Categorical Data

July 31, 2019

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!

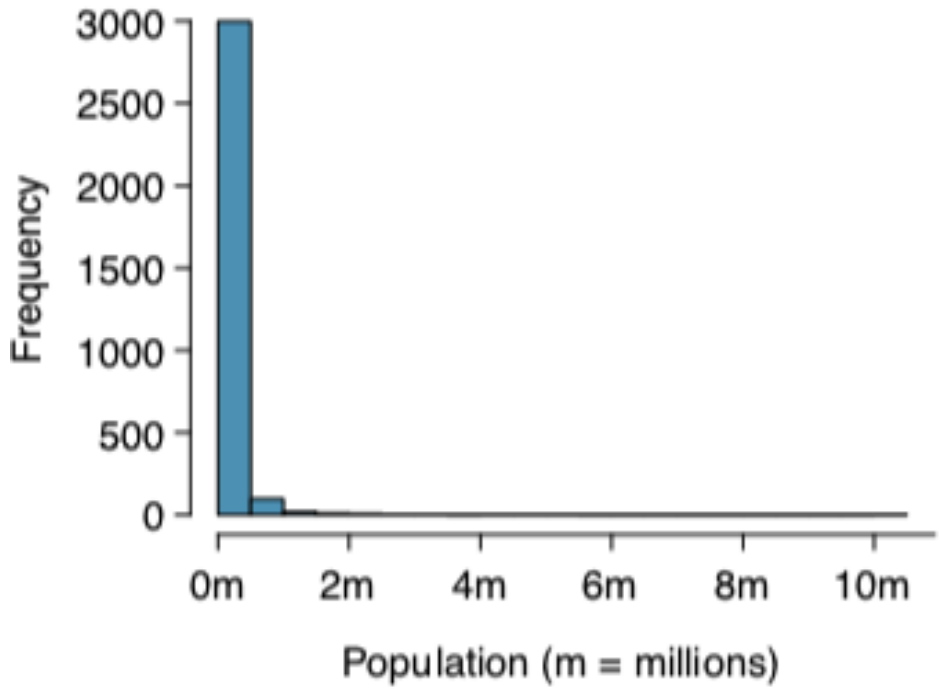
Transforming Data

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.

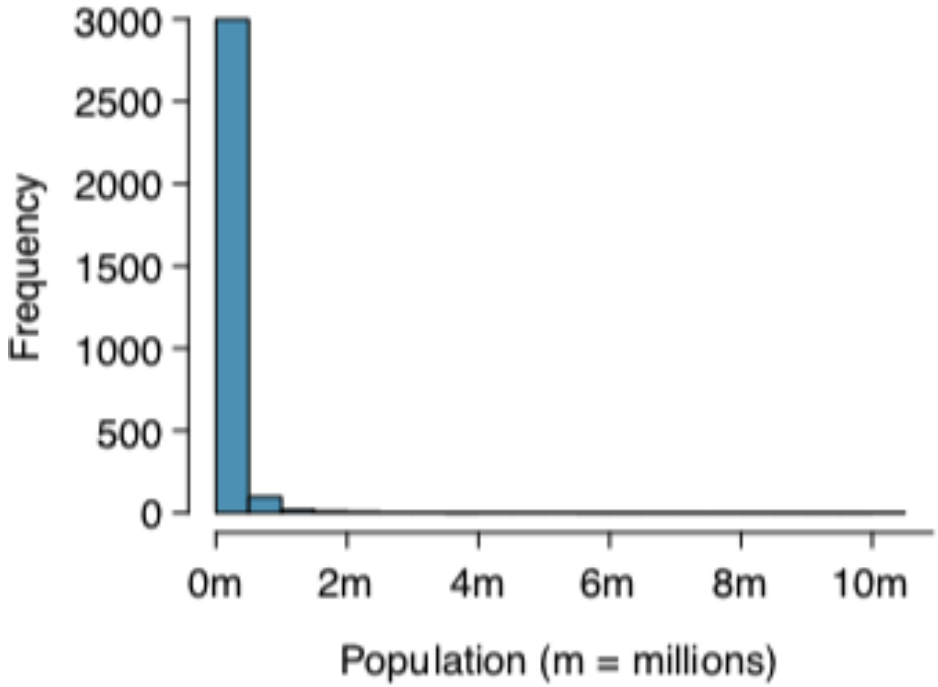
Example: Transforming Data

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!

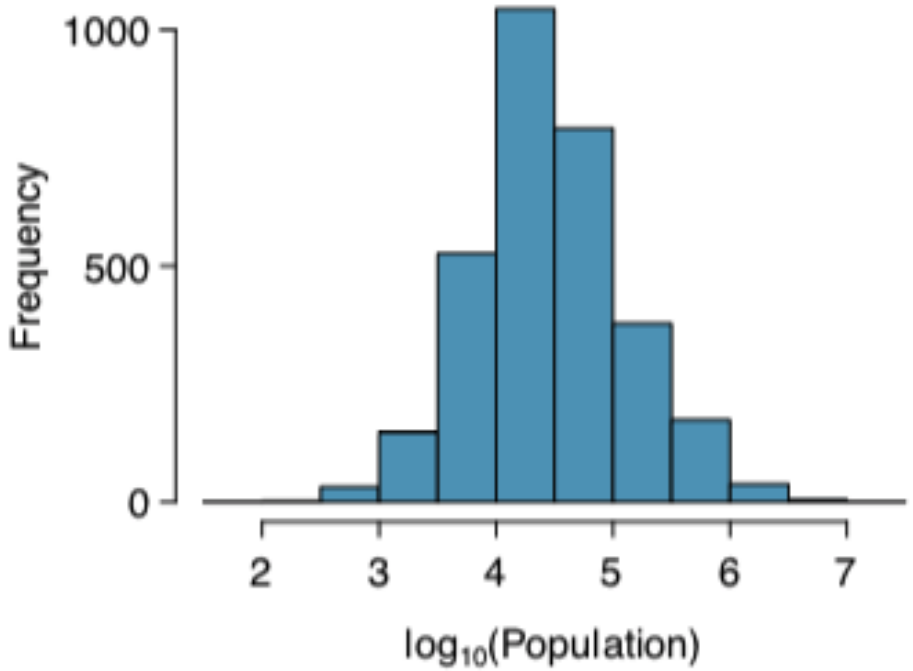
Example: Transforming Data



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.

