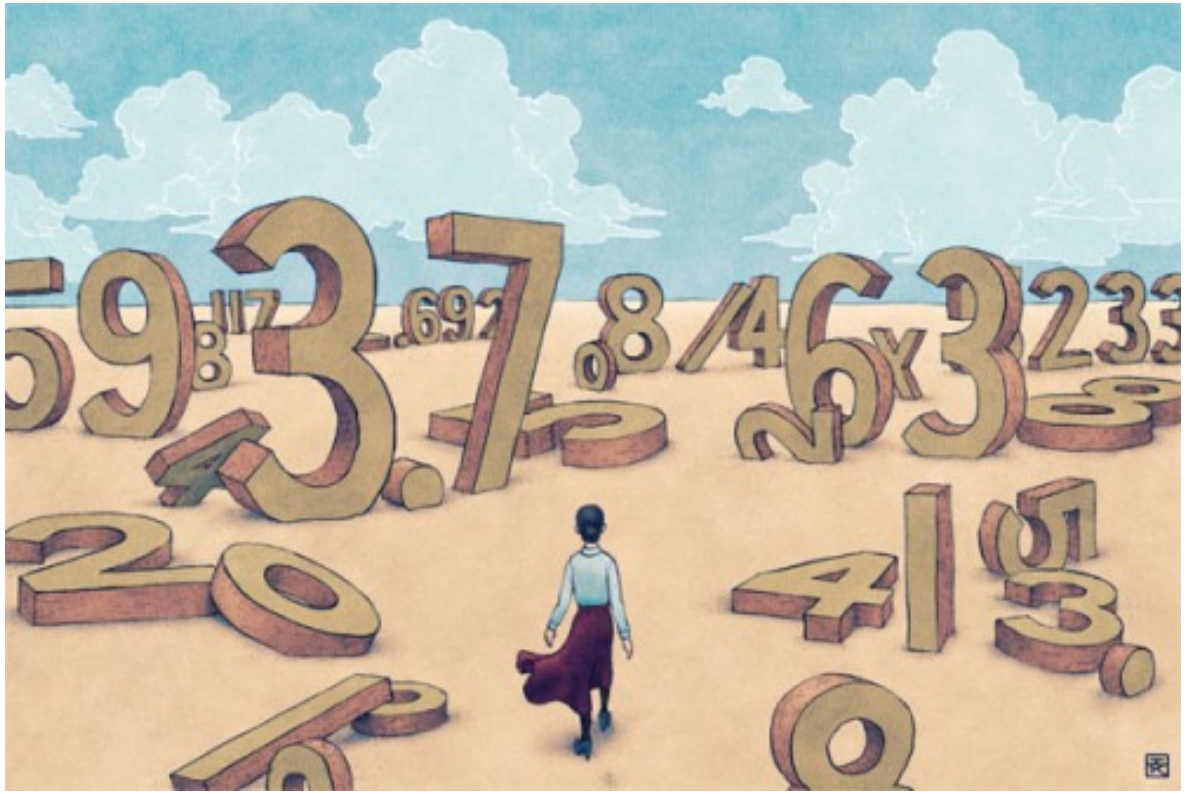
Quantitative data: shape and measures of central tendency



Overview

Review of categorical data concepts and R

Quantitative data

- Graphing the shape: histograms and outliers
- Measures of the central tendency: mean and median

Review

Categorical variables

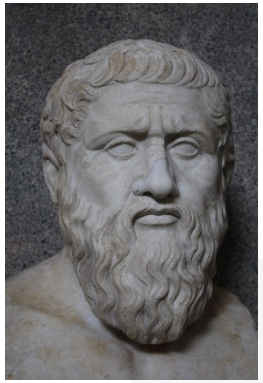
Quiz: Art time!

Please draw:

