Quantitative data and measures of central tendency

Overview

Review of analyzing categorical data

- Concepts and R

statistics for a quantitative variable

- Graphing the shape: dot plots, histograms and outliers
- Measures of the central tendency: mean and median
- Resistance
- Try it in R yourself

Announcement: if you’re on the waitlist you’re in

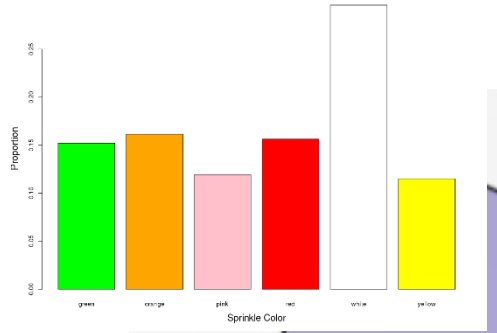
If you are still on the waitlist for this course and want to get in, please fill out an add form and have me sign it

How did worksheet 1 go?

Questions?

Good work!

What is a categorical distribution?

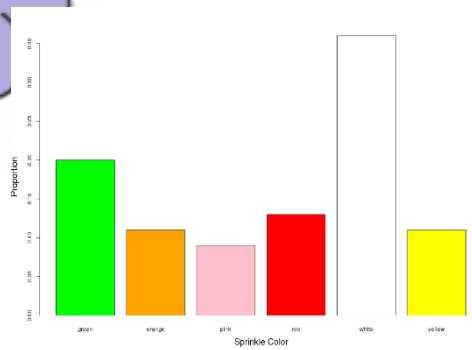


π_{red}

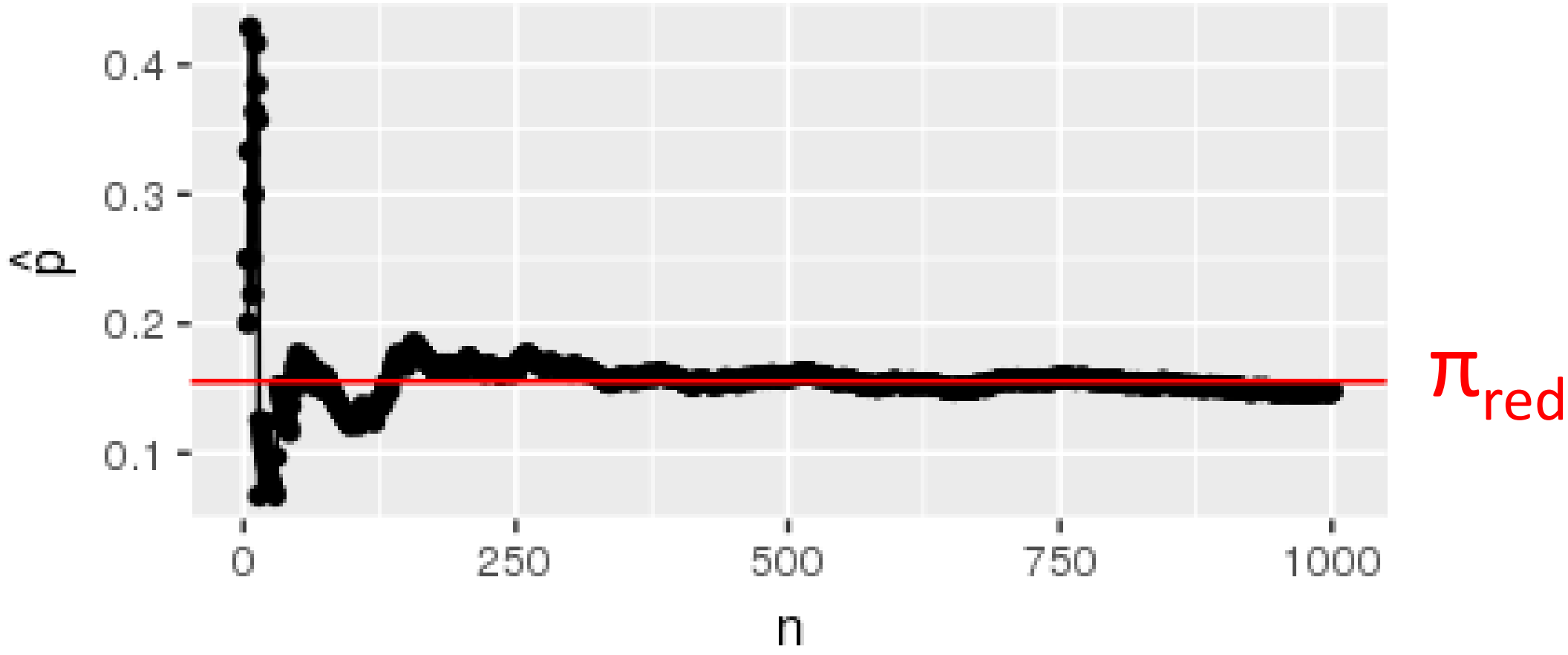
Categorical distribution

Bar chart

$\hat{p}_{red}$



Consistency



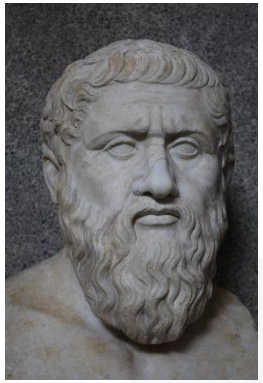
Review

Categorical variables

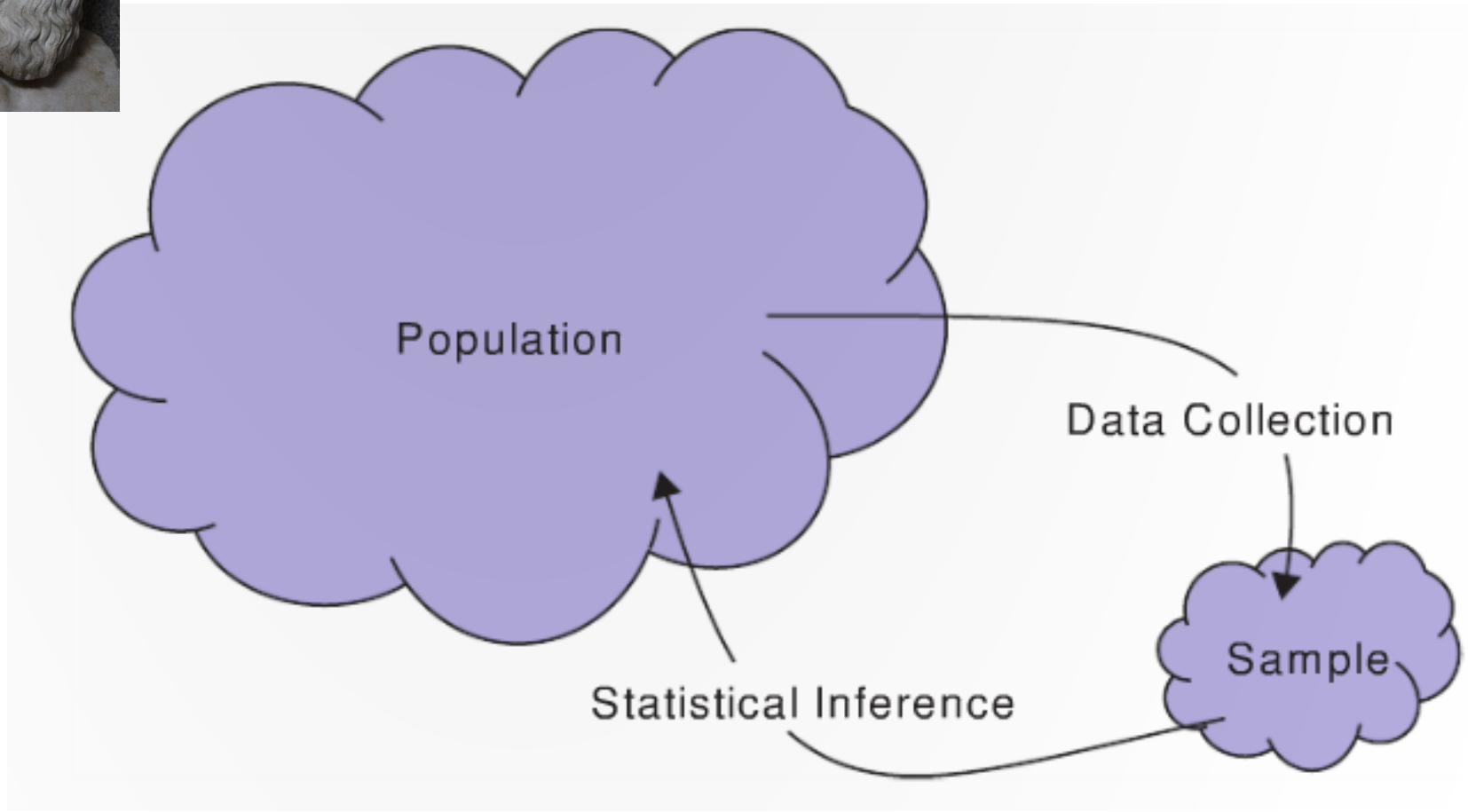
Quiz: Art time!

