Welcome to PH142: PPDAC and Starting to look at Data

Welcome to PH142: PPDAC and Starting to look at Data

Today's Goals

Welcome and orientation to the class - answer questions

My goals for our time together

Talk about the framework we use in the class (PPDAC)

Introduce some data visualizations using R

Who am I?

Who am I?

Figure 2: Pandemic Year

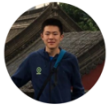
Our Teaching team





Our Fabulous Summer 2021 Teaching team!

https://ph142-ucb.github.io/su21/staff/

Logistics

Lecture/Section/Office Hours/Piazza

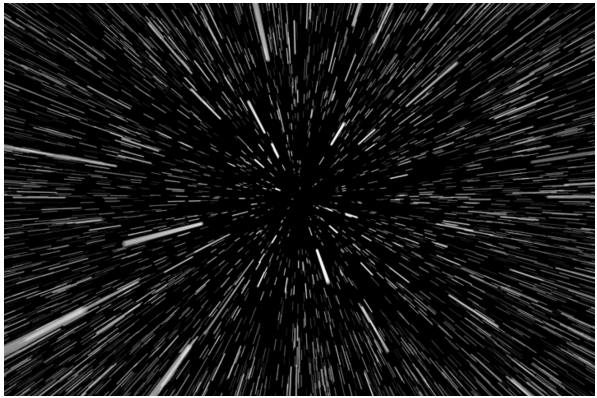
Zoom usage for this class

Rationale for structure

When in doubt - check the website and the piazza announcements

Logistics

Compressed format ....



Frequently asked questions so far

Do I have to attend lecture/section?

Do I need the textbook?

Do I need to know programming?

How to get help with code

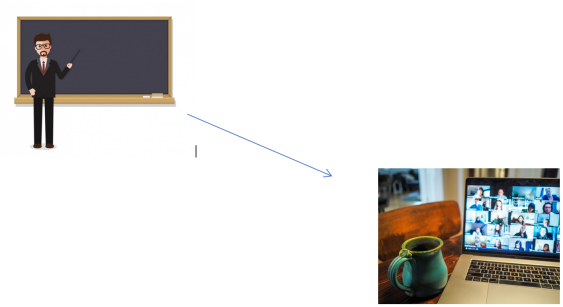
- Ask questions during labs, homework parties, GSI office hours, or on Piazza discussion forum. Use the appropriate thread!
- Develop your online search skills. For example if you have a `ggplot2` question, begin your google search with "r ggplot" and then describe your issues, e.g., "r ggplot how do I make separate lines by a second variable".
- The most common links that will appear are:
  - https://stackoverflow.com: Crowd-sourced answers that have been upvoted. The top answer is often the best one.
  - https://ggplot2.tidyverse.org/: The official ggplot2 webpage is very helpful.
  - https://community.rstudio.com/: The RStudio community page.
  - https://rpubs.com/: Web pages made by R users that often contain helpful tutorials.

Frequently asked questions so far

Figure 3: Will I get an A?

There's an app for that...

Ongoing evolution of the course



From Derivation to hands on programming

Co-Development of course with Dr. Riddell

a pre-emptive appology

(credit to xkcd.com for the comic)

What is this class?

What is this class?

Figure 4: What do you think of when you think about statistics?

My goals for you

Foundational concepts in probability and biostatistics

How to answer questions with data:

- your ability to critically assess statistical information presented to you in scientific and non-scientific fora
- your sense of how to approach answering real world questions with data
- develop your statistical intuition around variability and chance
- develop your toolkit for visualization, summarizing and testing simple relationships
- your ability to concisely and accurately describe statistical methods and results

This is not a math class

Statistics is often classified as a branch of math, but I'd argue that it is more important to focus on the connections that statistics has with science (how we can learn about the world through data)

Though it is true that statistics uses math (and sometimes fairly advanced math!), not much math is needed to learn introductory statistics

In this class we will try, as much as possible, to emphasize concepts and help you develop your statistical intuition

This is not a programming class

Statistics is often viewed as "just computer programming," but this is an incorrect and dangerous characterization: computer programming is simply a tool for conducting statistical analysis

The use of computer programming in statistics is—and should be—quite different than approaches to non-statistical programming

We are using r programming in this course because it is an extremely useful skill, facilitates computation, and is desired in the job market

This is a relevant class

I hope to convince everyone here that statistics is relevant to everyone

As is more and more apparent, public health statistics have relevance to important policy decisions

You also make many decisions during your day that are influenced by statistics

Statistics is not just relevant for public health, but also for other professions, including: education, journalism and law

As we'll try to illustrate via the recurring "statistics is everywhere" segments, statistics is useful for understanding the news and the world around us

Statistics is Everywhere

Advice about how to be happier

How to have a better day during the pandemic

June 30, 2020

(photo courtesy of Pexels)

Passively browsing social media is not good for you — and other useful findings on resilience and happiness from the Positive Emotions and Psychophysiology Lab.

