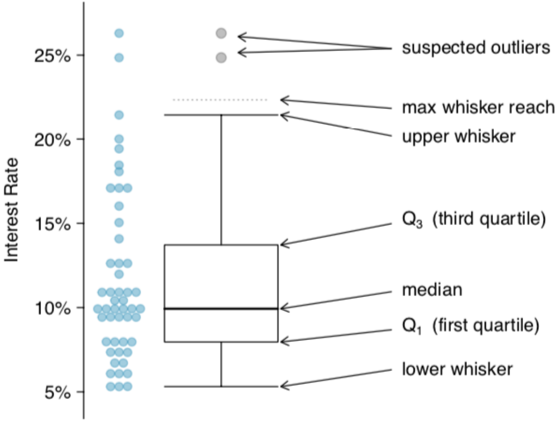
  and the lower whisker no farther than

  $$Q1 - 1.5 \times IQR$$

We may choose not to include the maximum upper reach and minimum lower reach on our box plot, but we always include the whiskers themselves.

2.1.5 Outliers

Finally, we add any outliers by labeling each one with a dot.



2.1.5 Outliers

Since we've already built the rest of our boxplot, we can start to think about outliers as whatever is left out.

- We label these observations specifically because they are *unusual* or *extreme*.
- Observations that are unusually far from the rest of the data are referred to as **outliers**.

2.1.6 Why Examine Outliers?

- Identify sources of strong skew.
- Provide insight into potentially interesting properties of the data.
- Identify possible data collection or data entry errors.

2.1.6 Robust Statistics

Would outliers affect the median a lot? Would outliers affect the first, and third quartile a lot? The median and the IQR are called **Robust Statistics** because extreme observations have little effect on their values. The mean and standard deviations are more heavily influenced by changes in extreme observations.

2.1.6 When Are Robust Statistics Important?

- Suppose you wanted to know about the typical home price in the United States in 2018.
- Recall that the mean and median are both measures of center.
- Would you look at the mean or the median? Why?

2.1.6 When Are Robust Statistics Important?

As long as you can defend your answer, there is value to each option!

- If we wanted to know what the typical homeowner is spending, the median would be more useful.
- If we wanted our estimate to scale, e.g., to estimate how much total money was spent on homes in 2018, the mean might be a better option.

2.1.7 Transforming Data

- When data are very strongly skewed, we sometimes transform them to make them easier to model.
- For our purposes, data is easiest to model when it is
  - Mostly symmetric
  - Unimodal
  - "Bell-shaped"
- We want to be able to use our mean and standard deviation instead of our median and IQR!

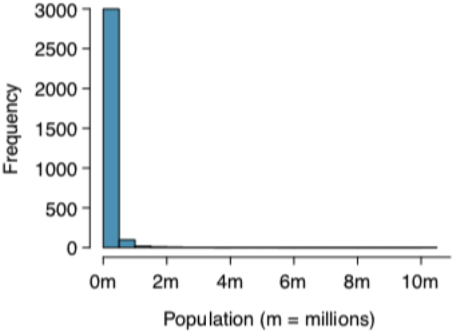
2.1.7 Transforming Data

What does it mean to ”transform” the data?

- Essentially, we apply some mathematical function to our data in order to rescale it.
- Technically, we want transformations that are continuous and invertible.
- Fortunately, there are a number of standard transformations that we use.

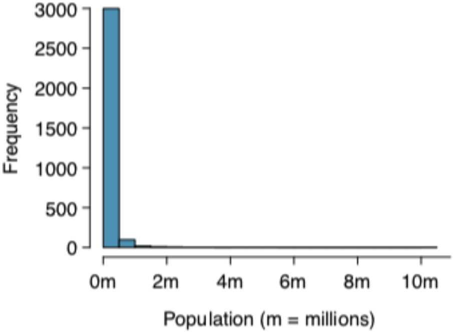
2.1.7 Example: Transforming Data

A histogram of the populations of all US counties.



For perspective, Riverside County has 2.4 million people and Los Angeles County has 10.2 million people!