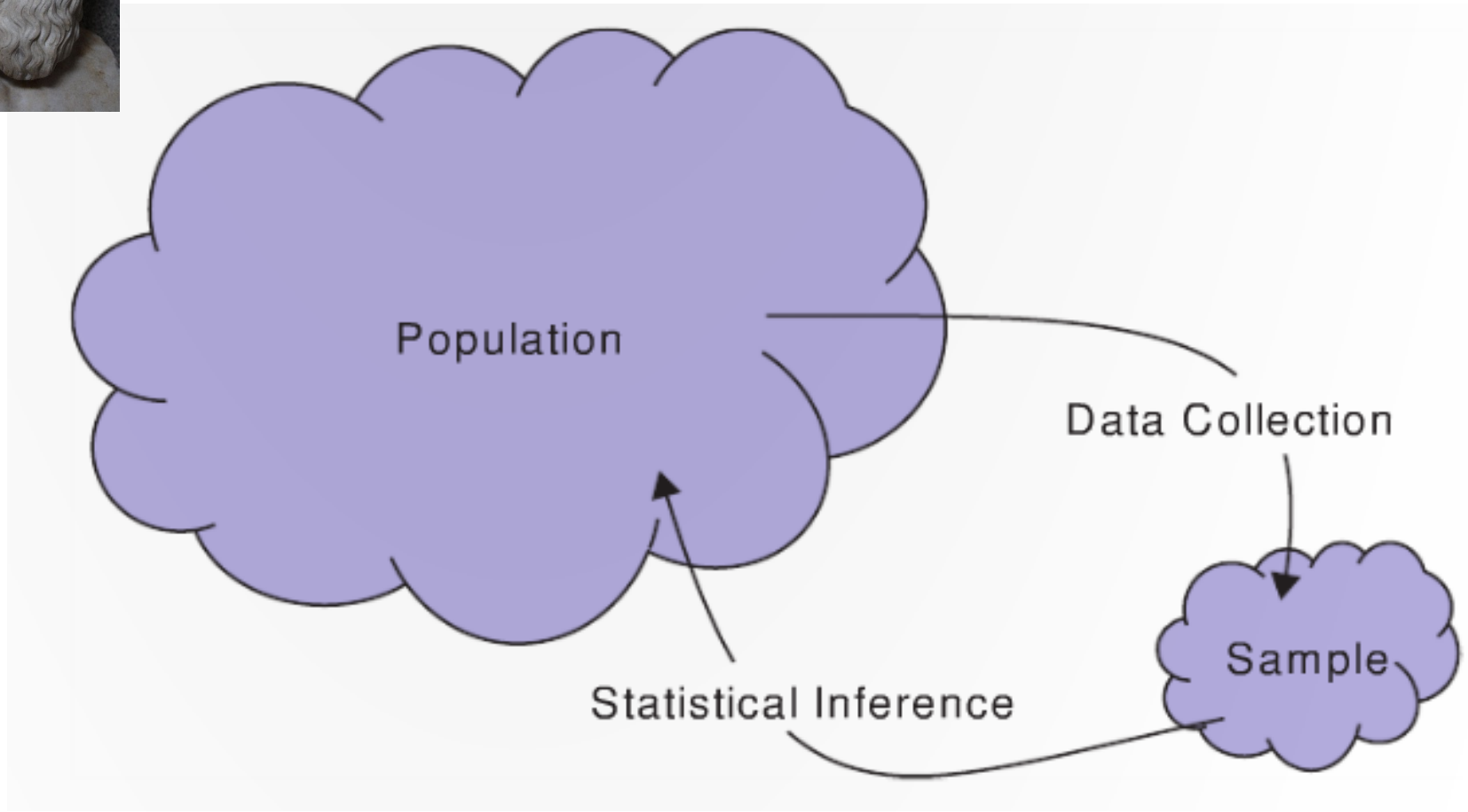
1. A population – and label it a “population”
2. A sample – and label it “sample”
3. Add the label “parameter” in the appropriate location
4. Add the label “statistic” in the appropriate location
5. Add the symbol for a population proportion in the appropriate location
6. Add the symbol for a sample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

Upload your file to Canvas survey

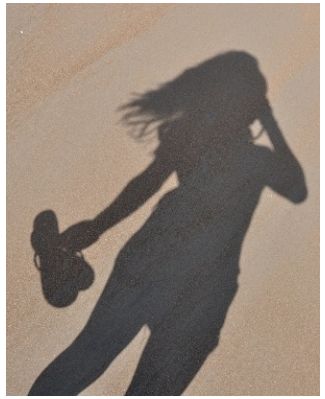
- bragging rights to the best drawing!