Please draw:

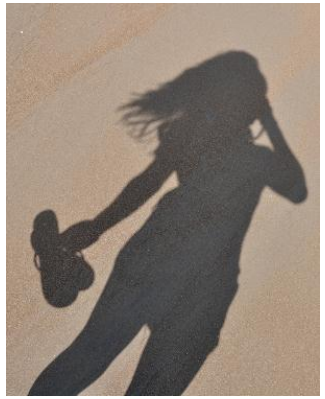
1. A population – and label it a “population”
2. A sample – and label it “sample”
3. Add the label “parameter” in the appropriate location
4. Add the label “statistic” in the appropriate location
5. Add the symbol for a population proportion in the appropriate location
6. Add the symbol for a sample statistic for proportion in the appropriate location
7. Add Plato in the appropriate location
8. Add the shadows in the appropriate location



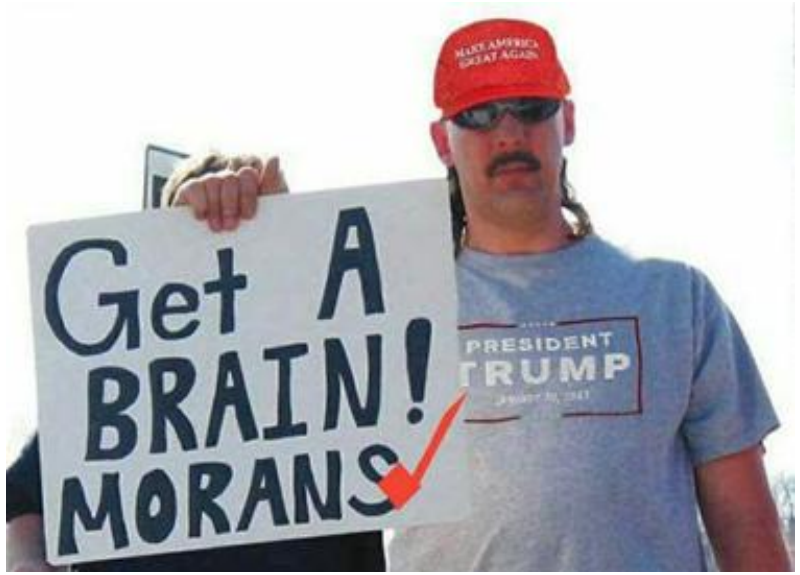
parameter: π



statistic: $\hat{p}$



Example of categorical data: Trump approval ratings



Approve

Disapprove

Example: Trump approval rating

```
# get Trump's approval rating from 1,000 simulated voters
> source("/home/shared/intro_stats/get_approval_sample.R")
> approval_sample <- get_approval_sample(1000)
```

1	approve
2	disapprove
3	disapprove
4	disapprove
5	disapprove
6	approve
7	disapprove

Questions:

1) What are the observational units (cases)?

2) What is the variable?

3) What is the population?

Example: Trump approval rating

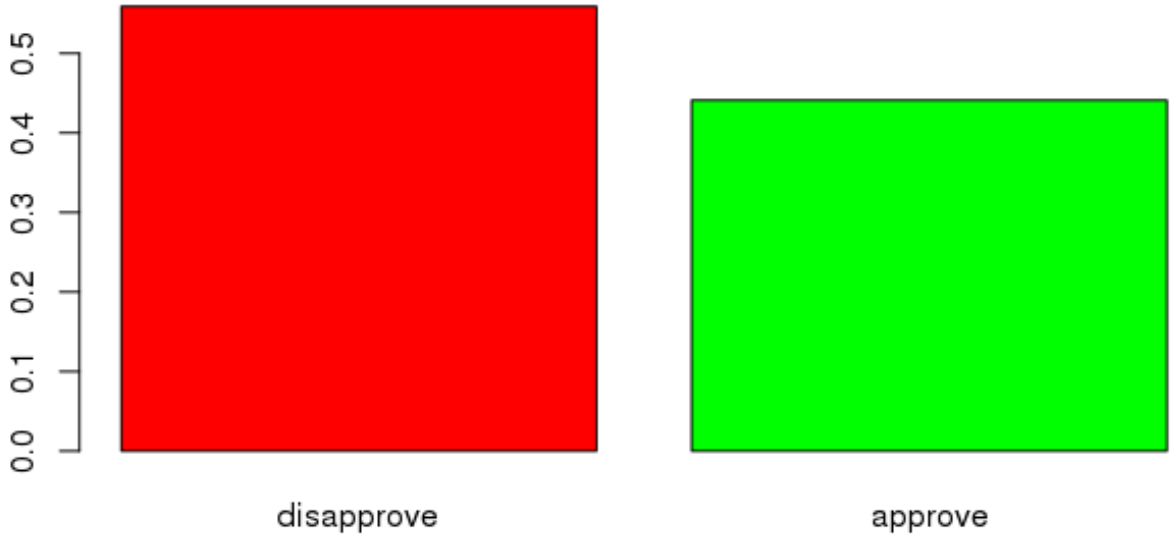
Can you calculate $\hat{p}$ for Trump's approval?