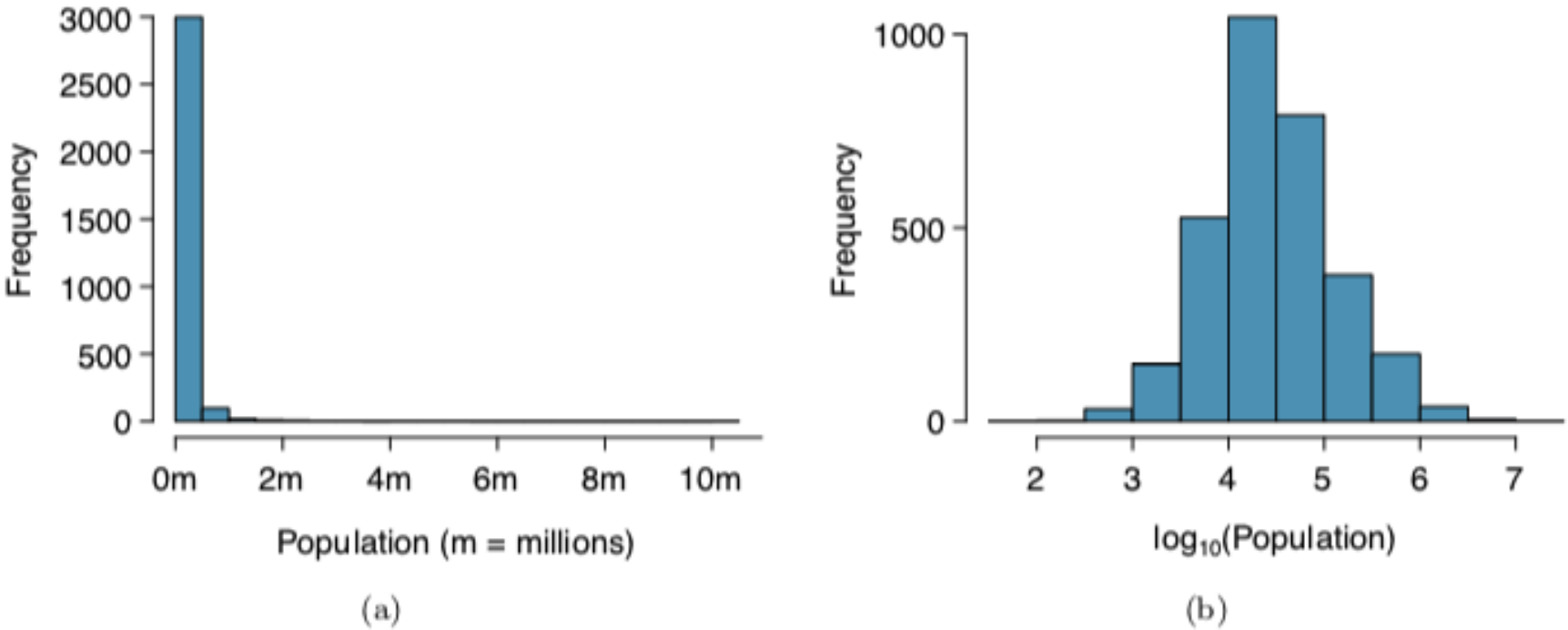
Example: Transforming Data

To transform the data we take $\log_{10}(Population)$. The histogram of the transformed data looks like this:



Example: Transforming Data

Before and after transformation:



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

Transformations

We may also apply

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

Transformations

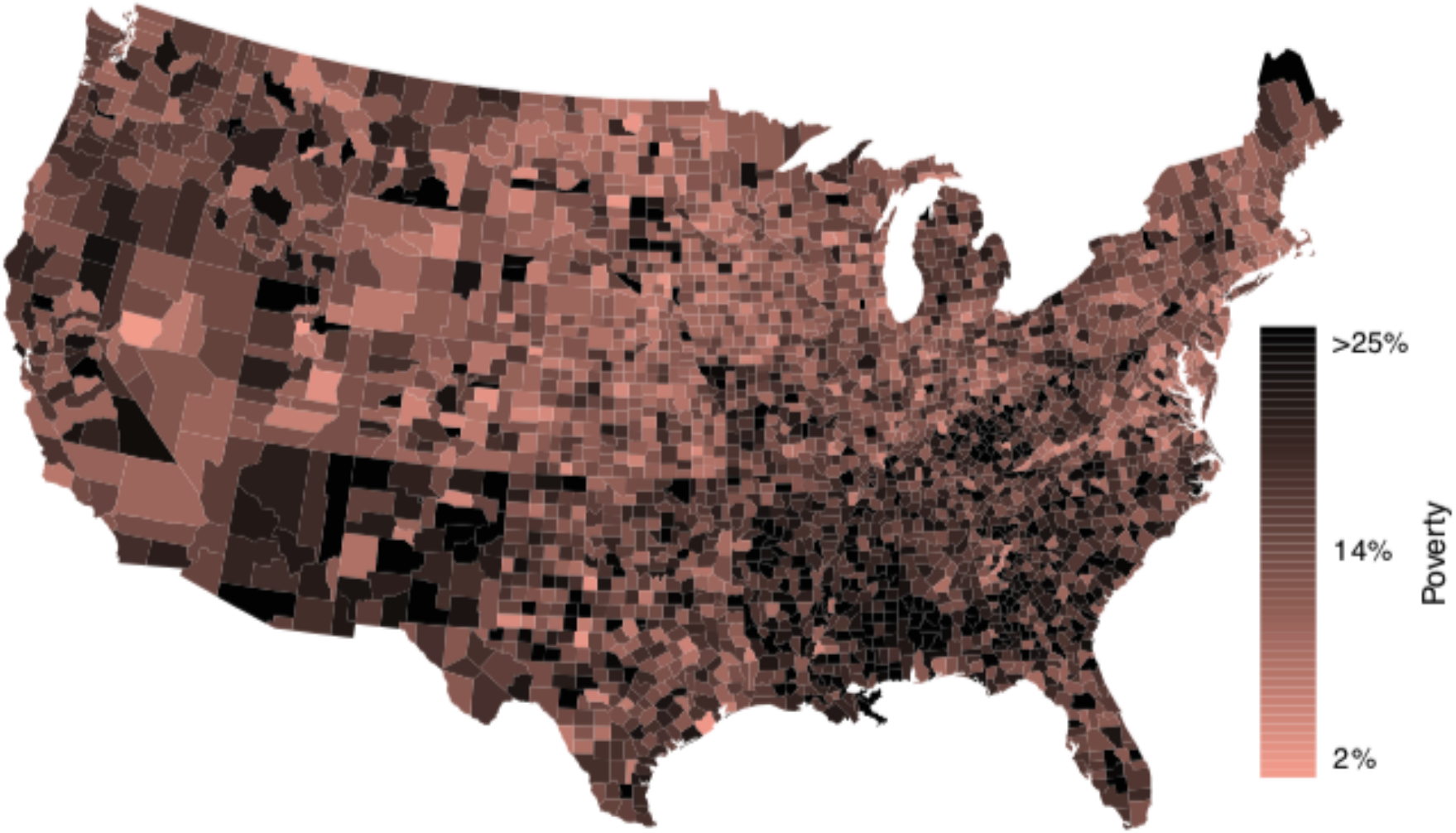
In general, transformations:

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

Visualizing Geographic Data

- Geographic data can be plotted using the data visualization techniques we've already seen.
- We might instead want to create an intensity plot.
- These plots allow us to show higher and lower values of a variable using colors on a map.
- Intensity plots are good for seeing geographic trends.