parameter: π



statistic: $\hat{p}$

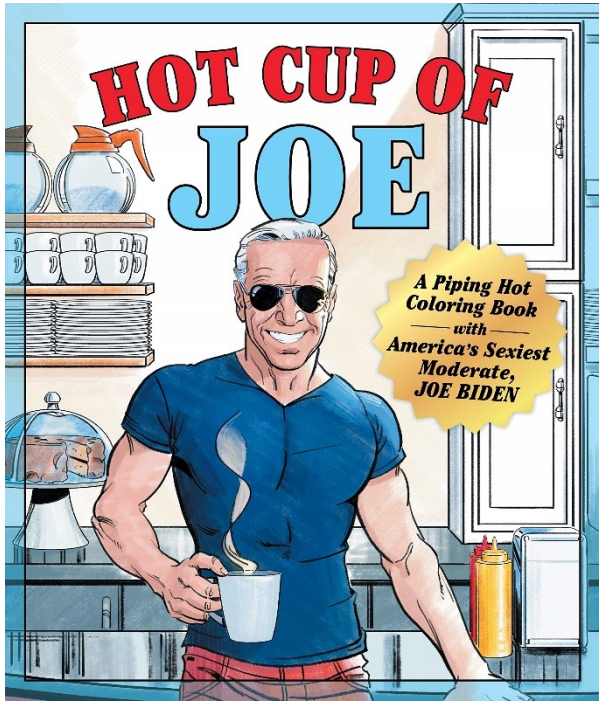


Example of categorical data: Presidential approval ratings

Approve

Disapprove

Example of categorical data: Presidential approval ratings



Approve

Disapprove

Example: Biden's approval rating

```
# get Biden's approval rating from 1,000 simulated voters
> library(SDS100)
> approval_sample <- get_approval_sample(1000)
```

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

<u>Questions:</u>

1. What are the observational units (cases)?
2. What is the variable?
3. What is the population?

Example: Biden's approval rating

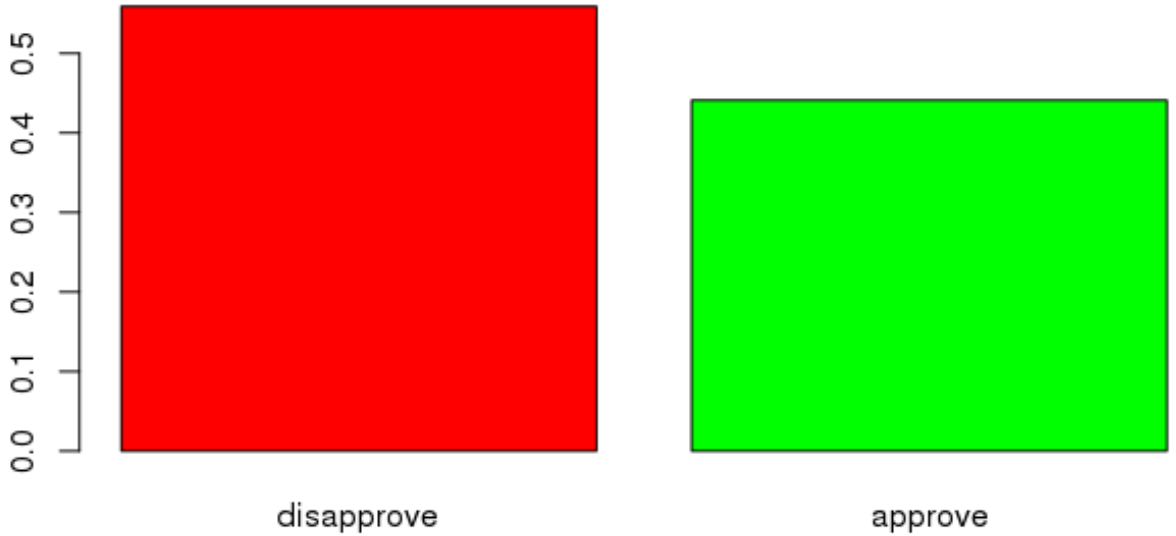
Can you calculate $\hat{p}$ for Biden's approval?

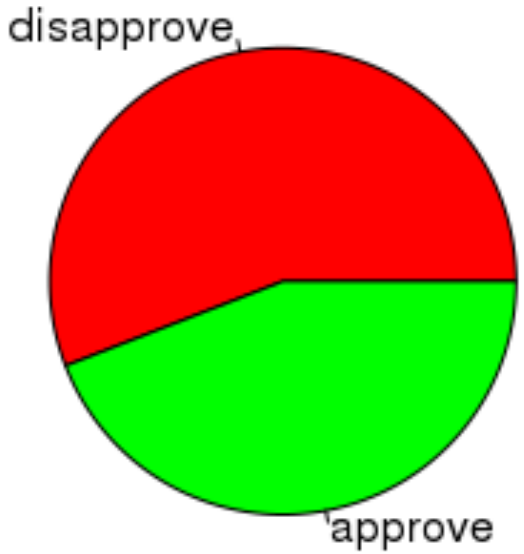
```
> approval_table <- table(approval_sample)
> approval_proportions <- prop.table(approval_table)
> approval_proportions["approve"]
```