Website link

Advice about how to be happier

In their web posting they say:

We'd like to join those calling for a terminological change. What is needed is not social distancing but physical distancing and social solidarity. During these "challenging times", it's even more important than usual that people stay connected and help each other. So to have a better day during the pandemic, it's vital that everyone MARCH together:

- Minimize passive scrolling through social media.
- Accept negative emotion.
- Really connect with people.
- Care for yourself.
- Help others."

Advice about how to be happier

How did they get to these conclusions?

What should we ask about their methods?

Should we be convinced to change our behavior based on these data?

Advice about how to be happier

Methods:

At the Positive Emotions and Psychophysiology Lab at UNC Chapel Hill, our team recently collected data from over 600 adults around the United States, asking about their experiences and behaviors from the past day.

Problems with scientific reporting

Consequences of poor communication

PPDAC - the approach we will use to answering questions with statistics

Problem

A clear statement of what we are trying to achieve.

Three main problem types

- Descriptive: learning about some particular attribute of a population
- Causative/Etiologic: do changes in an explanatory variable cause changes in a response variable?
- Predictive: how can we best predict the value of the response variable for an individual?

Problem type?

- Insurance company: What is the probability (how likely is it) that a 25 year old unmarried male driver has a car accident?
- Health department: How many cases of influenza have we seen this season compared to last season?
- Health care system: If we treat patients with diabetes using medication X, will thier insulin regulation be better or worse than medication y?

Plan

The procedures we use to carry out the study.

- Census or sample from the target population?
  - How was the sampling conducted?
  - Was the sample random?
- Is the study prospective or retrospective?
- Is the study observational or experimental?

Data

The data which is collected according to the Plan.

- How many observations to we have?
- How relaible are the measures?

Analysis

The data is summarized and analysed to answer the questions posed by the Problem.

We use our knowledge about probabilities to assess the role of chance in our findings.

Conclusion

Conclusions are drawn about what has been learned about answering the Problem.

PPDAC Example 1: A smoking behaviour study

PPDAC Example

Problem: Suppose we wish to study the smoking behavior of California residents aged 14-20 years.

In particular, we are interested in the *prevalence* of current smoking by gender.

What type of problem is this?

PPDAC Example

Plan: We need to first choose a time period, because we know that smoking behavior has changed immensely over time. It is unfeasible to gather these data for all residents in California who are 14-20 years old.

Instead we conduct a *random sample* of size n persons. We collect their: age, gender, and smoking status.

Note that we need to decide how large n should be, and how to obtain the random sample. The latter question is, in particular, very important if we want to ensure that our sample is representative of the population of interest. Time and money also constrain how the sample will be collected.

PPDAC Example

Data: Suppose that a random sample of 200 persons aged 14-20 was selected, yielding these data:

Gender	Number of smokers	Number of non-smokers	Total
Teen girls and women	32	66	98
Teen boys and men	27	75	102
Total	59	141	200

PPDAC Example

Analysis: The proportion of women in the sample who smoke is $32/98 = 33\%$. The proportion of men in the sample who smoke is $27/102 = 26\%$.

We would also like some idea as to how close this estimate is likely to be from the actual proportion in the population.

If we selected a second random sample of the same size, we would likely estimate different proportions for men and women. We will learn how to estimate the precision of these estimates.

PPDAC Example

Conclusion: 33% of girls and women aged 14-20 and 26% of boys and men of the same age group are current smokers in California in 2018 (plus a measure of uncertainty).

Example 2: Life expectancy for non-Hispanic black and white men and women in California between 1969-2013

Introduction

Life expectancy is one of the core measures used in public health to comment on the well-being of groups of people. Differences in life expectancy by race/ethnicity, for individuals living in the same region can reflect underlying inequalities in policies, access to care, food environments, structural and systemic racism, among other potential causes.

Research objective (Problem)

The purpose of this short report is to visualize life expectancy among black and white men and women in California between 1969 and 2013.

We are interested in whether there are differences by group and whether these differences have changed over time.

What type of problem is this?

Plan

Death certificates in the United States include race/ethnicity, age at death, and date of death and capture all deaths of US residents. These data are aggregated by the CDC's National Cancer Institute into the SEER*Stat software. Previously, Riddell et al.[1], analyzed these data to compute estimated trends in life expectancy for non-Hispanic black and white men and women, for 40 US states between 1969 and 2013. States without enough data were excluded from these analyses.

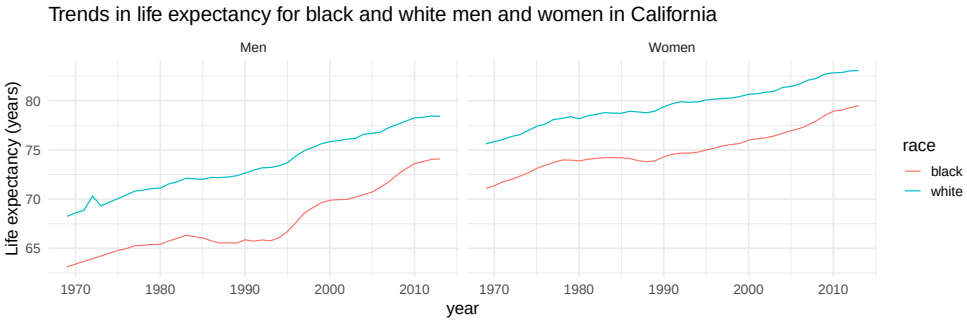
To carry out this short report, we will use data from Riddell et al. to visualize trends in life expectancy as part of an exploratory data analysis. In particular, we will plot time trends for black and white men and women in California.