Mapping Data



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.

Word Clouds

If we have text that we're interested in, we can turn words into categories. Here are the top seven words from the survey question about slaying a dragon:

Word	Frequency
sword	9
dragon	9
stab	6
kind	5
heart	4
fire	3
dont	3

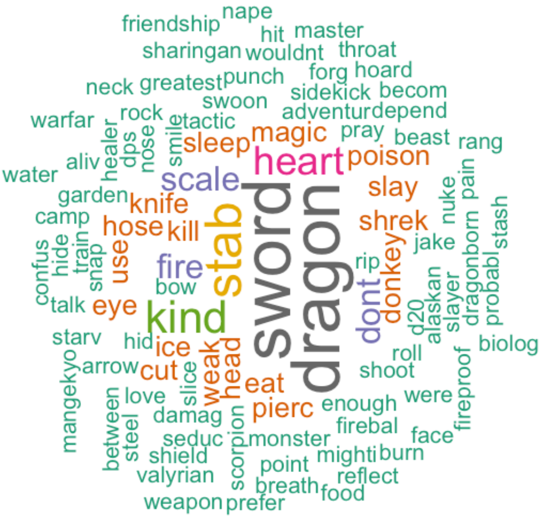
Word Clouds

Here are a few things I did before finding the top words:

- Removed responses like ”N/A” and ”I don’t know”.
- Removed low-information words like ”the” and ”and”.
- Removed punctuation.
- Converted all text to lowercase.
- Reduced words to their roots - ”kindness” becomes ”kind” - to group those words together.

Now we’re ready to create a word cloud out of the responses.

Word Clouds


friendship
nape
hit
master
sharingan
wouldnt
throat
forg
hoard
neck
greatest
punch
sidekick
becom
swoon
adventur
depend
warfar
rock
tactic
magic
pray
beast
rang
healer
dps
nose
smile
sleep
heart
poison
water
aliv
scale
slay
pain
garden
knife
nuke
stash
camp
hose
kill
shrek
jake
confus
hide
train
snap
use
fire
stab
sword
dragon
rip
donkey
alaskan
slayer
dragonborn
probabl
bow
dont
talk
eye
kind
d20
starv
hid
ice
weak
head
roll
biolog
mangekyo
arrow
cut
slice
eat
shoot
fireproof
between
love
were
steel
damag
pierc
enough
seduc
scorpion
monster
firebal
face
shield
point
mighti
burn
valyrian
breath
reflect
weapon
prefer
food

Summary Tables

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

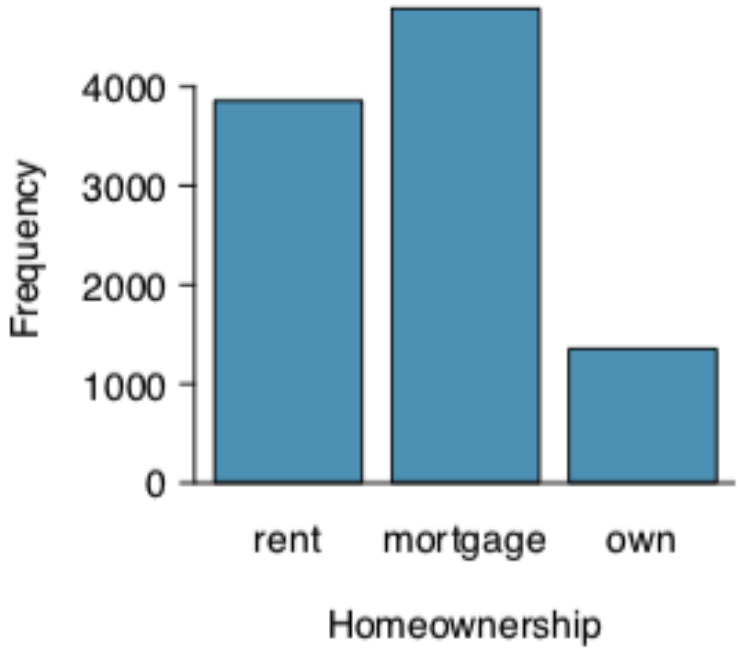
`homeownership`	Count
Rent	3858
Mortgage	4789
Own	1353
Total	10000

`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.

Bar Plots

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



Summary Tables: Proportions

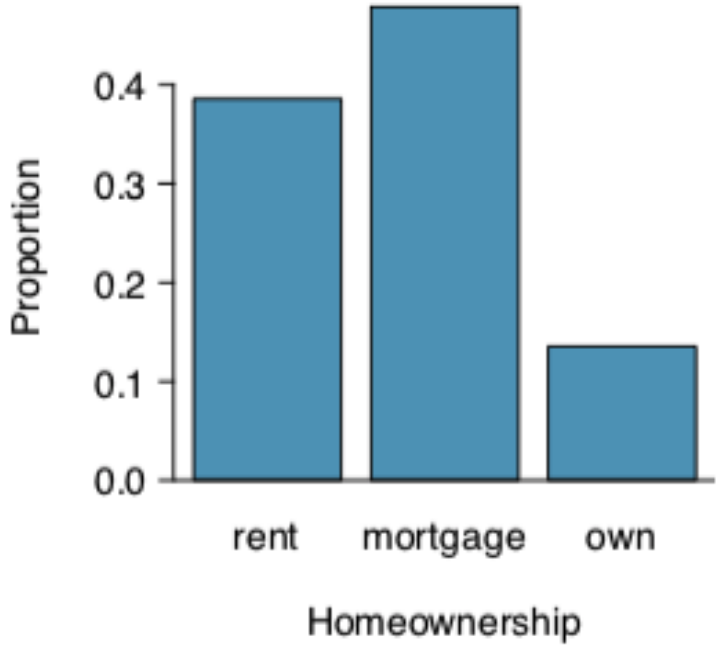
We may occasionally prefer to see our data summarized by proportions (see the fractional breakdown of our data).

`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

Bar Plots

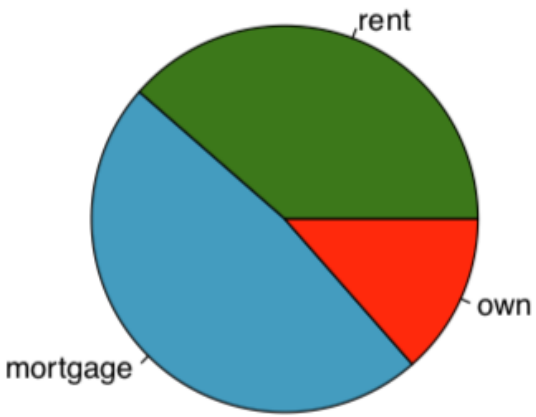
We can again use a bar plot to visualize this information.



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

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.

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

Contingency Tables

Contingency tables allow us to summarize two categorical variables together by breaking them down into subcategories.

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

Contingency Tables

Notice that the column of totals is the same as the summary table for `apptype` and the row of totals has the same information as the summary table for `homeownership`.