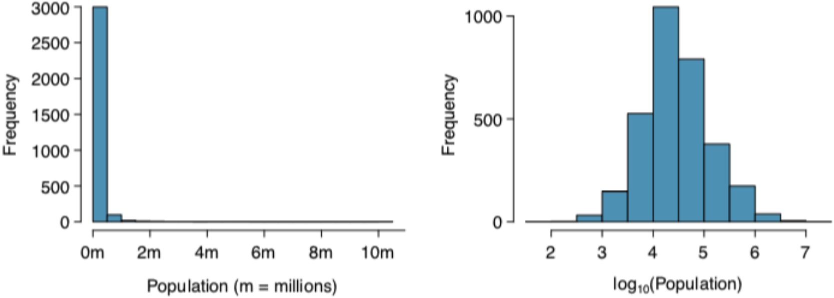
Example: Transforming Data



These data are very strongly skewed! Almost all of the counties have populations between 0 and 1 million people, but a few have over 10 million.

2.1.7 Example: Transforming Data

Before and after transformation:



(a) (b)

In histogram (b), it is much more reasonable to use the mean and standard deviation to measure the center and spread of our data.

2.1.7 Transformations

We may also apply

- A square root transformation
  - $\sqrt{\text{original variable}}$
- An inverse transformation
  - $(\text{original variable})^{-1}$

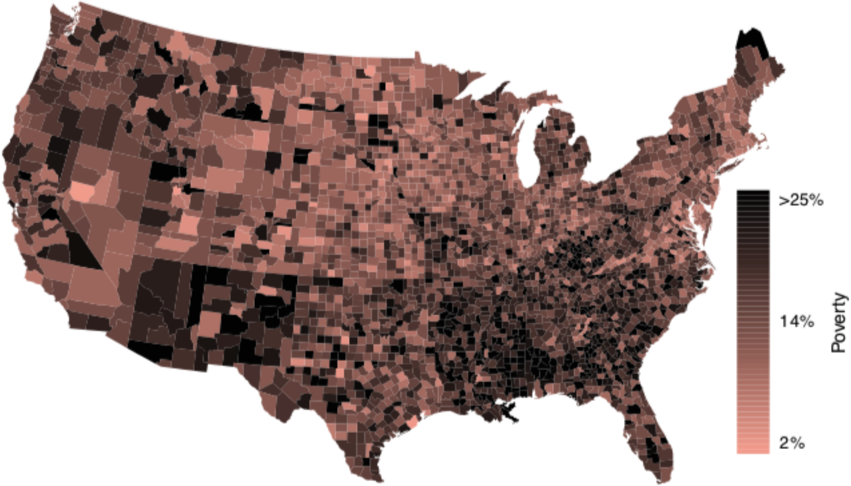
2.1.7 Transformations

In general, transformations:

- Let us see data structure differently.
- Reduce skew.
- Assist in modeling.
- Straighten nonlinear relationships in scatterplots.

2.1.8 Mapping Data

Geographic data can be plotted show higher and lower values of a variable using colors on a map.



2.2 Categorical Data

In the previous section, we focused on numerical data. We now turn our attention to categorical data.

This section includes more tools and language that we will use throughout the course.

2.1.1 Summary Tables

A basic **summary table** *summarizes* a categorical variable by showing the frequency, or count, of each category.

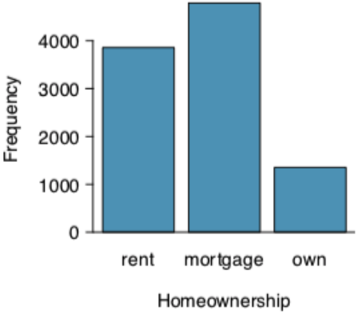
`homeownership`	Count
Rent	3858
Mortgage	4789
Own	1353
Total	10000

`apptype`	Count
Individual	8505
Joint	1495
Total	10000

Note: `homeownership` refers to whether or not someone owns a home and `apptype` indicates whether a loan application was made individually or jointly.

2.2.1 Bar Plots

A **bar plot** is a common way to visualize the information in a summary table.