Data

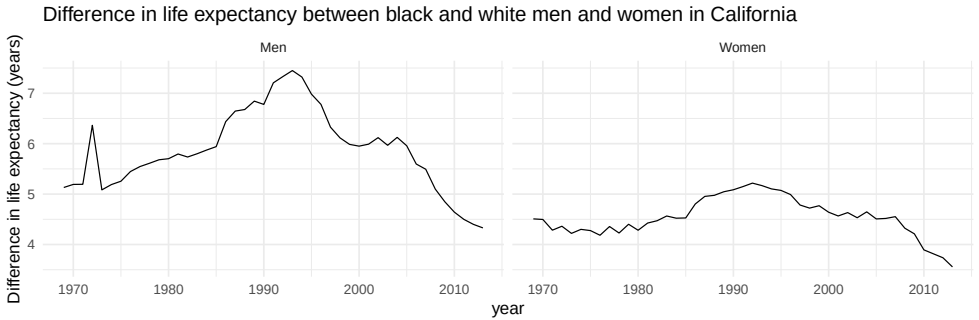
Here are the first few rows of these data for California:

state	stabbrs	year	sex	Census_Region	Census_Division	LE	race
California	CA	1969	Female	West	Pacific	75.61137	white
California	CA	1969	Male	West	Pacific	68.24766	white
California	CA	1970	Female	West	Pacific	75.84916	white
California	CA	1970	Male	West	Pacific	68.59865	white
California	CA	1971	Female	West	Pacific	76.05663	white

Analysis



Analysis



Conclusion

The difference in life expectancy in 1969 between non-Hispanic blacks and whites was 5.1 years for men and 4.5 for women in California.

By 2013, the difference was 4.3 years for men and 3.6 for women in California.

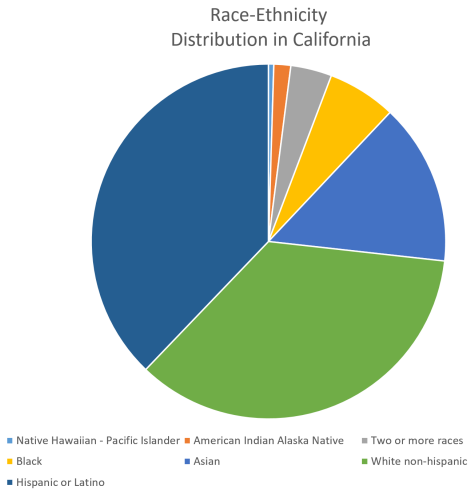
Visualizations for categorical data

Visualization of categorical data

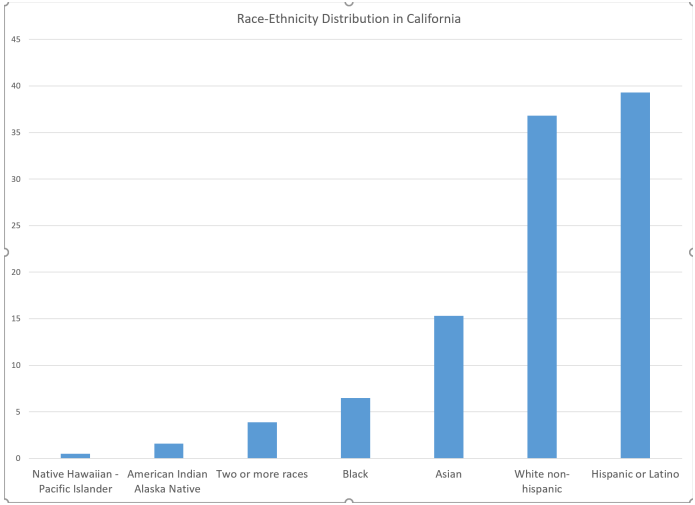
- What is the best way to visualize one categorical variable at a time?
- Generally speaking, it is not a good idea to use pie charts

Visualziation of categorical data

Can you judge the area of the slices?