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

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.

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.

Contingency Tables for 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

We can tell at a glance that this is for the *row proportions* because all of the *row totals* are 1.

The rows are total breakdown of `homeownership`, so the bottom row of totals is the same as the home ownership summary table with proportions (see slide 15). They are *not* the additive total for the row of proportions.

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.

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.

Contingency Tables for Row 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

We can tell at a glance that this is for the *column proportions* because all of the *column totals* are 1.

The rows are the total breakdown of `apptype`, so the bottom row of totals is the same as the application type ownership summary table with proportions (see slide 15). They are *not* the additive total for the column of proportions.

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.

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

- If there is no association, the proportions will be (approximately) the same across the row.
- We say that loan types *vary between* different **levels** of home ownership.
- (Using the column proportions, we can also say that home ownership status varies between levels of loan type.)

Example: Student Survey

Let's look at a contingency table for some of our survey data.

		Year				
		Sophomore	Junior	Senior	Other	Total
Want	1	1	0	2	1	4
	2	0	4	3	1	8
	3	1	11	1	0	23
	4	7	10	8	1	26
	5	0	3	1	0	4
	Total	9	28	25	3	65

Is there a relationship between year and desire to take this course?

Example: Student Survey

It's hard to tell! Let's look at whether there is a change in `want` by `year` (does want vary between levels of year).

		Year				
		Sophomore	Junior	Senior	Other	Total
Want	1	0.11	0.00	0.08	0.33	0.06
	2	0.00	0.14	0.12	0.33	0.12
	3	0.11	0.39	0.44	0.00	0.35
	4	0.78	0.36	0.32	0.33	0.40
	5	0.00	0.11	0.04	0.00	0.06
	Total	1.00	1.00	1.00	1.00	1.00

Is there a relationship between year and desire to take this course?

Example: Student Survey

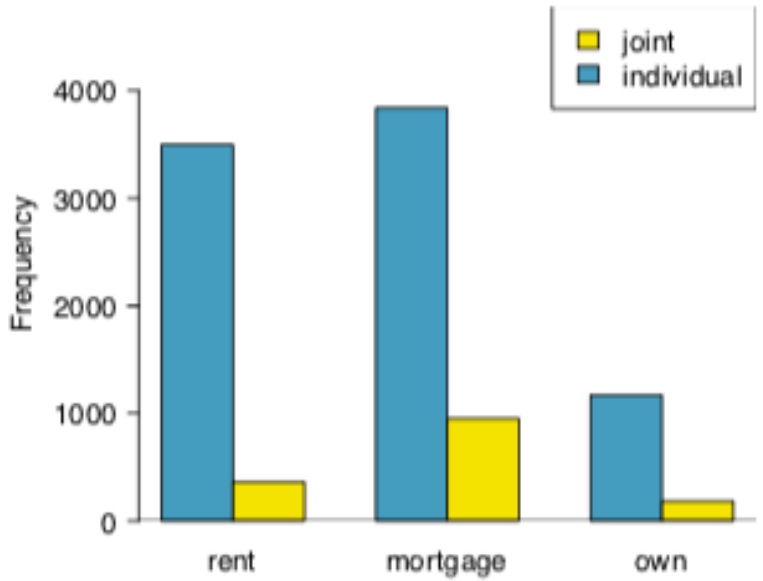
What if we looked at whether there is a change in `year` by `want` (does year vary between levels of want)?

- We should still see a relationship.
- It makes sense to think about whether year affects your desire to take this course.
- However, it probably doesn't make sense to think about whether desire to take this course affects your year in school.
  - In this scenario, you'd have to have done something extreme like taken a year off and fallen behind just because you really didn't want to take this course. Hopefully that's not the case!

Two-Variable Bar Plots

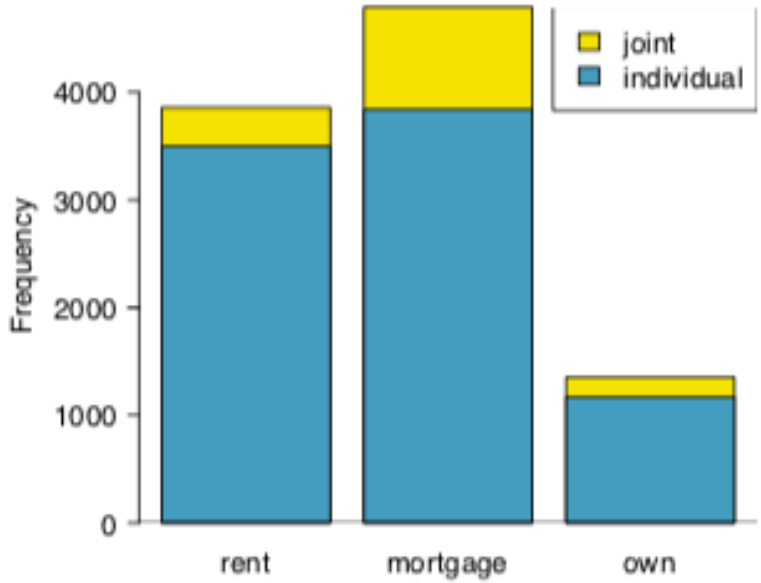
- We can extend our bar plots to help visualize the information in a contingency table by creating
  - **Stacked bar plots**.
  - **Side-by-side bar plots**.
- A stacked bar plot takes our one-variable bar plot and breaks up the bars to show a second variable.
- A side-by-side bar plot takes our one-variable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

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.

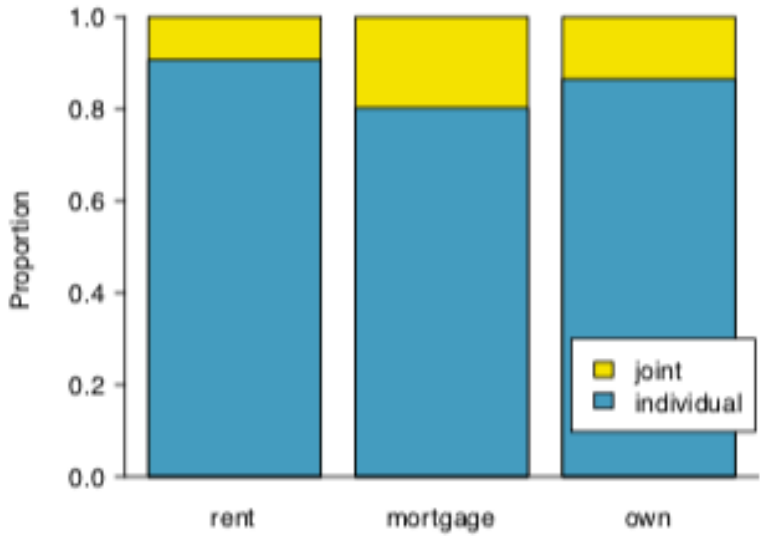
Stacked Bar Plots



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.