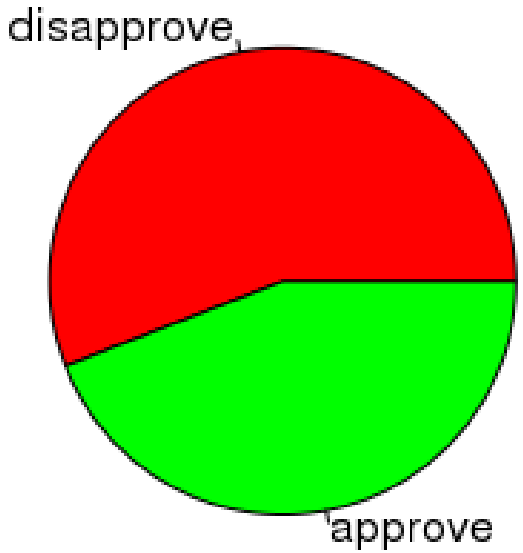
```
> approval_table <- table(approval_sample)
> approval_proportions <- prop.table(approval_table)
> approval_proportions["approve"]
```

Can you make a bar plot and pie chart for his approval proportion?

```
> barplot(approval_proportions)
> pie(approval_proportions)
```

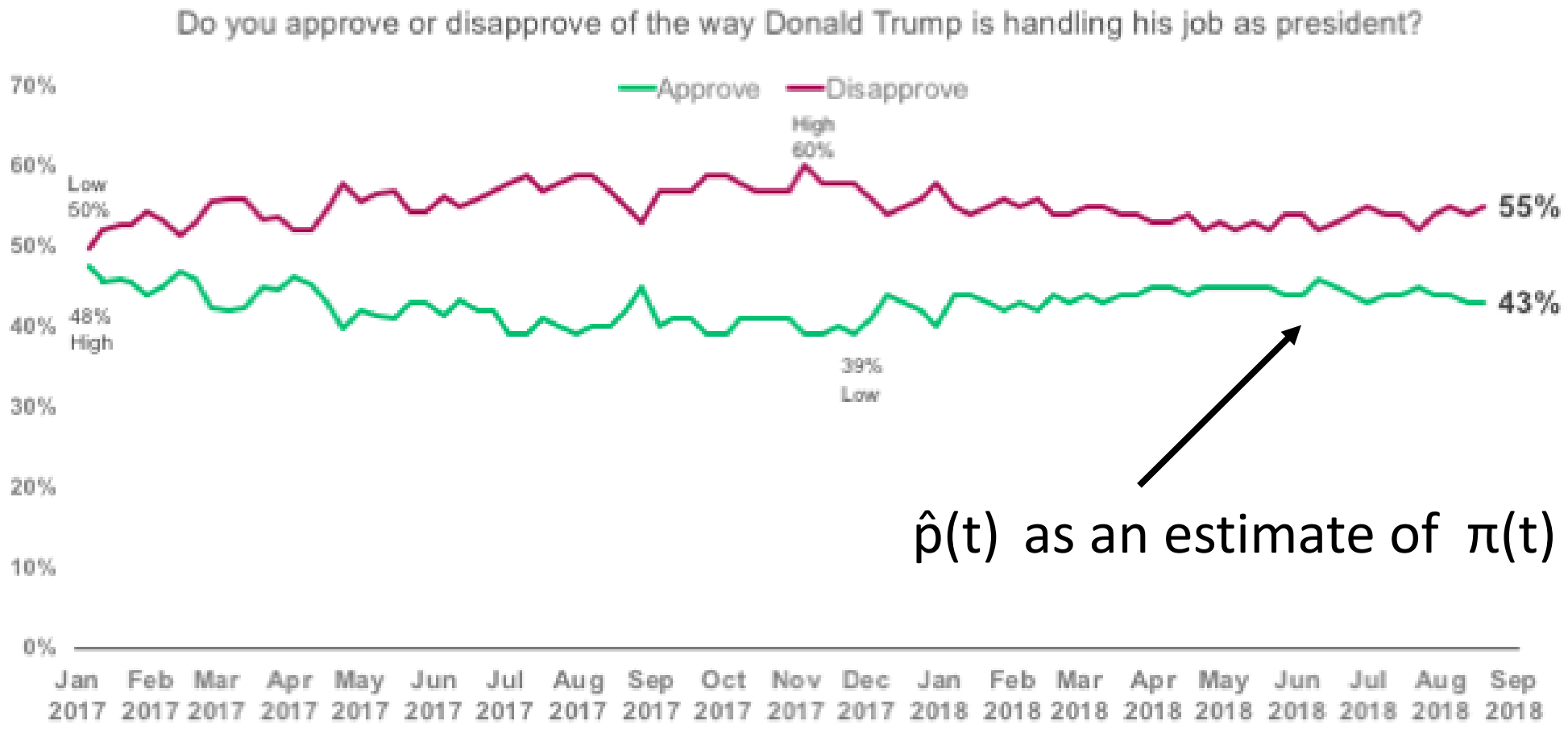
Example: Trump approval rating





Is this $\pi_{approve}$ or $\hat{p}_{approve}$?

Trump Job Approval



$\hat{p}(t)$ as an estimate of $\pi(t)$

SurveyMonkey

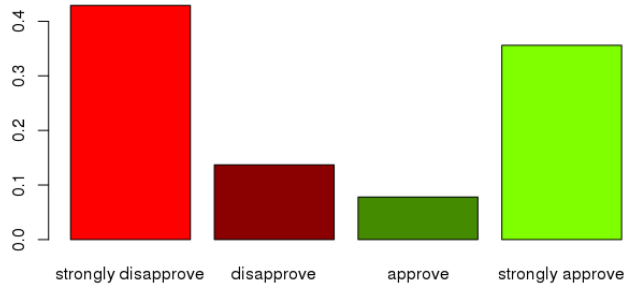
Practice at home

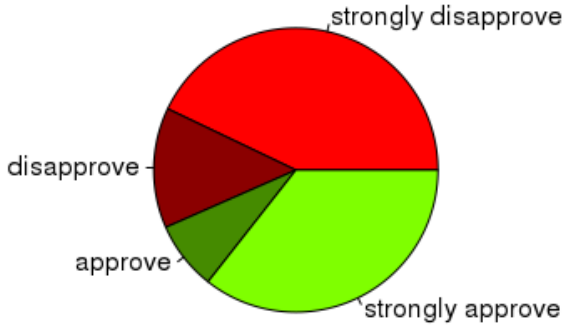
Get the degree to which likely voters approve of trump:

```
> approval_sample <- get_approval_sample(1000, degree_of_approval = TRUE)
```

Practice at home:

- Calculate a relative frequency table for the degree of trump's approval
- Make a bar plot and pie chart of this data





Quantitative variables

Descriptive statistics for one quantitative variable

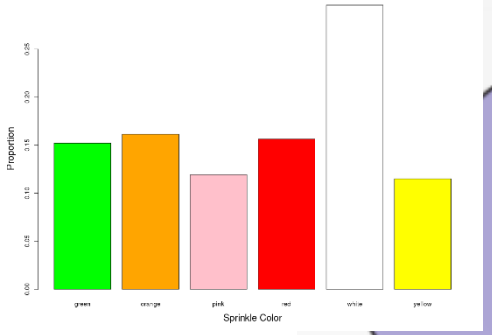
We will be looking at:

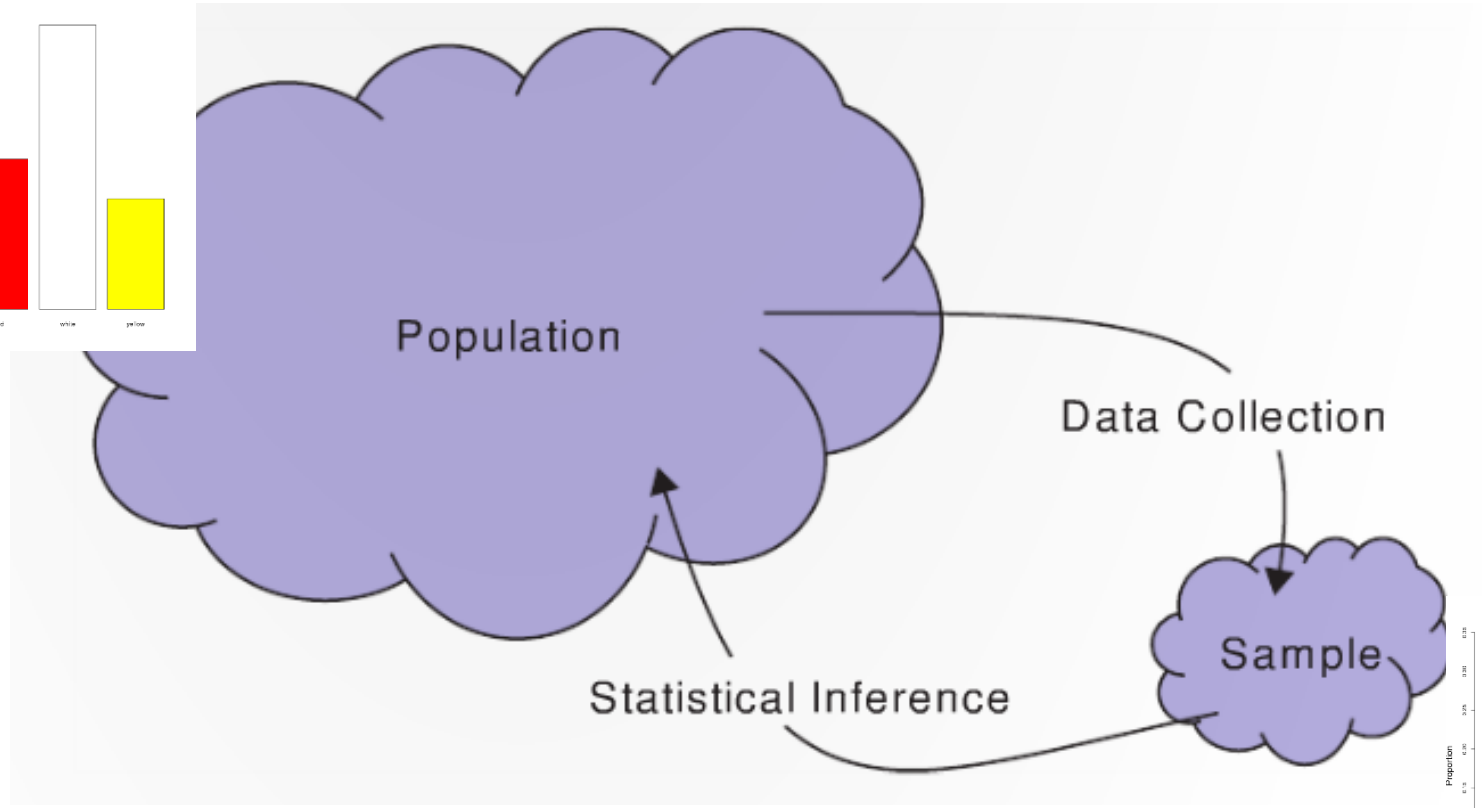
- What is the general ‘shape’ of the data
- Where are the values centered
- How do the data vary

There are all properties of how the data is ***distributed***

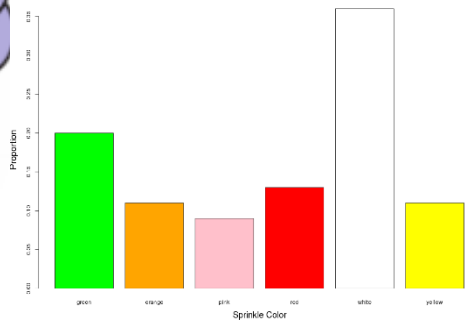
Last class: for categorical data we had…

Categorical Distribution (π)