Visualziation of categorical data



Visualization of categorical data

- We prefer bar graphs (also called bar charts) for the display of categorical data.
- Bar charts display the number or percent of data for each level of the categorical variable being plotted

Example: infectious disease data

- Task: Make a bar chart of the percent of cases on infectious disease for each category of disease.
- First, read and view the infectious disease data from Baldi and Moore:

```
id_data <- read_csv("Ch01_ID-data.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   disease = col_character(),
##   type = col_character(),
##   number_cases = col_double(),
##   percent_cases = col_double()
## )
```

Example: infectious disease data

```
id_data
```

```
## # A tibble: 7 x 4
##   disease            type       number_cases percent_cases
##   <chr>              <chr>             <dbl>         <dbl>
## 1 Chlamydia          STI              174557          66.4
## 2 Gonorrhea          STI               44974          17.1
## 3 Pertussis          Pertussis         11219           4.27
## 4 Campylobacteriosis Foodborne          7919           3.01
## 5 Early syphilis     STI                7191           2.74
## 6 Salmonellosis      Foodborne          5361           2.04
## 7 Other              Other             11559           4.40
```

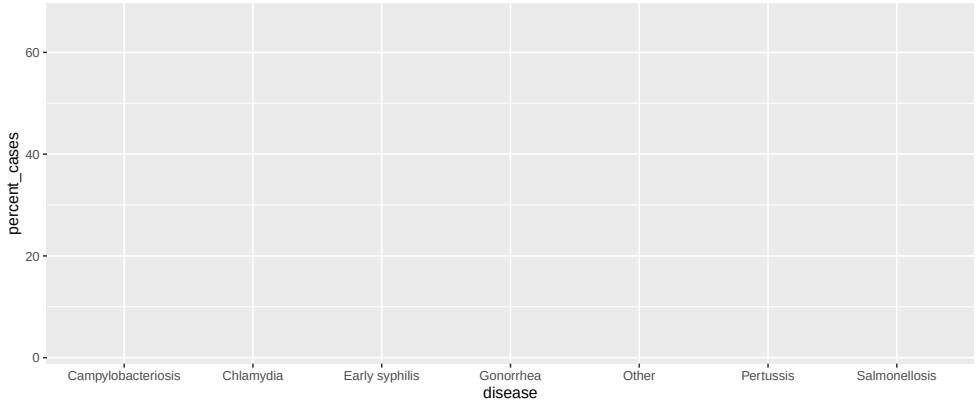
Example: infectious disease data

- Note the variables `number_cases` and `percent_cases`
- What do you want the bar chart to display? What is the x and y variables for a bar chart?

Introducing ggplot

First step to building a `ggplot()`: set up the canvas

- The first line of code below pulls in the ggplot package
- The second line of code below specifies the `data` set and what goes on the `x` and `y` axes



Next choose a function

- We will use a `geom_` function to create our chart

`ggplot()`'s `geom_bar()` makes a bar chart

Syntax for bar charts

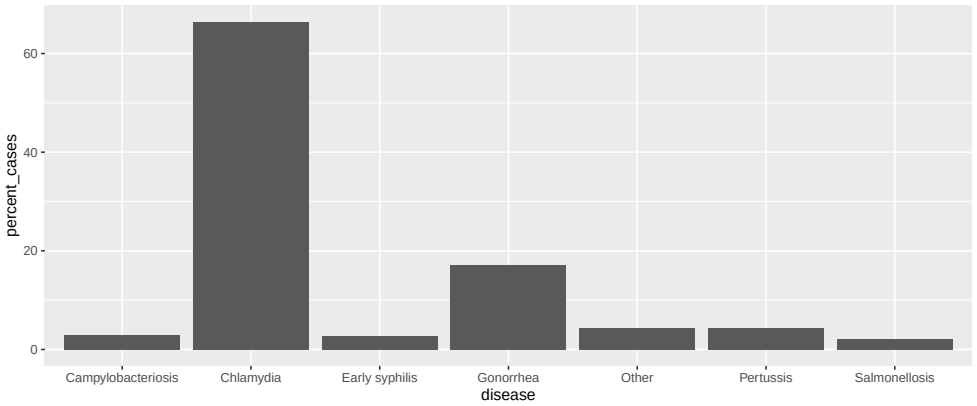
ggplot(id_data, aes(x = disease, y = percent_cases)) +

geom_bar(stat = "identity")

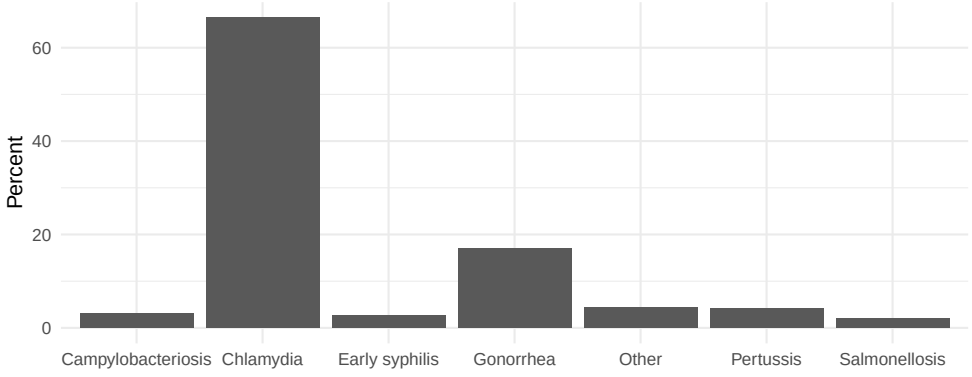
stat = "identity" tells geom_bar that we supplied a y variable that is exactly what we want to plot.