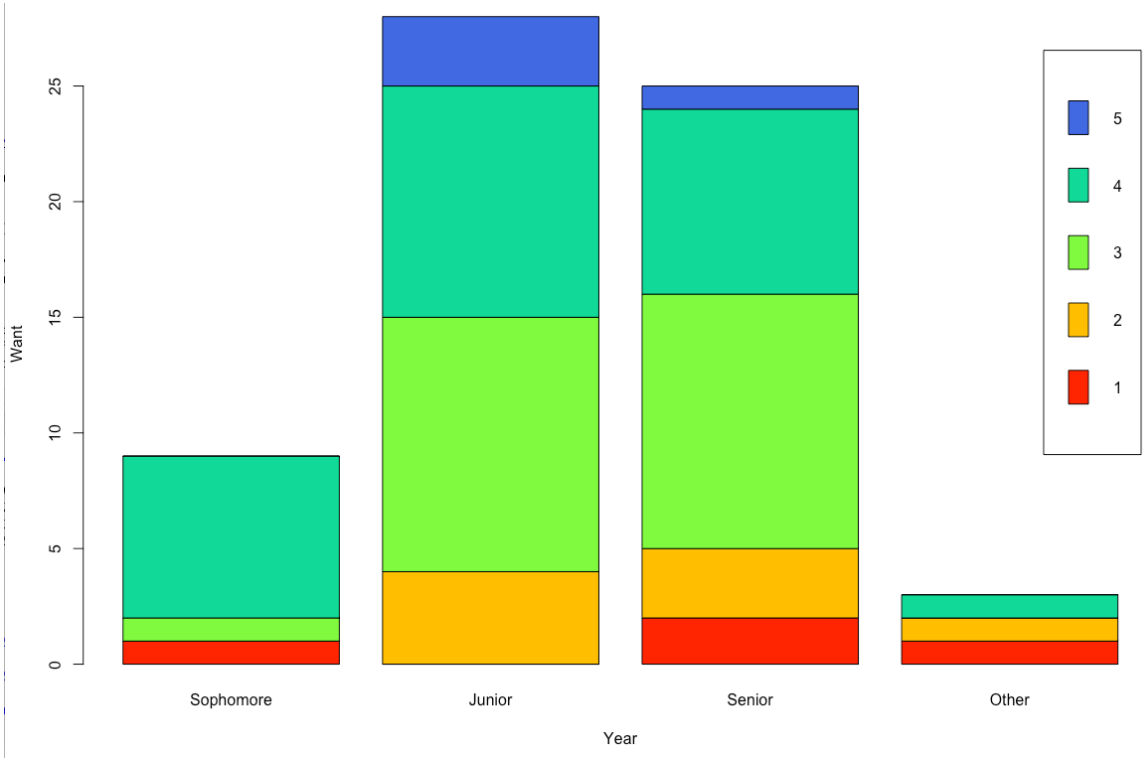
Stacked Bar Plots: Frequencies



- 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.

Example: Student Survey Data

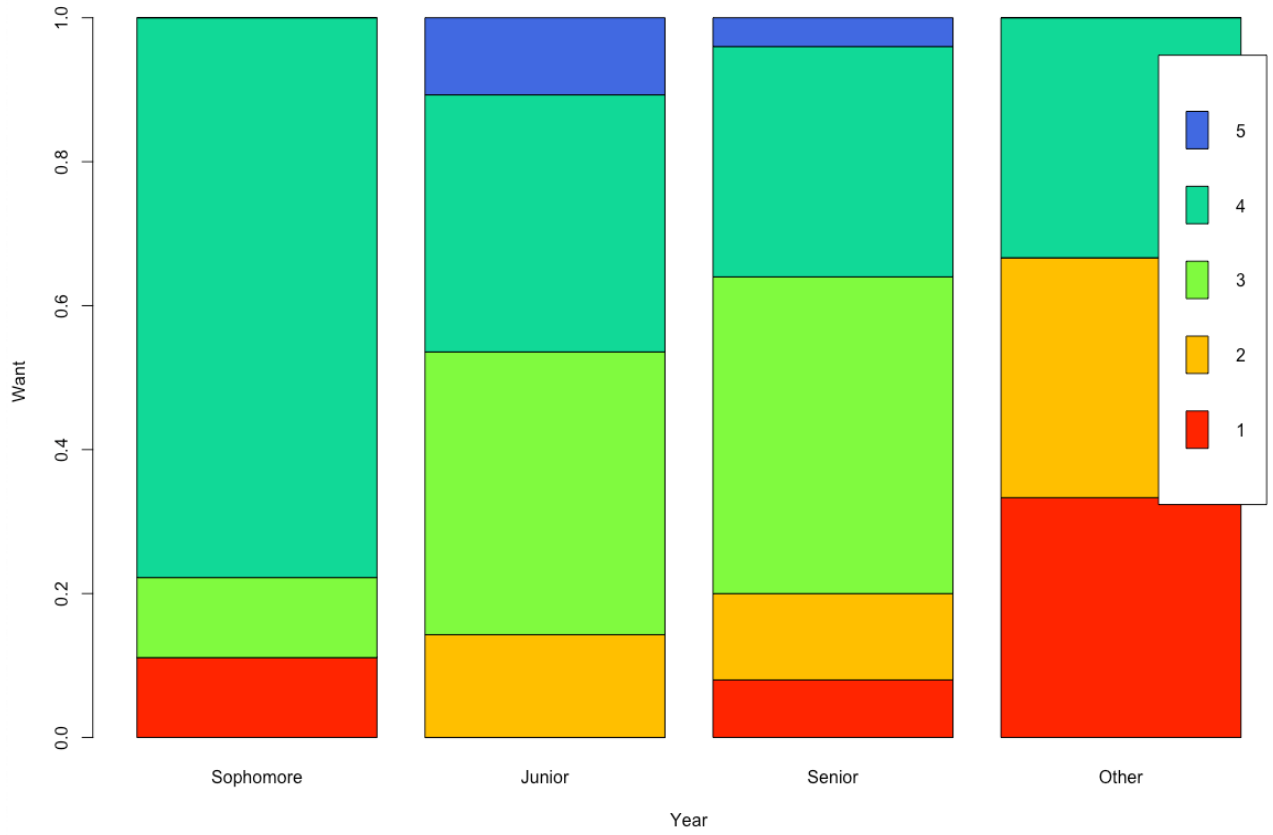
Let's turn our contingency table into a stacked bar plot:



Here, we can see that most of you are juniors and seniors (and that there's a decent spread of how much you want to be here).