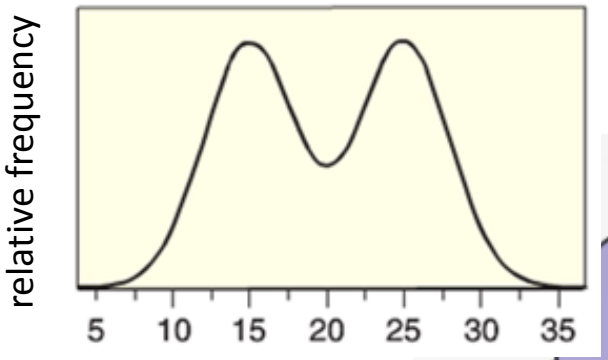


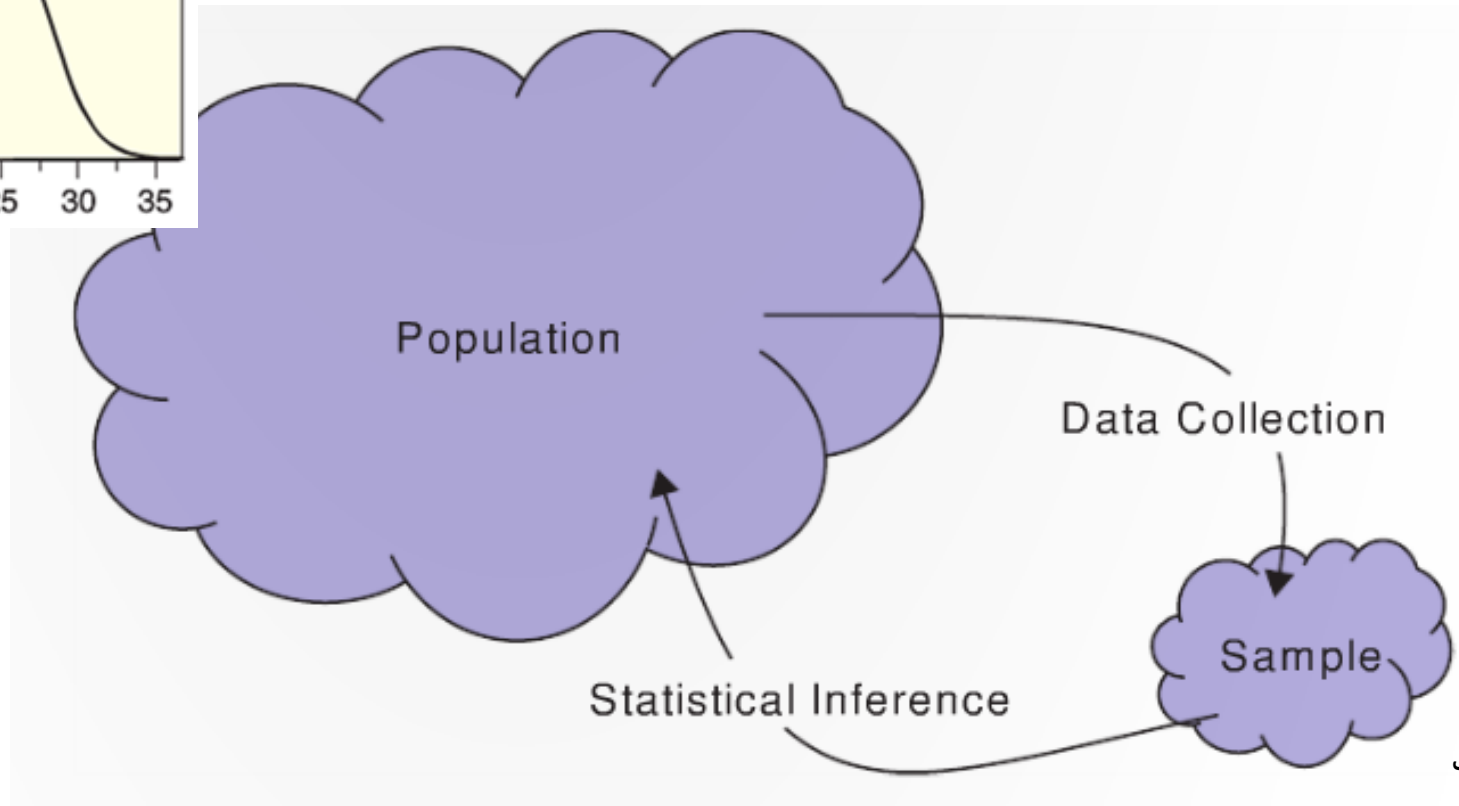


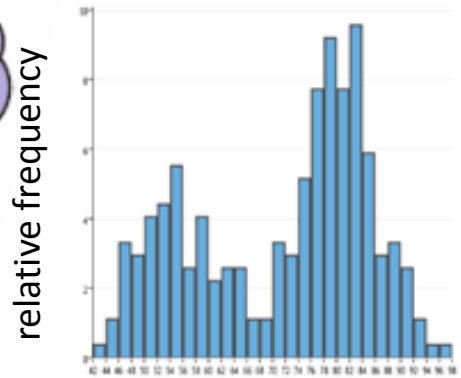
Bar chart ($\hat{p}$)



Population distributions and sample histograms





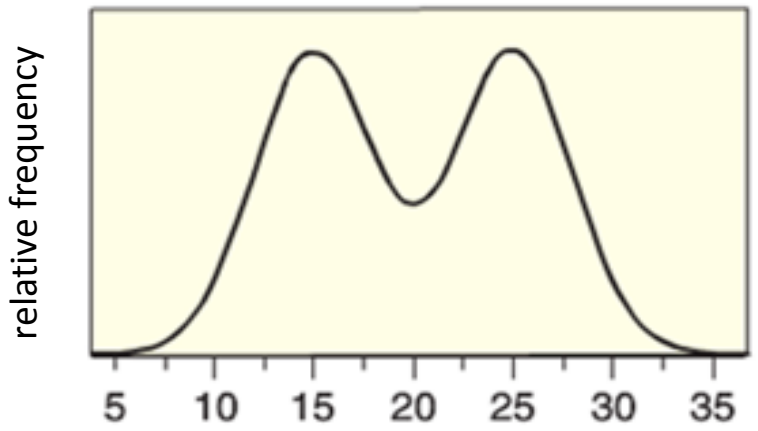


Histograms

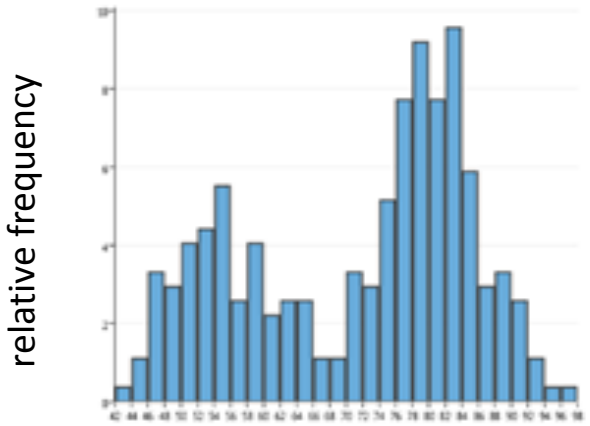
Histograms are a way of visualizing a sample of quantitative data

- They are similar to bar charts but for quantitative variables
- They aim to give a picture of how the data is distributed

Continuous distribution



Histogram



Gapminder data and data frames

```
# get a data frame with information about the countries in the world
  > source("/home/shared/intro_stats/cs206_functions.R")
  > country_data <- get_cs206_data("gapminder_2007")
  > View(country_data)
```

	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	2007	43.828	31889923	974.5803
2	Albania	Europe	2007	76.423	3600523	5937.0295
3	Algeria	Africa	2007	72.301	33333216	6223.3675
4	Angola	Africa	2007	42.731	12420476	4797.2313
5	Argentina	Americas	2007	75.320	40301927	12779.3796

Hans Rosling's gapminder

Gapminder data

Questions:

1) What are the observational units (cases)?

2) What are the variables?

3) Are the variable categorical or quantitative?

4) What is the population?

	country	continent	year	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	2007	43.828	31889923	974.5803
2	Albania	Europe	2007	76.423	3600523	5937.0295
3	Algeria	Africa	2007	72.301	33333216	6223.3675
4	Angola	Africa	2007	42.731	12420476	4797.2313
5	Argentina	Americas	2007	75.320	40301927	12779.3796

Gapminder data

	country	continent	year	lifeExp
1	Afghanistan	Asia	2007	43.828
2	Albania	Europe	2007	76.423
3	Algeria	Africa	2007	72.301
4	Angola	Africa	2007	42.731
5	Argentina	Americas	2007	75.320

Data frames are the way R represents structured data

Data frames can be thought of as collections of related vectors

- Each vector corresponds to a variable in the structured data

We can access individual vectors of data using the $ symbol

```
# we can look at the number of countries in each continent
  > continents <- country_data$continent     # continent is a categorical variable
  > continent_table <- table(continents)
  > barplot(continent_table)
```

Gapminder: life expectancy in different countries

Let’s look at the life expectancy in different countries, which is a quantitative variable

```
# pull a vector of life expectancies from the data frame
    > life_expectancy <-  country_data$lifeExp
```

Histograms – countries life expectancy in 2007

Life expectancy for different countries for 142 countries in the world:

- 43.83, 72.30, 76.42, 42.73, …

To create a histogram we create a set of intervals

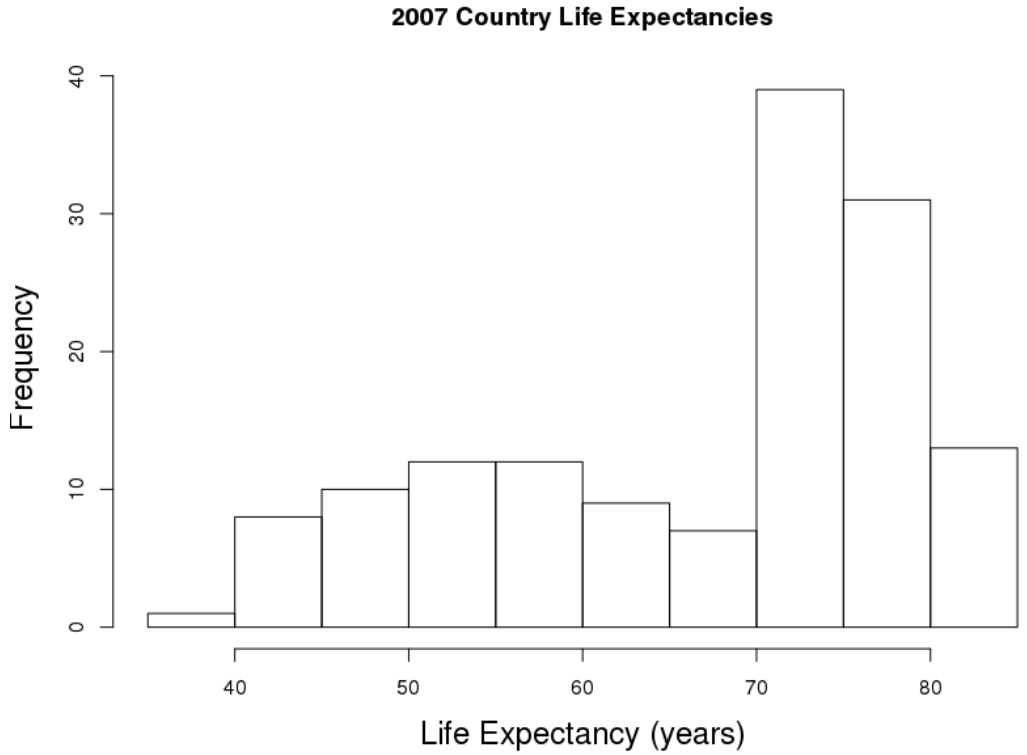
- 35-40, 40-45, 45-50, … 75-80, 80-85

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

Histograms – countries life expectancy in 2007

Life Expectancy	Frequency Count
(35 – 40]	1
(40 – 45]	8
(45 – 50]	10
(50 – 55]	12
(55 – 60]	12
(60 – 65]	9
(65 – 70]	7
(70 – 75]	39
(75 – 80]	31
(80 – 85]	13