We do not need geom_bar() to calculate the number or percent for us.

ggplot()'s geom_bar() makes a bar chart



some additions to ggplot for style



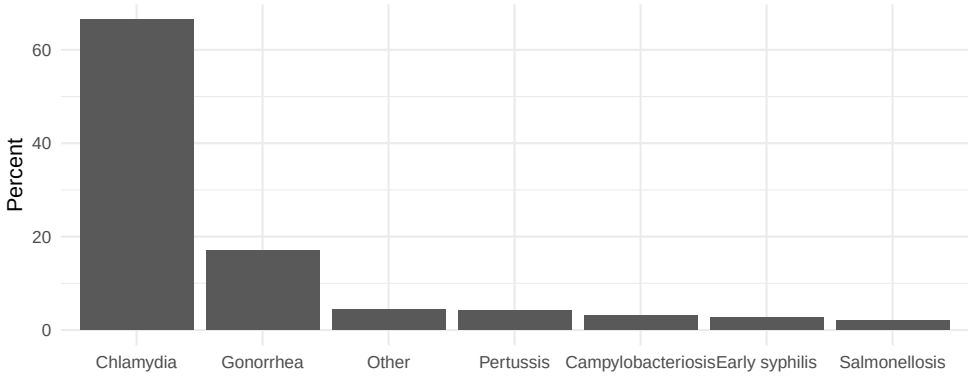
base_size controls the font size on these plots

theme_minimal affects the "look" of the plot it removes the grey background and adds grey gridlines

fct_reorder reorders disease according to value of percent_cases

```
id_data <- id_data %>%
  mutate(disease_ordered = fct_reorder(disease, percent_cases, .desc = T))
```

Re-ordered plot

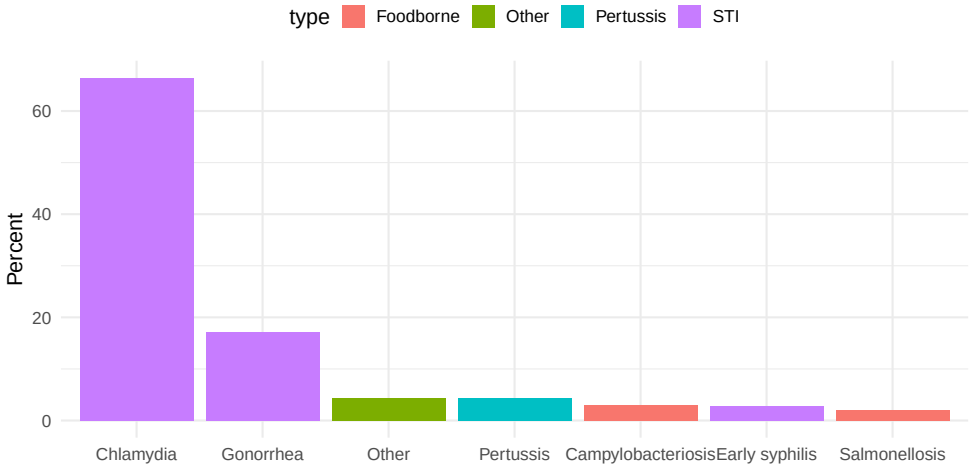


Use `aes(fill = type)` to link the bar's fill to the disease type

geom_bar(stat = "identity", aes(fill = type)) +

theme(legend.position = "top")

Use `aes(fill = type)` to link the bar's fill to the disease type



Visualizing quantitative variables

Visualize quantitative variables using histograms

- Histograms look a lot like bar charts, except that the bars touch because the underlying scale is continuous and the order of the bars matters
- In order to make a histogram, the underlying data needs to be binned into categories and the number or percent of data in each category becomes the height of each bar.
- the bins devide the entire range of data into a series of intervals and counts the number of observations in each interval
- the intervals must be consecutive and non-overlapping and are almost always chosen to be of equal size

Example: opioid state prescription rates

- The textbook gives an example using data from 2012.
- In the data folder, there is updated data from 2018. It came from the paper: "Opioid Prescribing Rates by Congressional Districts, United States, 2016", by Rolheiser et al. link

Example: opioid state prescription rates

Problem: To determine the extent to which opioid prescribing rates vary across US congressional districts.

Plan: In an observational cross-sectional framework using secondary data, they constructed 2016 congressional district–level opioid prescribing rate estimates using a population-weighted methodology.

Data: In the data structure we have State as the unit of analysis, and measured perscription rates as the variable of interest

Example: opioid state prescription rates

```
opi_data <- read.csv("Ch01_opioid-data.csv")
head(opi_data)
```

```
##   Rank State   Mean Median    SD    Min    Max Num_Districts
## 1    1    AL 121.31 113.09 21.87 105.58 166.69             7
## 2    2    AR 115.22 115.13  8.59 104.80 125.79             4
## 3    3    TN 108.12 108.26 19.16  73.60 133.00             9
## 4    4    MS 105.64 106.25 17.36  83.90 126.14             4
## 5    5    LA  98.38  98.88 10.34  83.22 112.65             6
## 6    6    KY  98.13  85.76 26.72  77.62 147.00             6
```

- Mean provides the mean prescribing rate per 100 individuals. Thus, a mean of 121.31 implies that in Alabama, there were 121.31 opioid prescriptions per 100 persons, an average across the 7 congressional districts.