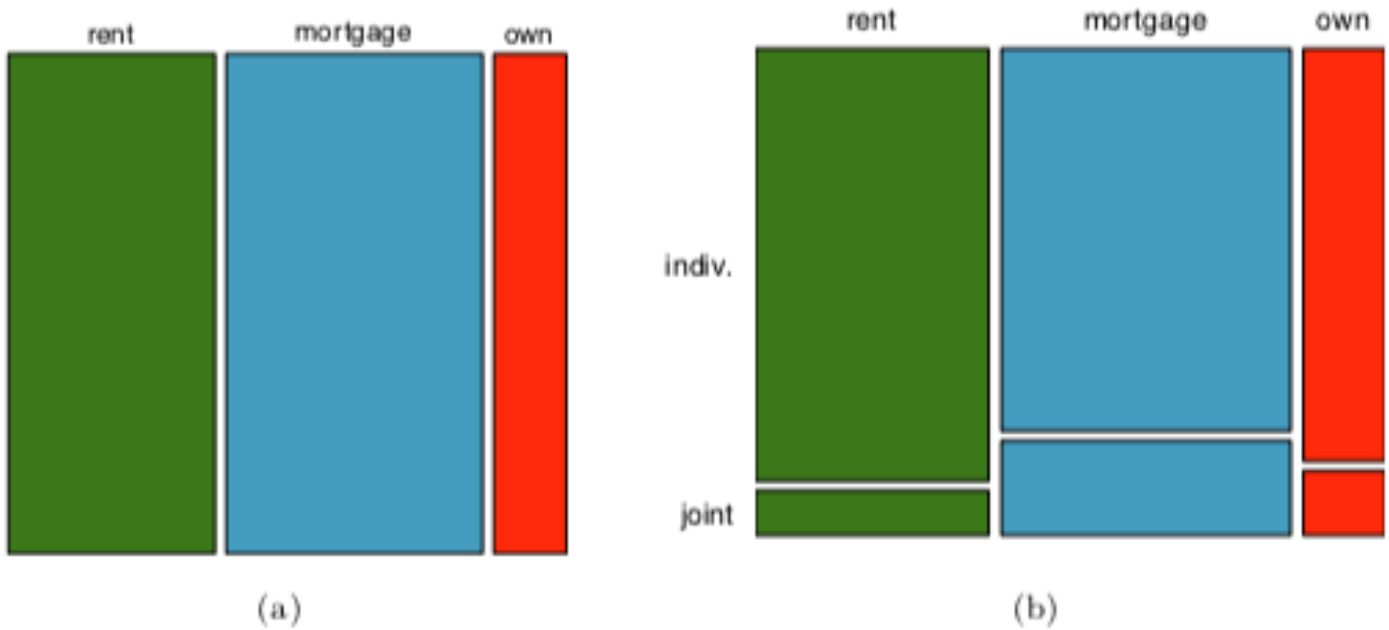
Example: Student Survey Data

Let's do the same with the proportion-based table:



Now we can quickly visualize the differences between the years.

Mosaic Plots

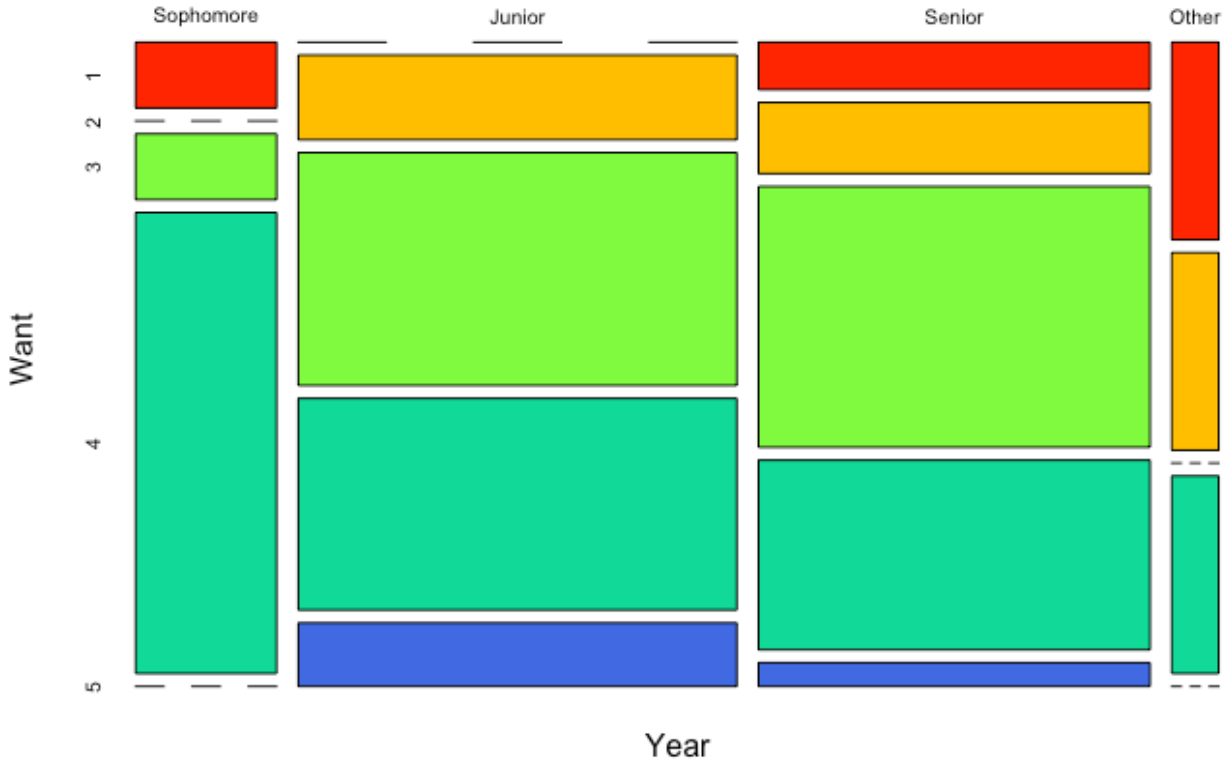


(a) is a one-variable mosaic plot for `homeownership`.
(b) is a two-variable mosaic plot for `homeownership` and `app_type`.

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.

Example: Student Survey Data



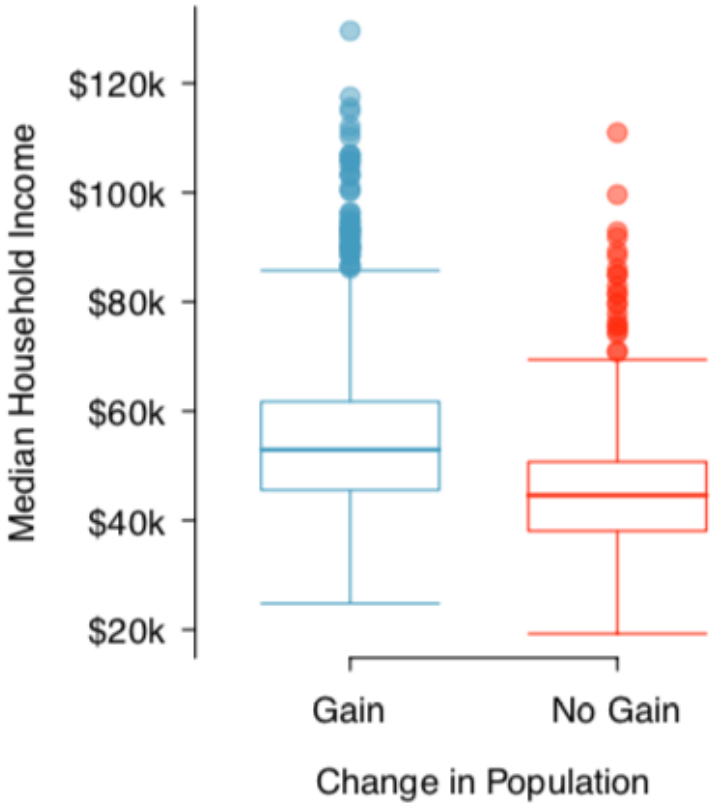
We can again see that there are more juniors & seniors in the class and that sophomores are more likely to want to take this course beyond its being a requirement.