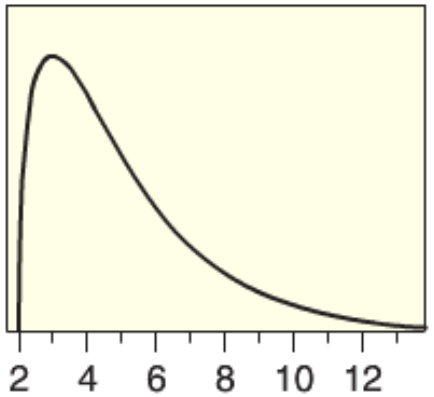
R: `hist(v)`

Gapminder: life expectancy in different countries

Try creating a histogram of the life expectancy in different countries using the hist() function

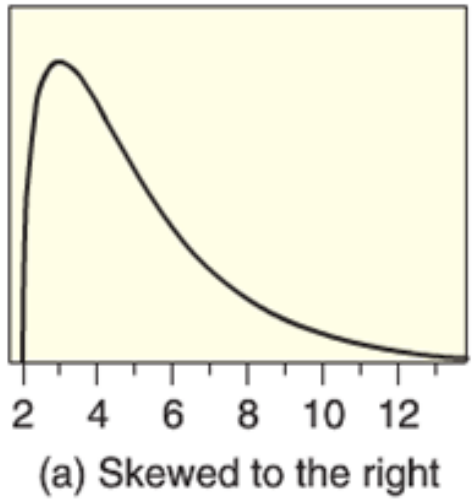
```
# pull a vector of life expectancies from the data frame
> life_expectancy <-  country_data$lifeExp
> hist(life_expectancy)
```

Common shapes for distributions



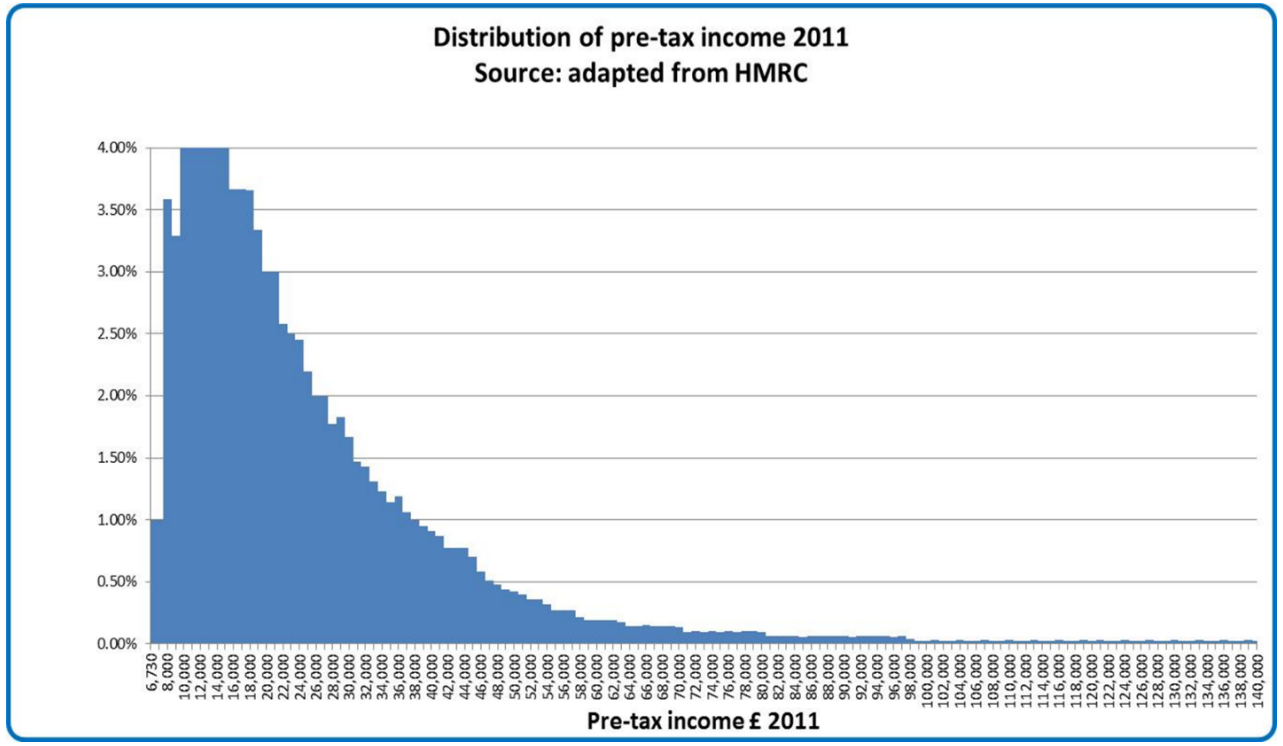
(a) Skewed to the right

Can you think of a distribution that is right skewed?