Histogram of opioid prescription rates

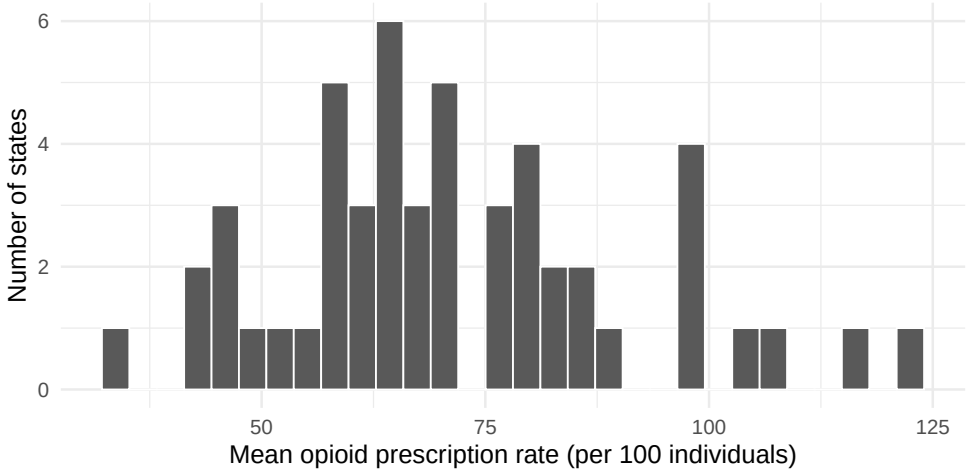
- Task: Make a histogram of the average prescribing rates across US states
- What is the x variable? What is the y variable?
- What `geom` should be used?

Histogram of opioid prescription rates - default is 30 bins

```
ggplot(data = opi_data, aes(x = Mean)) +
geom_histogram(col = "white") +
labs(x = "Mean opioid prescription rate (per 100 individuals)",
    y = "Number of states") +
theme_minimal(base_size = 15)
```

Histogram of opioid prescription rates

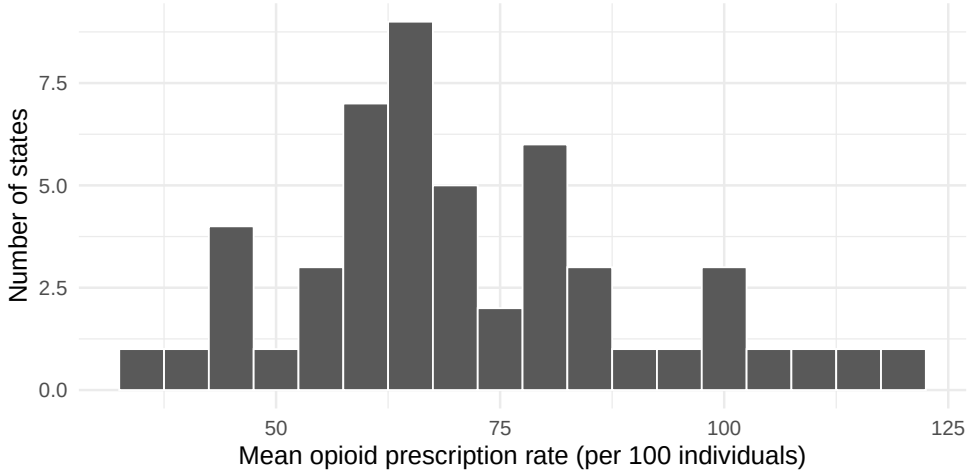
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



same graph, change the bins `geom_histogram(binwidth = 5)`

```
ggplot(data = opi_data, aes(x = Mean)) +
geom_histogram(col = "white", binwidth = 5) +
labs(x = "Mean opioid prescription rate (per 100 individuals)",
    y = "Number of states") +
theme_minimal(base_size = 15)
```