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

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

Results fromTrumps' approval rating

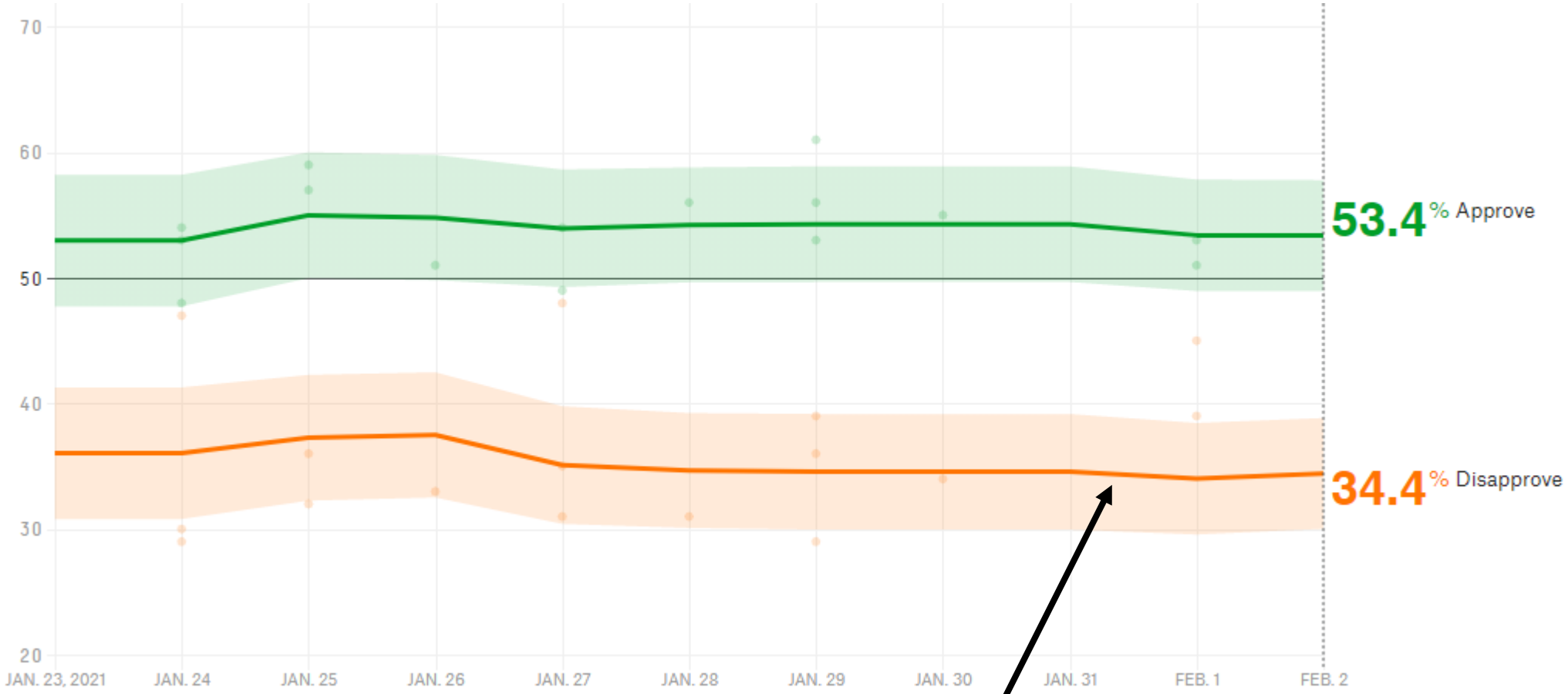




Is this $\pi_{approve}$ or $\hat{p}_{approve}$?

What do Biden's results look like?

Example: Biden's approval rating



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

Can we ever know π?

Usually we are interested in knowing about properties of ***an infinite processes*** so we can never perfectly know a parameter value

- i.e., we can never know π

However, for ***finite populations,*** it is possible to know the value of a parameter exactly

For example, if π is the proportion of voters who will vote for Biden in the 2024 election, then should know π in November 2024

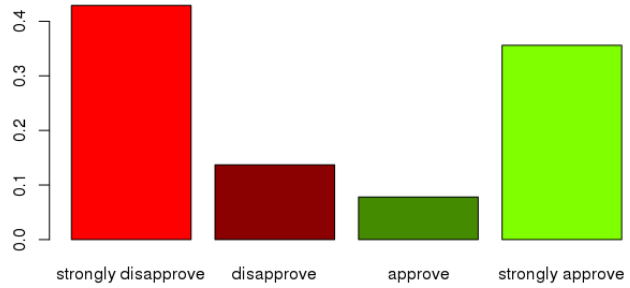
Practice at home

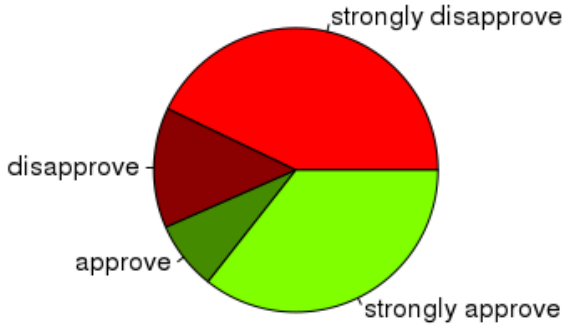
Get the degree to which likely voters approve of Biden:

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

Practice at home:

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





Quantitative variables

Descriptive statistics for one quantitative variable

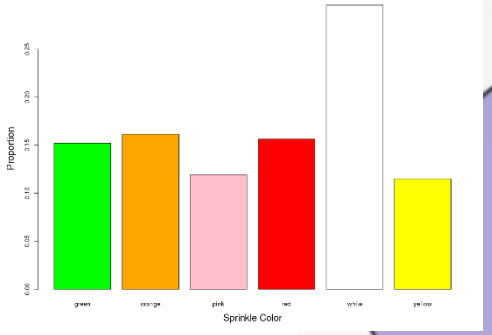
We will be looking at:

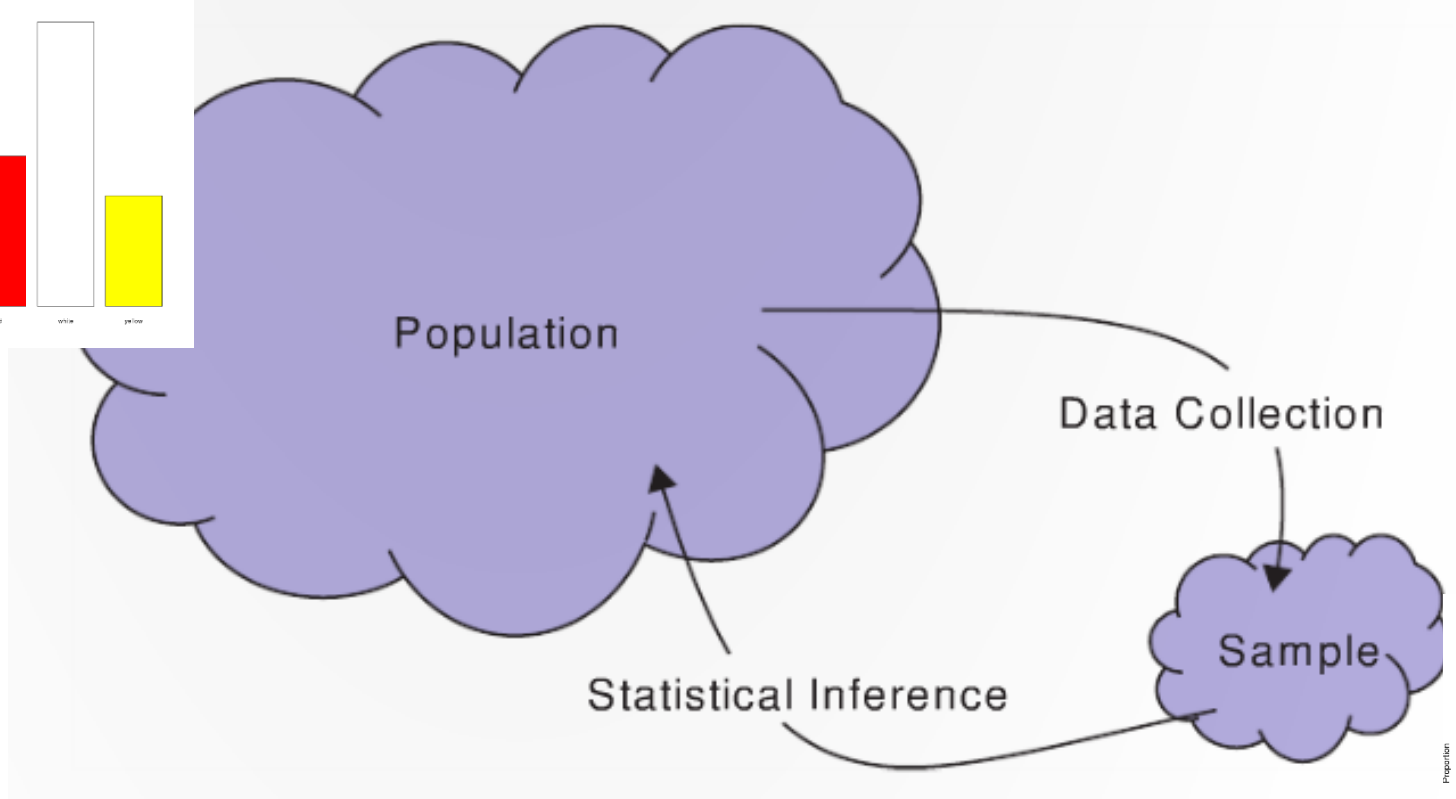
- What is the general ‘shape’ of the data
- Where are the values centered
- How do the data vary