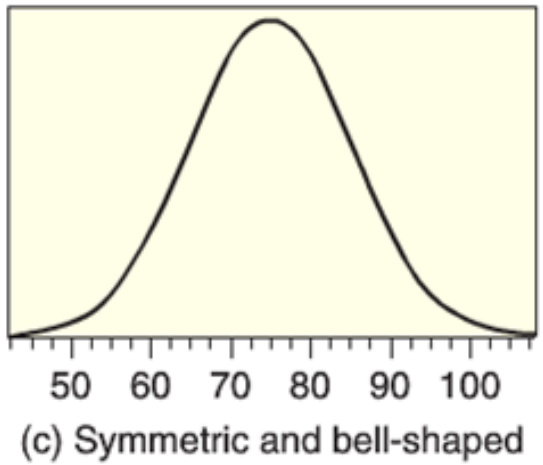


(a) Skewed to the right

Income distribution

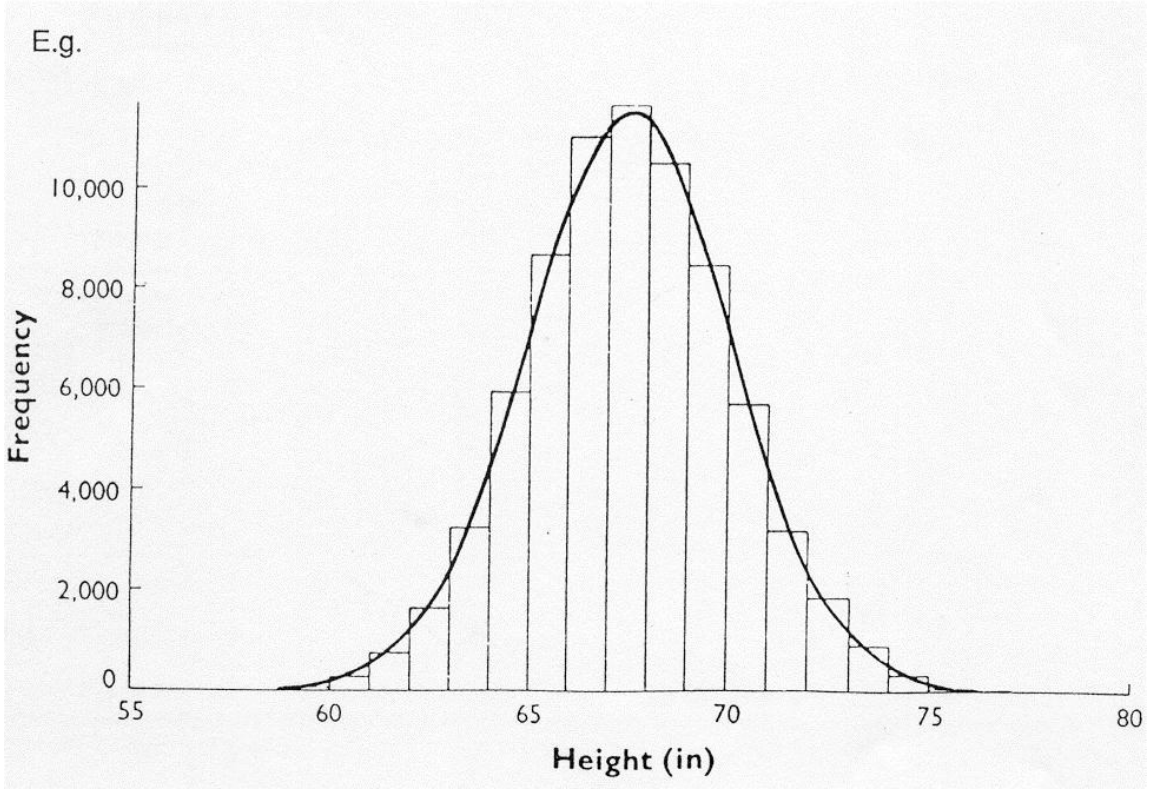


Can you think of a distribution that is symmetric and bell-shaped?

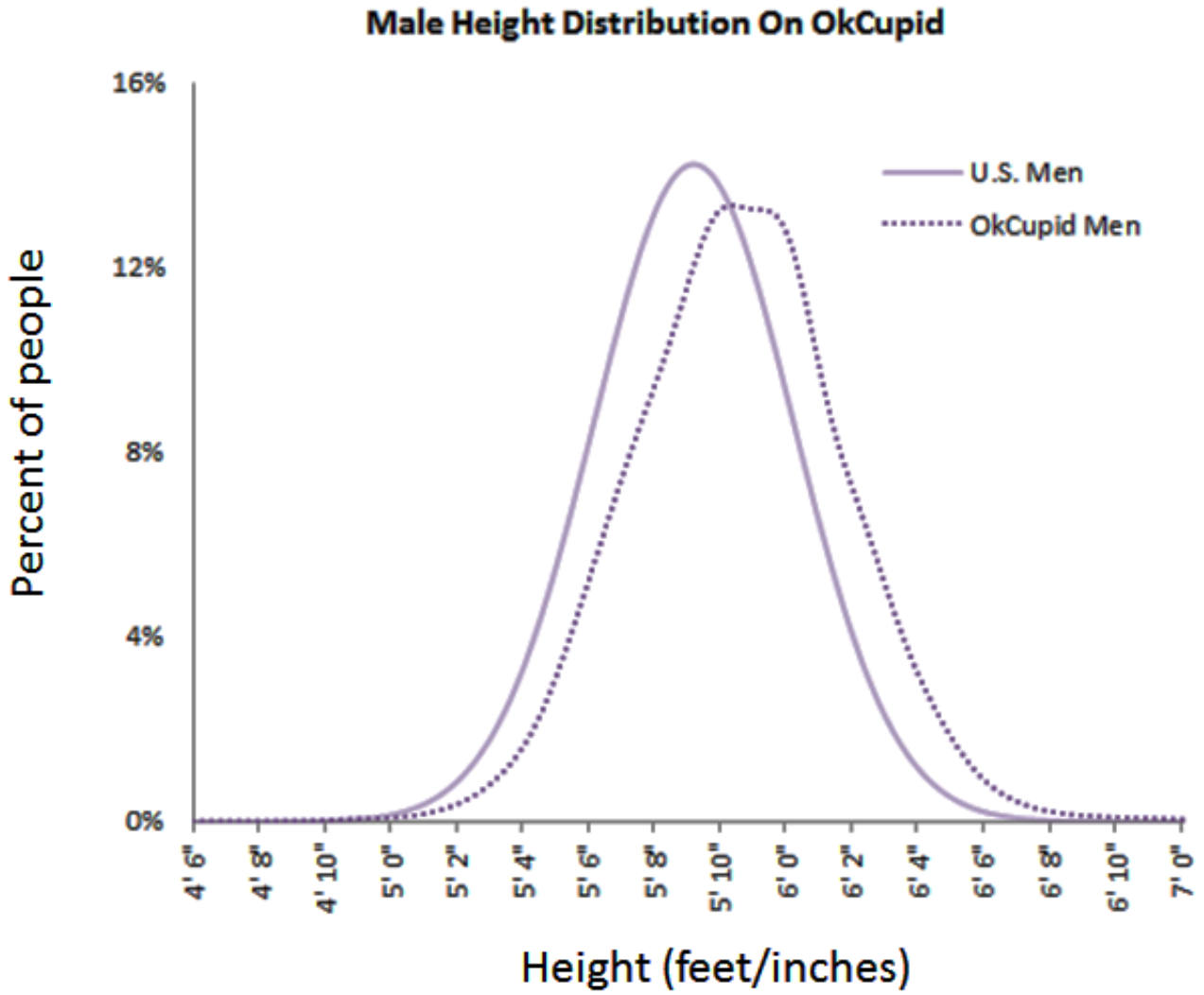


(c) Symmetric and bell-shaped

Young adult male heights (Martin, 1949)

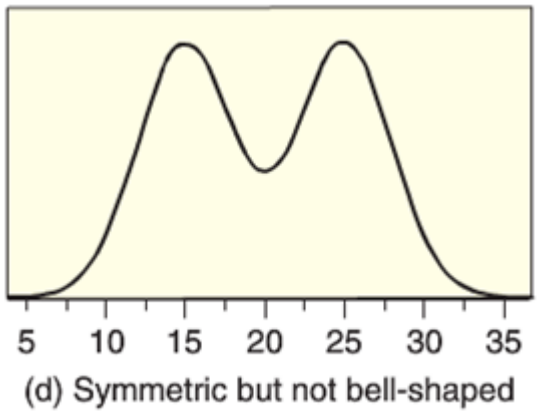


Men on OkCupid are taller!