Comparing Numerical Data Across Groups

- Our question of interest often involves comparing numerical data across categories.
- Whenever we are interested in comparing some numeric outcome across treatment groups, this is our goal!
- In general, these comparisons require that we make side-by-side or stacked versions of our data visualization techniques for numerical data.

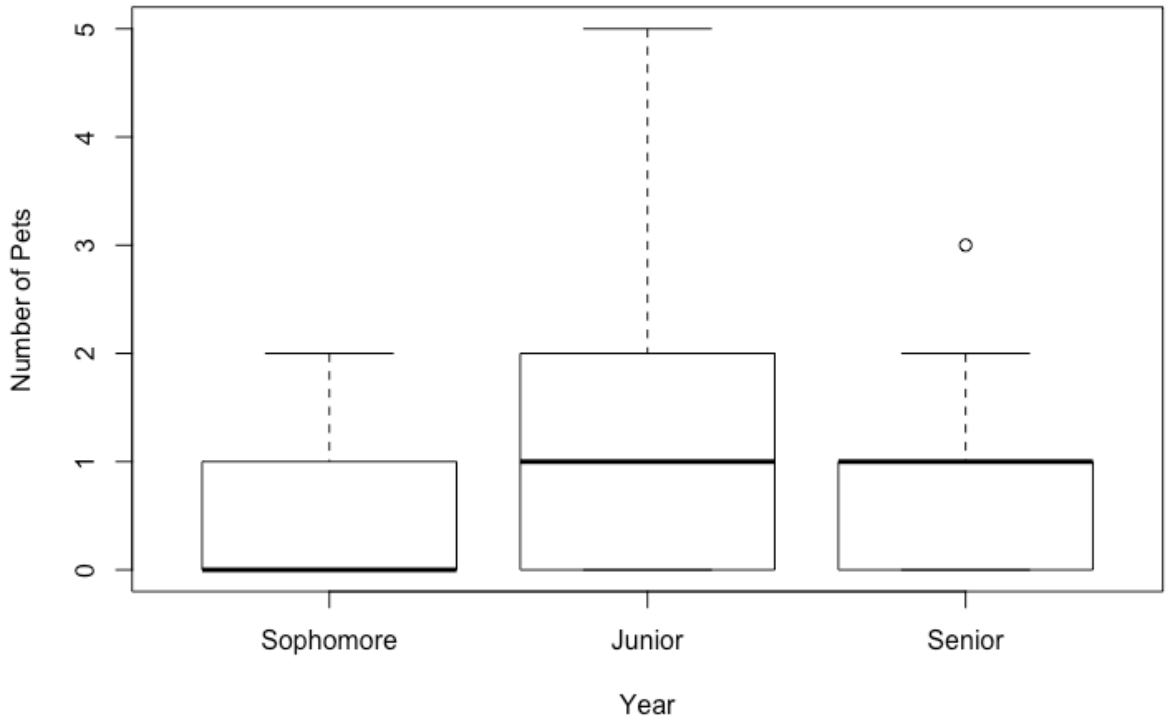
Side-By-Side Box Plots

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



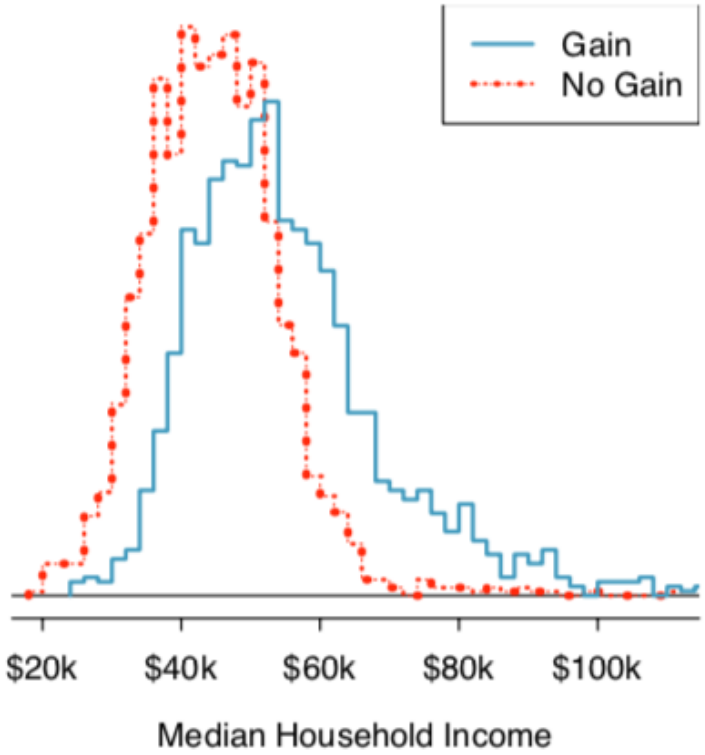
Example: Student Survey Data

Let's look at how number of pets differs between year:



Juniors have a larger IQR and longer whiskers, suggesting that they have a larger spread in number of pets.

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.

Case Study

- Suppose we split the class into two groups by drawing a line down the middle of the classroom.
- Let $\hat{p}_L$ be the proportion of students on the left side who own an Apple product.
- Let $\hat{p}_R$ be the proportion of students on the right side who own an Apple product.

Would you expect these two proportions to be *exactly* the same?

Case Study

- There's no reason to believe that Apple users tend to sit on one side of the room or another*, so we would expect the proportions to be pretty similar.
- But we probably wouldn't expect these numbers to be exactly the same.
- This small expected variation is due to random chance.

* What assumption are we making about how these variables relate to one another?

Case Study: Malaria Vaccine

We consider a study on the malaria vaccine, PfSPZ.

- Volunteer patients randomized into one of two experimental groups.
  - 14 patients received the vaccine.
  - 6 patients recieved a placebo.
- After 19 weeks, all patients are exposed to a (drug-sensitive) strain of malaria.

Case Study: Malaria Vaccine

These are the results:

		outcome		
		Infection	No Infection	Total
treatment	Vaccine	5	9	14
	Placebo	6	0	6
	Total	11	9	20

This suggests infection rates of 35.7% for the treatment group and 100% for the control (placebo) group.