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

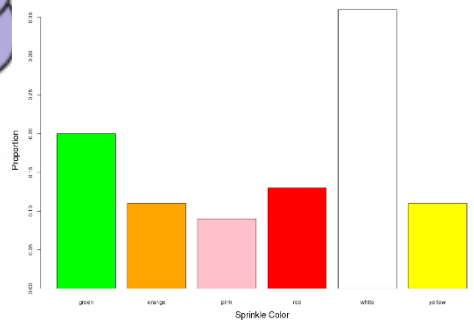
For categorical data we had…

Categorical Distribution (π)

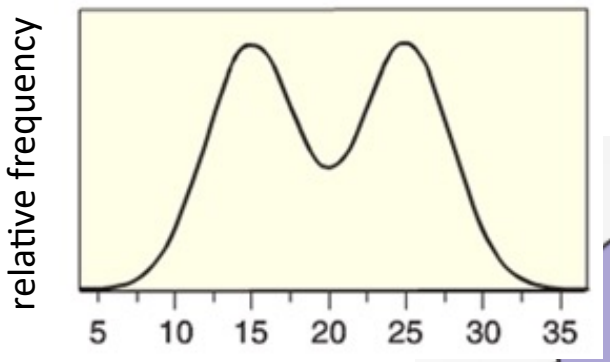


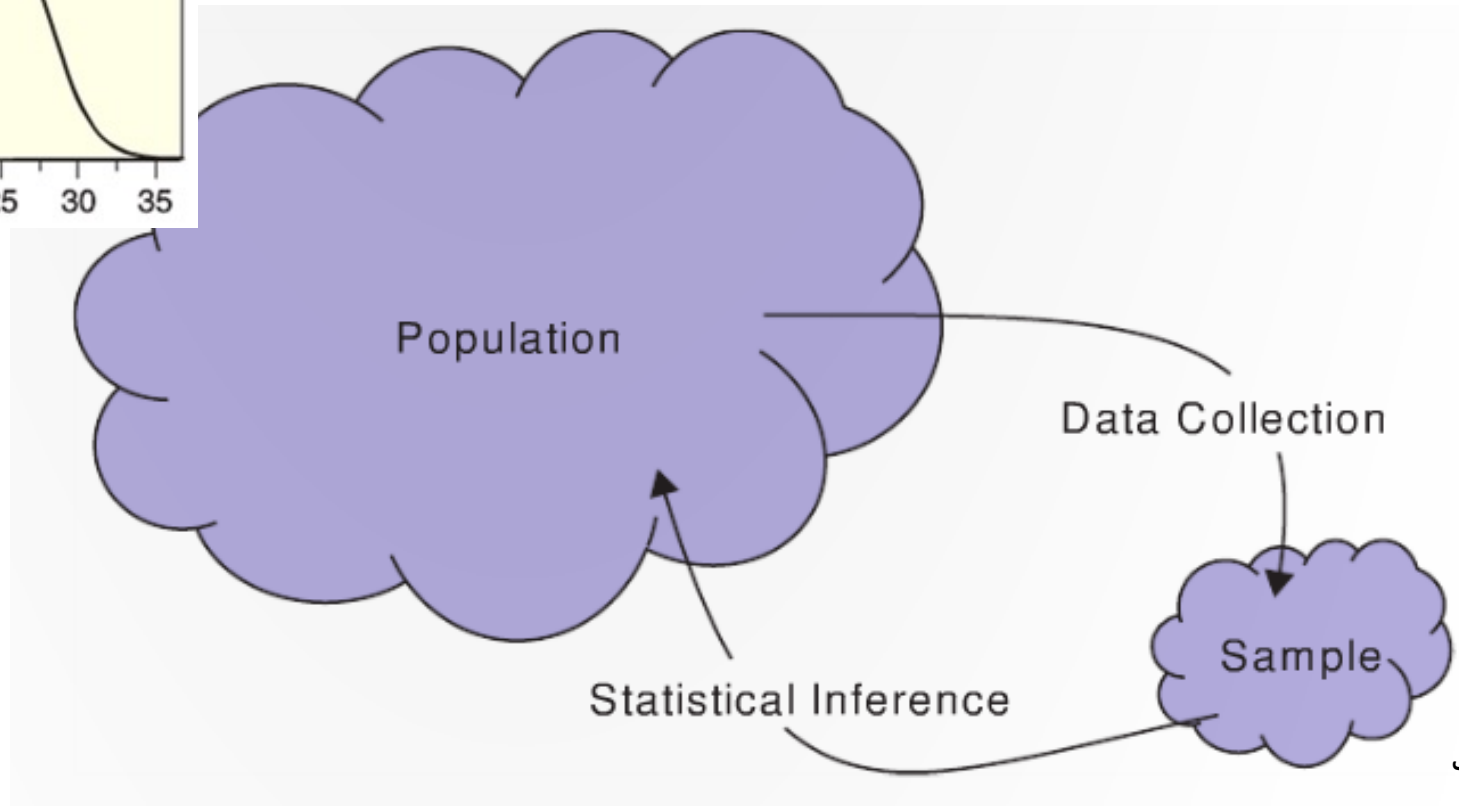