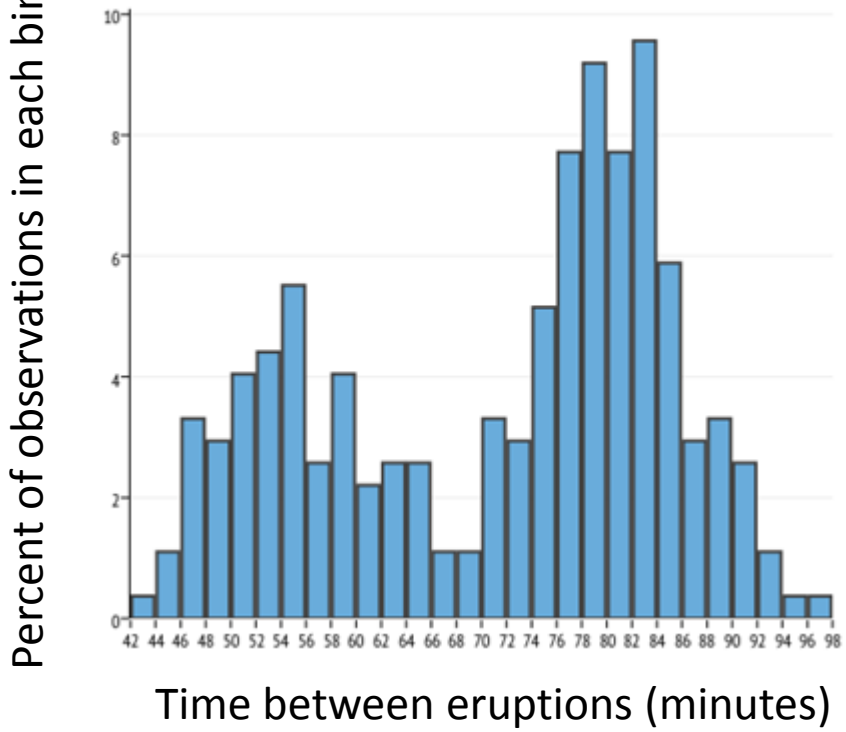
Bias?

Can you think of a distribution that is symmetric but not bell-shaped?

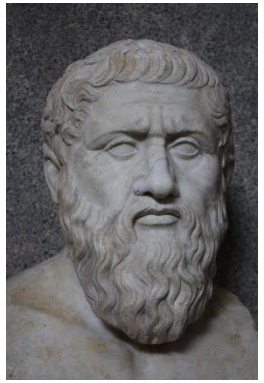


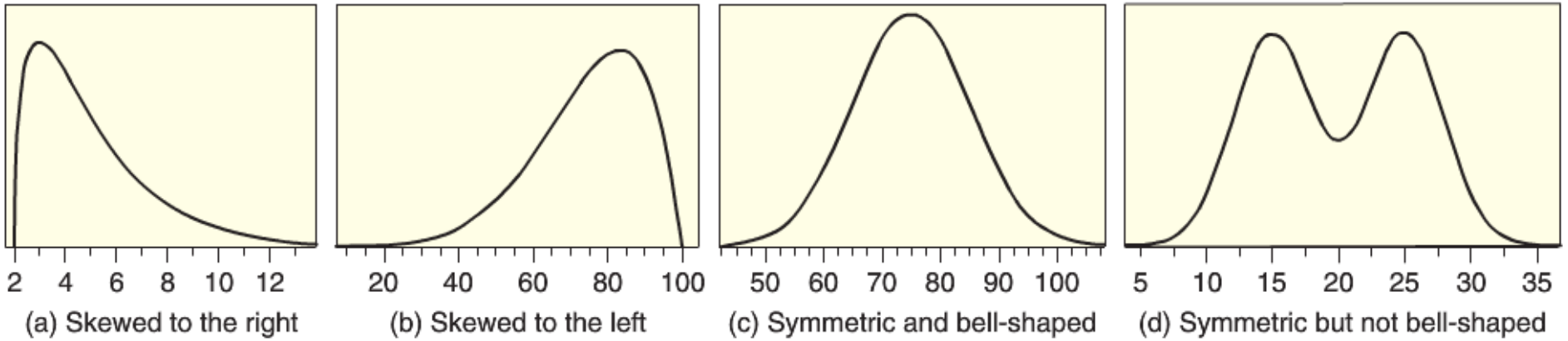
(d) Symmetric but not bell-shaped

Old Faithful eruption times

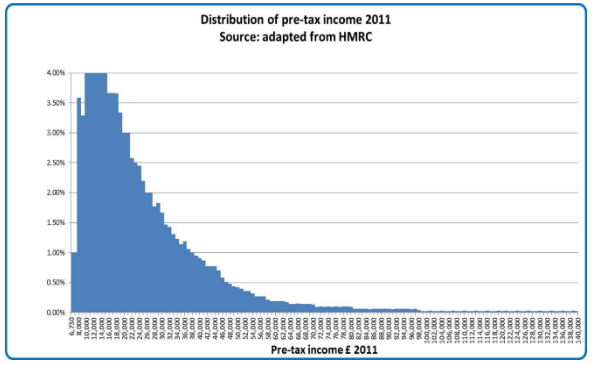


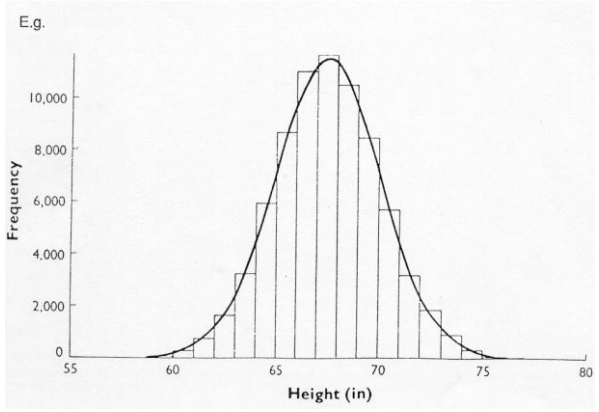
Plato and shadows: distributions and histograms

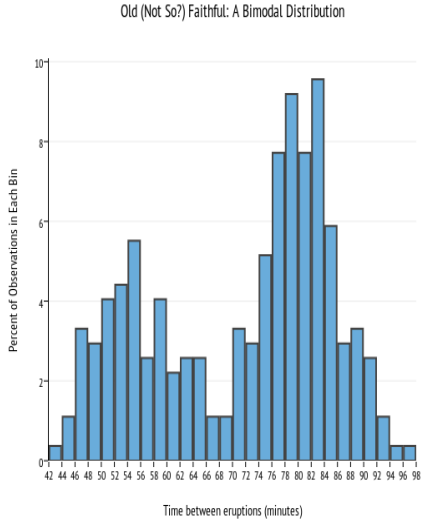




Income distribution







Summary of concepts

1. A ***probability distribution*** shows the ***relative likelihood*** that we will get a data point in the population with a particular value
    - (for a more precise definition take a class in probability)

2. Distributions can have different shapes
    - E.g., left skewed, right skewed, bell shaped, etc.

Summary of R

Data frames contain structured data

- We can view a data frame in R Studio (not in Markdown) using

  > View(my_data_frame)

- We can extract vectors from a data frame using:

  > my_vec <- my_data_frame$my_var

We can get a sense of how quantitative data is distributed by creating a histogram

> hist(my_vec)

Practice at home

Lock5 questions:

- Proportions
  - warmups: 2.1, 2.3, 2.5, 2.7, 2.9 (both editions)
  - 2.13 (2nd edition 2.15) Rock papers scissors

- Quantitative data (shape and central tendency)
  - 2.33, 2.35, 2.37 (2nd edition 2.43, 2.45, 2.47)

Experiment with the Gapminder data frame:

- Create some bar and pie charts for the categorical data
- Create some histograms for the quantitative data