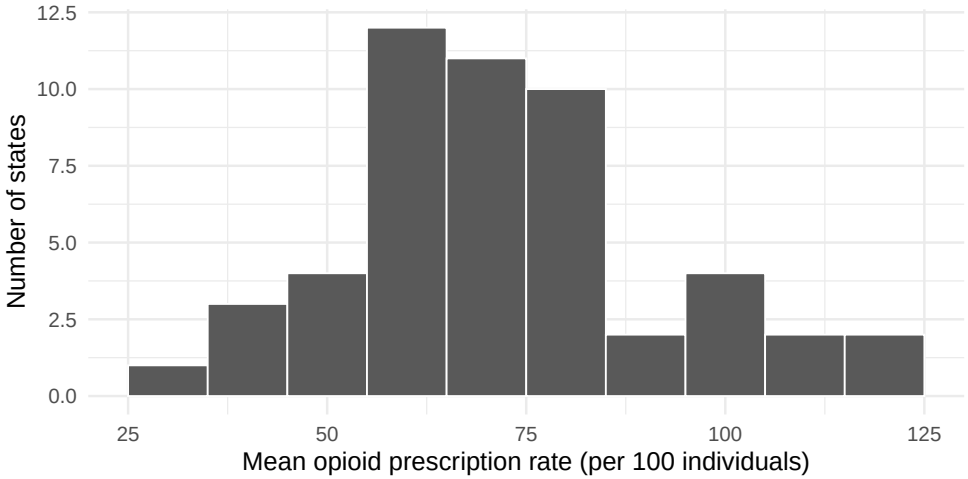
same graph, change the bins `geom_histogram(binwidth = 5)`



change the bins again geom_histogram(binwidth = 10)

```
ggplot(data = opi_data, aes(x = Mean)) +
geom_histogram(col = "white", binwidth = 10) +
labs(x = "Mean opioid prescription rate (per 100 individuals)",
    y = "Number of states") +
theme_minimal(base_size = 15)
```

change the bins again `geom_histogram(binwidth = 10)`

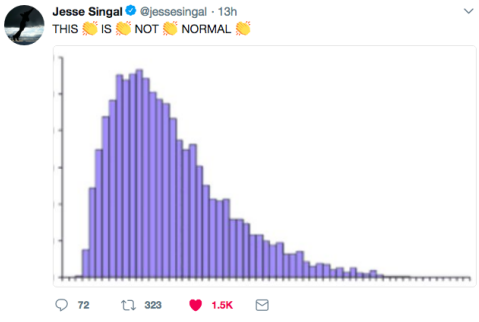


Describing your distribution - what are we looking for?

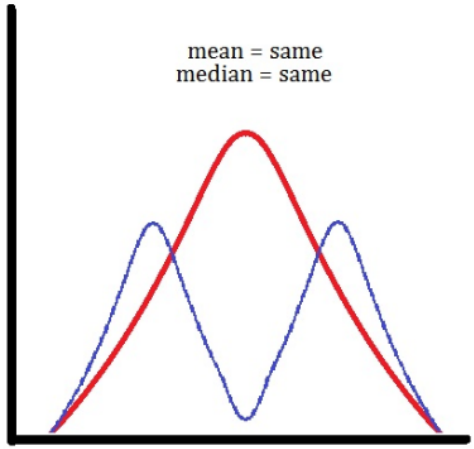
Shape, Center, Spread

- When we examine histograms, we can make comments on a distribution's:
  - Shape: Is the distribution symmetric or skewed to the left or right?
  - Center: Does the histogram have one peak (unimodal), or two (bimodal) or more?
  - Spread: How spread out are the values? What is the range of the data?
  - Outliers: Do any of the measurements fall outside of the range of most of the data points?

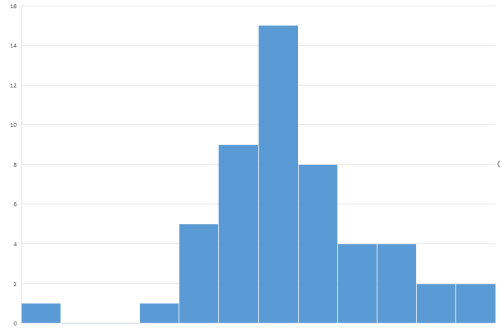
Is this skewed left or skewed right?



Center - one hump or two?



Outlier



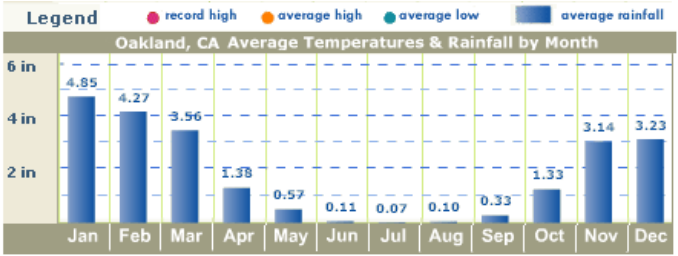
Time plots

Visualize quantitative variables over time using time plots

- Time plots are a specific subset of line plots where the x variable is time.
- Unlike the previous plots, the time plot shows a relationship between two variables:
  i) a quantitative variable
  ii) time
- Often times, these plots can be used to look for cycles (e.g., seasonal patterns that recur each year) or trends (e.g., overall increases or decreases seen over time).

Time plot

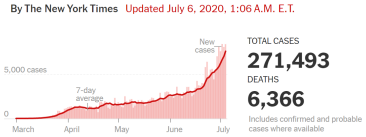
- from *See California.com*, January 2019:



Time plot

- from *'r link("NYtimes.com','https://www.nytimes.com/interactive/2020/us/california-coronavirus-cases.html')'*

California Coronavirus Map and Case Count



Life expectancy for White men in California

Make a scatter plot of the life expectancy for White men in California over time.