Case Study: Malaria Vaccine

- This study is an experiment, because treatment levels were assigned by the researchers.
- Therefore we can evaluate a causal relationship between the vaccine and incidence of malaria.
- It is not clear what level of blinding was used, but since they used a placebo, it is probably blind.

Strength of Evidence

- We expect there to be some differences in our sample estimates, even if the true values are exactly equal.
- The sample size is small, so it's not clear whether the vaccine would be effective in the population at large.
- It's impossible to know whether the observed difference is due to the vaccine's efficacy or random chance.
- It's possible that such a large difference is normal (due to chance alone) in such a small sample.

Note: In reality, clinical trials suggest that PfSPZ is effective, but storage and transportation costs make it difficult to distribute to areas where malaria is prevalent.

Variability in the Data

This is a good reminder that our observed data may not perfectly reflect the truth!

- This is due to **random noise**, the variability between values due to random chance.
- Random noise and sample size are things we take into account when statistically analyzing scientific claims.

Competing Claims

Whenever we ask a research question, we always have two competing claims, or **hypotheses**. These are labeled H_0 ("H-nought") and H_A ("H-A").

H_0: **Independence model**. The variables treatment and outcome are independent. They have no relationship. Any observed difference between the proportion of patients who developed an infection in the two groups is due to chance.

H_A: **Alternative model**. The variables are not independent. The difference in infection rates is not due to chance. The vaccine affected the rate of infection.

Independence Model

If H_0, the independence model, is true

- The vaccine is irrelevant to infection status.
- The 11 patients who developed an infection would have develop an infection regardless of which group they were assigned to.
- The 9 who did not develop an infection wouldn't have developed an infection regardless of which group they were assigned to.
- The difference in infection rates was due to chance alone.

Alternative Model

If H_A, the alternative model, is true

- Infection rates are influenced by whether or not a person received the vaccine.

Which Model is Correct?

We draw conclusions about which model is more likely to be true by assessing how strong our evidence is

- Do the data conflict with H_0 strongly enough to conclude H_A?
- This depends on
  1. How different the groups are.
  2. How variable the groups are.
  3. How much data we have.

Simulations

We can start to think about the strength of our evidence using simulations.

- Our simulations will assume that our independence model is true.
- We want to know if it is common to see differences as large as the one we saw in our study.
- If it is common, it is more likely that the difference was due to random chance.
- If it is uncommon, it is more likely that the vaccine is helpful in preventing malaria.

Simulations

Simulations sound complicated, but the idea here is just to assume that the vaccine has no effect and then re-randomize the patients to the treatment and control groups.

- If the vaccine has no effect, we assume that the 11 patients who developed an infection would have done so no matter what.
- We also assume that the 9 who did not develop an infection would have no infection no matter what.

Simulations

We can approach this simulation like this:

1. Take 20 note cards to represent the 20 patients
2. Write each infection status on a note card (11 will say "infection"; 9 will say "no infection").
3. Shuffle the note cards and then randomly pull out 14 for the `vaccine` pile. Put the other 6 into the `placebo` pile.
4. Count up how many infections are in each pile.

Simulations

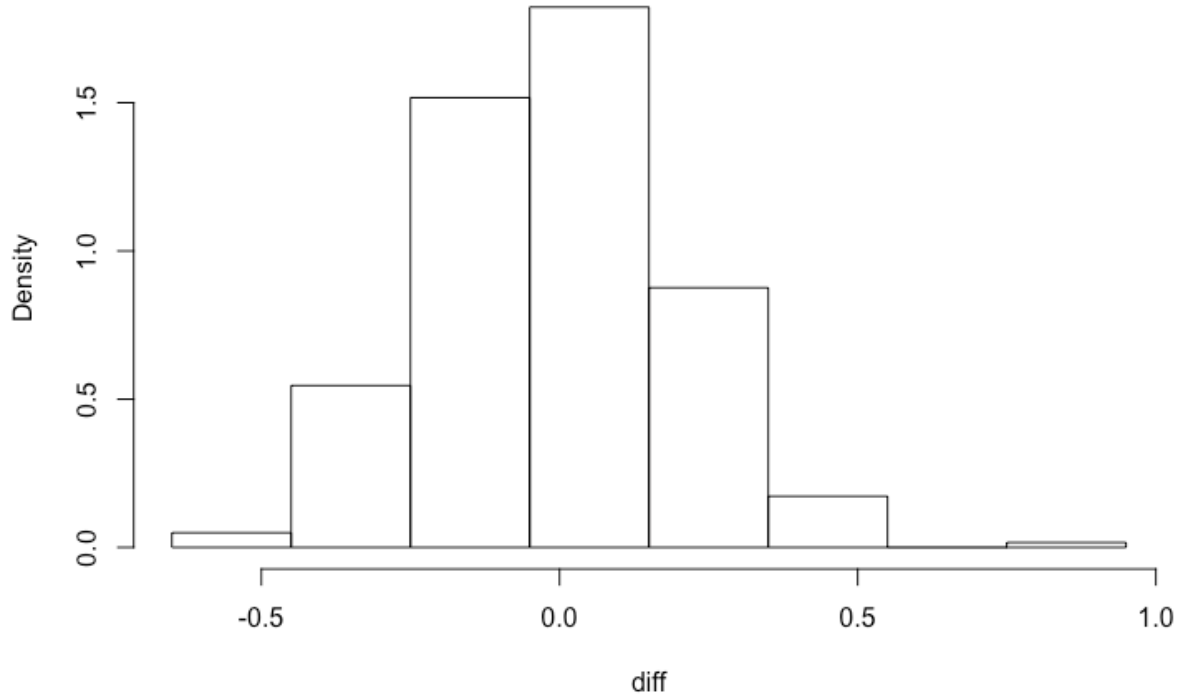
Doing this once, we get

		outcome		
		Infection	No Infection	Total
treatment	Vaccine	7	7	14
	Placebo	4	2	6
	Total	11	9	20

Here, there is an infection rate of 50% for the treatment group and 66.6% in the placebo group, a difference of 16.7%. This is much smaller than in the actual study!

Checking For Independence

The real power of simulations comes from repetition. Using R, I repeated this simulation 10,000 times.



Histogram of the differences across 10,000 repetitions.

Checking For Independence

- In the actual study, the difference in infection rates was 64.3%.
- In my simulations, the average difference was only 0.06%.
- I found a difference as big as the one in the study only 33 times.
  - This means that, if the vaccine is not useful, a difference of 64.3% happens by chance less than 1% of the time!
- This suggests that we have pretty good information despite the small sample size.

Moving Forward

The concepts we've been talking about with our case study are what we want to get at with this course!

- Hypotheses
- Testing claims (testing for independence)
- Figuring out how uncertain we are about our results

Eventually, we will formalize these concepts and talk about how to test our claims without simulations.

A Note About R Code

I've been using a lot of code to write these slides!

- I've added a new page to the course website that will contain links to all of this R code.
- This code will be heavily commented to make it easier to follow and I will set it up so that you will not need to download any additional data.
- As always, learning R is completely optional, but the code is there if you're interested.