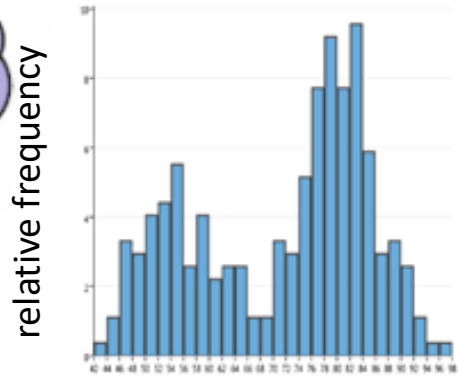
Bar chart ($\hat{p}$)



Population distributions and sample histograms





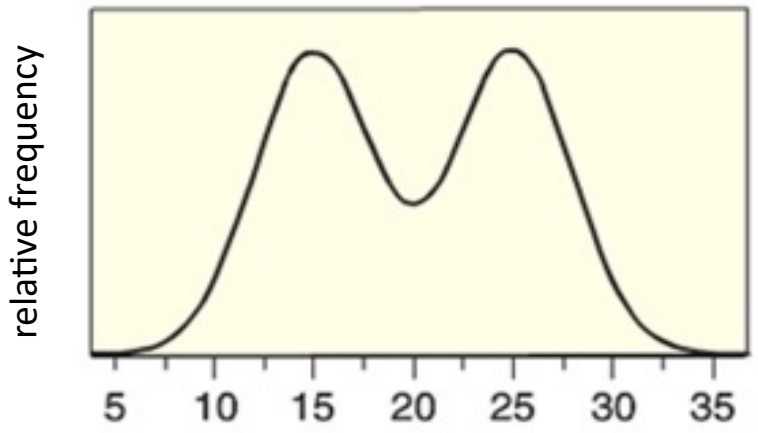


Histograms

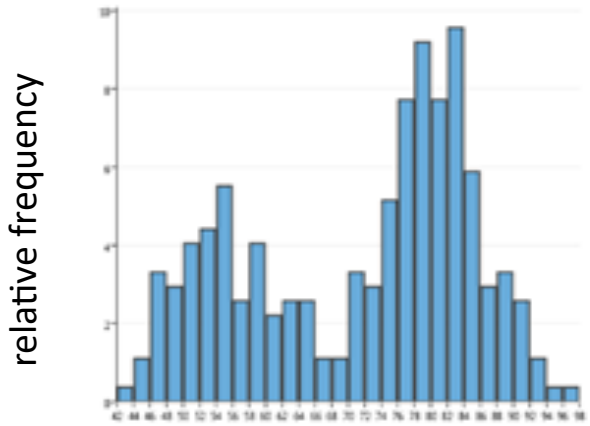
Histograms are a way of visualizing a sample of quantitative data