2.2.1 Summary Tables: Proportions

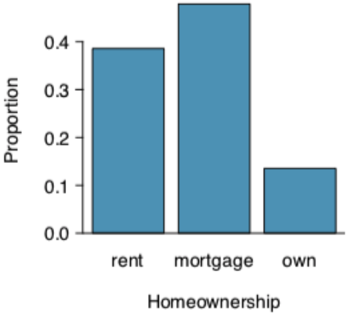
We may occasionally prefer to see our data summarized by proportions (see the fractional breakdown of our data).

homeownership	Proportion
Rent	0.3858
Mortgage	0.4789
Own	0.1353
Total	1.0000

apptype	Proportion
Individual	0.8505
Joint	0.1495
Total	1.0000

2.2.1 Bar Plots

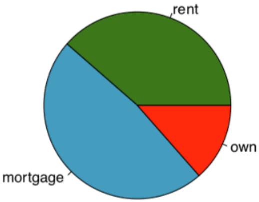
We can again use a bar plot to visualize this information.

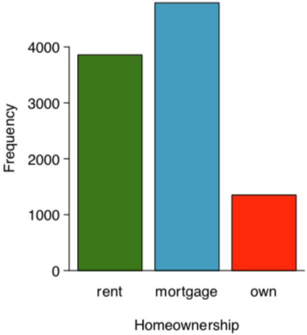


This bar plot looks exactly the same as the one with frequencies! The only difference is in the numbers along the vertical axis.

2.2.1 Pie Charts

Pie charts show the same information as bar charts, but are more difficult to discern details from.





They are good for infographics but are not well-suited to technical writing.

2.2.1 Contingency Tables

A **contingency table** is a table that summarizes two categorical variables. It looks something like this:

		homeownership			
		Rent	Mortgage	Own	Total
apptype	Individual	3496	3839	1170	8505
	Joint	362	950	183	1495
	Total	3858	4789	1353	10000

2.2.2 Row and Column Proportions

We may also want to examine the fractional breakdown of our contingency table data.

- **The row proportions are the row counts divided by the row total**.
- The column proportions are the column counts divided by the column total.
- The overall proportions are the counts divided by the total number of observations.

2.2.2 Contingency Tables for Row Proportions

We can now convert our previous contingency table into a contingency table *for the row proportions*:

		homeownership			
		Rent	Mortgage	Own	Total
apptype	Individual	0.411	0.451	0.138	1.000
	Joint	0.242	0.635	0.122	1.000
	Total	0.386	0.479	0.135	1.000

This breaks down each application type into home ownership status. We would say that, *among individual applications*, 41.1% are renters.

2.2.2 Row and Column Proportions

- The row proportions are the row counts divided by the row total.
- **The column proportions are the column counts divided by the column total**.
- The overall proportions are the counts divided by the total number of observations.

2.2.2 Contingency Tables for Column Proportions

We can also convert our contingency table into a contingency table *for the column proportions*:

		`homeownership`			
		Rent	Mortgage	Own	Total
`apptype`	Individual	0.906	0.802	0.865	0.851
	Joint	0.094	0.198	0.135	0.150
	Total	1.000	1.000	1.000	1.000

This breaks down each home ownership status into application types. We would say that, *among renters*, 90.6% filled out an individual loan application.

2.2.2 Contingency Tables for Row Proportions

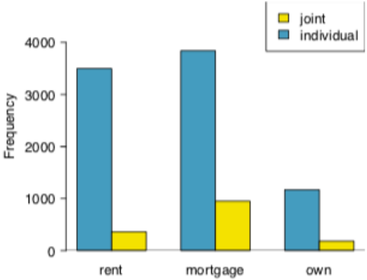
	Rent	Mortgage	Own	Total
Individual	0.906	0.802	0.865	0.851
Joint	0.094	0.198	0.135	0.150
Total	1.000	1.000	1.000	1.000

- We can use these contingency tables to check for an association between home ownership and loan type.
- Notice that, among individual applicants, 90.5% rent, but only 80.2% have a mortgage.