Since the dataset contains 39 states across two genders and two races, first use a function to subset the data to contain only White men in California.

Which function from last lecture do we need?

- `mutate()`, `select()`, `filter()`, `rename()`, or `arrange()`?

dplyr's filter() to select a subset of rows

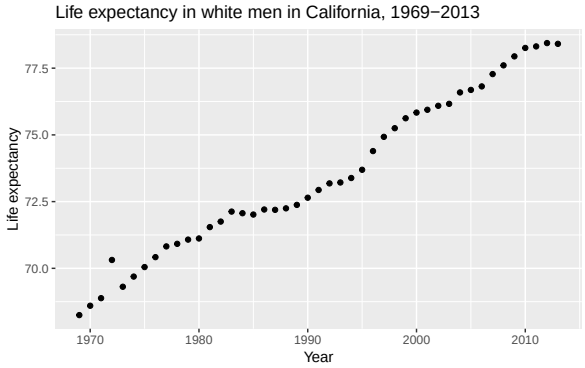
```
wm_cali <- le_data %>% filter(state == "California",
                              sex == "Male",
                              race == "white")

#this is equivalent:
wm_cali <- le_data %>% filter(state == "California" & sex == "Male" & race ==
```

Here we use geom_point to make a graph with dots

```
ggplot(data = wm_cali, aes(x = year, y = LE)) +
geom_point() +
labs(title = "Life expectancy in white men in California, 1969-2013",
    y = "Life expectancy",
    x = "Year",
    caption = "Data from Riddell et al. (2018)")
```

Here we use `geom_point` to make a graph with dots



geom_line() to make a line plot

ggplot(data = wm_cali, aes(x = year, y = LE)) +

geom_line(col = "blue") +

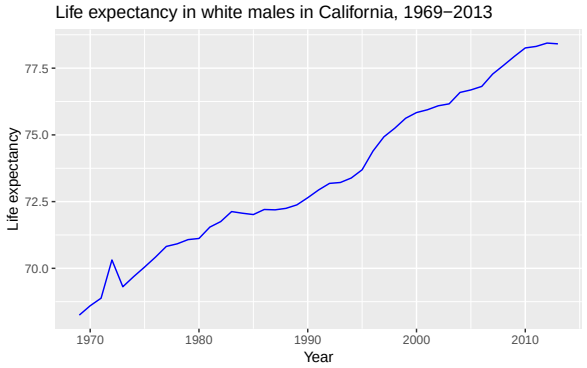
labs(title = "Life expectancy in white males in California, 1969-2013",

```
    y = "Life expectancy",

    x = "Year",

    caption = "Data from Riddell et al. (2018)")
```

geom_line() to make a line plot



R Recap: Code we used today

1. 'ggplot' to set up a canvas for graphics
2. `geom_bar(stat = "identity")` to make a bar chart when you specify the y variable
3. `geom_histogram()` to make a histogram for which ggplot needs to calculate the count
4. `fct_reorder(var1, var2)` to reorder a categorical variable (`var1`) by a numeric variable (`var2`)
   - from the `forcats` package
5. `geom_point()` to make a plot with dots
6. `geom_line()` to make a plot with lines

We only skimmed the surface!

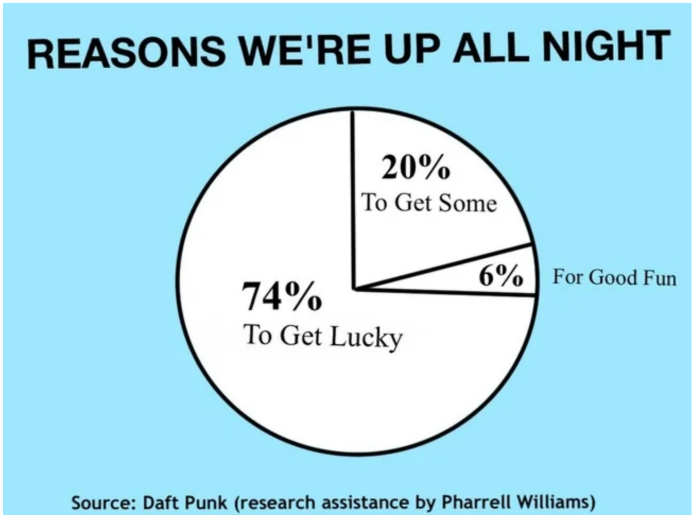
- Here is some extra material for those of you who love data visualization. This material won't be tested.
  - RStudio ggplot2 cheatsheet
  - Kieran Healy's data visualization book

References

The PPDAC method is described based on course notes from STAT 231 from the University of Waterloo (Ontario, Canada). Spring 2006 Course Packet.

1. Riddell CA, Morrison KT, Harper S, Kaufman JS. Trends in the contribution of major causes of death to the black-white life expectancy gap by US state. Health & Place. 2018. 52:85-100. doi: 10.1016/j.healthplace.2018.04.003.

Parting Humor



- from *Eric Tanoye Song Lyrics in Chart Form*