2.2.3 Two-Variable Bar Plots

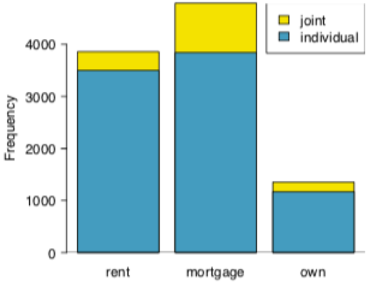
- We can extend our bar plots to help visualize the information in a contingency table by creating
  - **Stacked bar plots**.
  - **Side-by-side bar plots**.
- A stacked bar plot takes our one-variable bar plot and breaks up the bars to show a second variable.
- A side-by-side bar plot takes our one-variable var plot and splits each bar into two side-by-side bars.

2.2.3 Side-By-Side Bar Plots



This side-by-side bar plot shows home ownership with loan application type. Here, we're breaking the data into six categories and giving each one a bar.

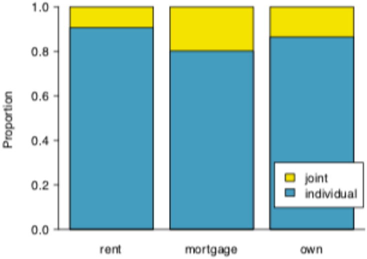
2.2.3 Stacked Bar Plots



This stacked bar plot shows home ownership broken down by loan application type.

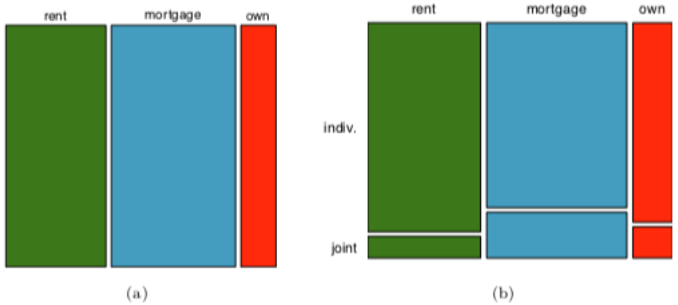
In both plots, it is easy to see that there are fewer people who own their homes and fewer people applying for joint loans.

2.2.3 Stacked Bar Plots: Frequencies



- Same information, but standardized based on home ownership.
- This is a visualization of the frequency-based contingency table for loan types varying between levels of home ownership (slide 30).
- Now we can see that the two variables are associated.

2.2.4 Mosaic Plots



(a)

(b)

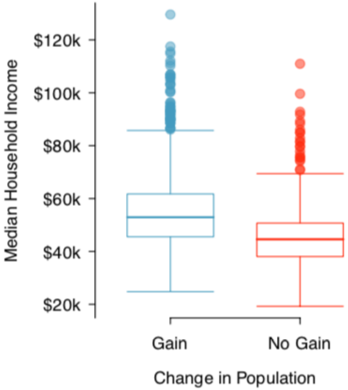
(a) is a one-variable mosaic plot for `homeownership`.
(b) is a two-variable mosaic plot for `homeownership` and `app_type`.

2.2.4 Mosaic Plots

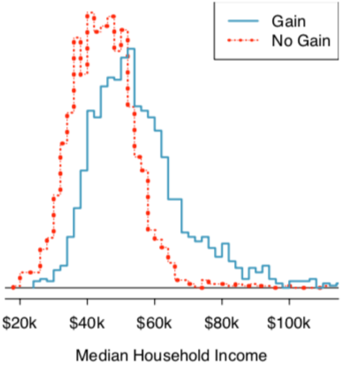
- Mosaic plots look a lot like bar plots, but now the *widths* of the bars depend on the group sizes.
- For two-variable mosaic plots, the boxes from the one-variable mosaic plot are divided up using the second variable.
- Now, the *heights* of the boxes also depend on group sizes.
- Thus, mosaic plots use *area* to represent the number of cases in each category.

2.2.6 Side-By-Side Box Plots

Side-by-side box plots are standard tools for visualizing numerical data broken down into categories.



2.2.6 Hollow (or Stacked) Histograms



Hollow histograms are a little bit harder to read, but they allow us to visualize what two distributions look like when layered on top of each other.