- 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
  > download_data("gapminder_2007.Rda")   # SDS100 function
  > load("gapminder_2007.Rda")
  > View(gapminder_2007)
```

	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 <- gapminder_2007$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 <-  gapminder_2007$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

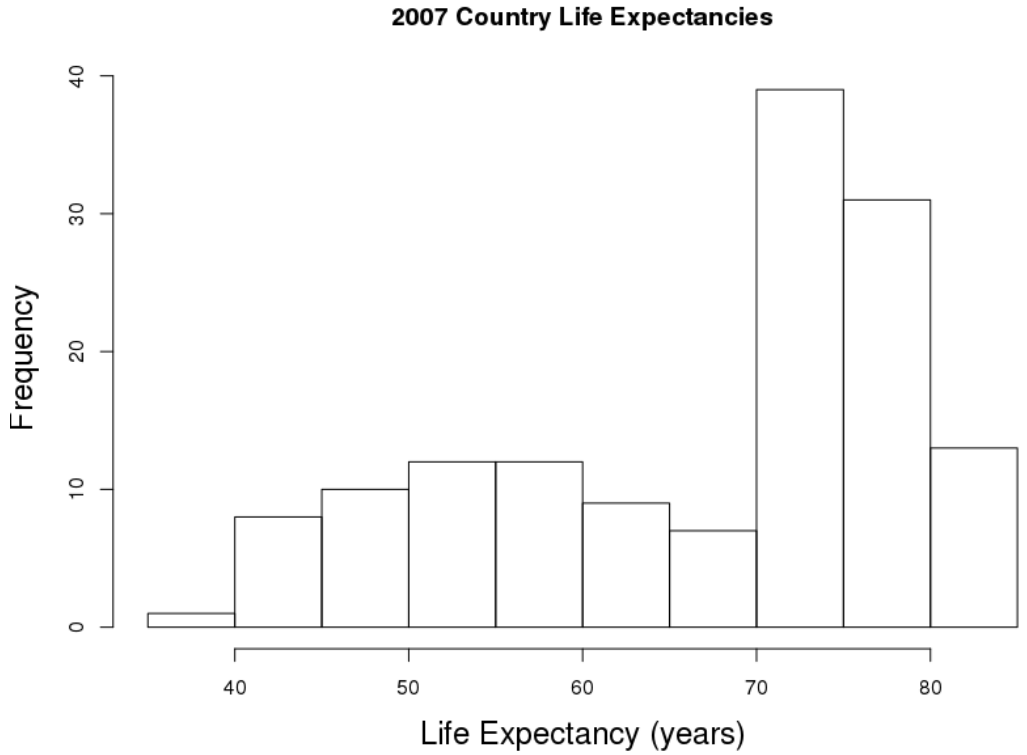
- 35-40, 40-45, 45-50, … 75-80, 80-85

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



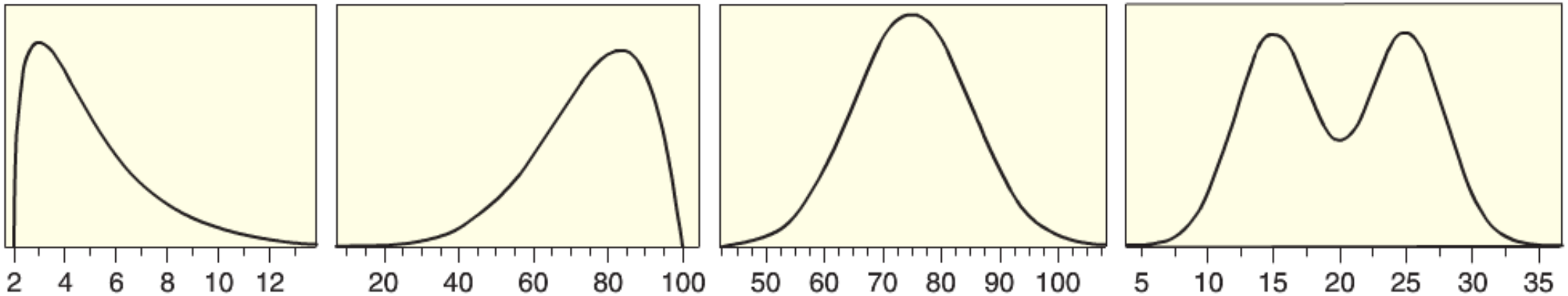
R: `hist(v)`

Gapminder: life expectancy in different countries

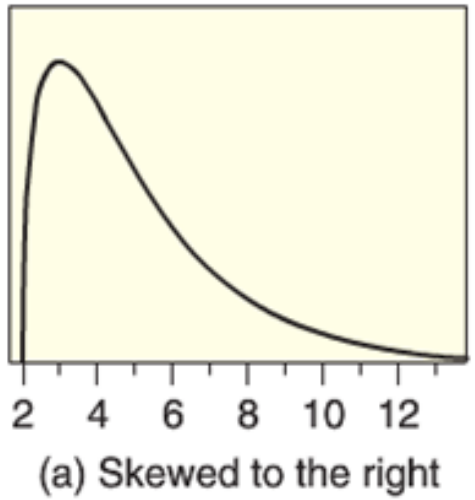
Try creating a histogram of the life expectancy in different countries using the hist() function

```
# pull a vector of life expectancies from the data frame
> life_expectancy <-  gapminder_2007$lifeExp
> hist(life_expectancy)
```

Common shapes for distributions

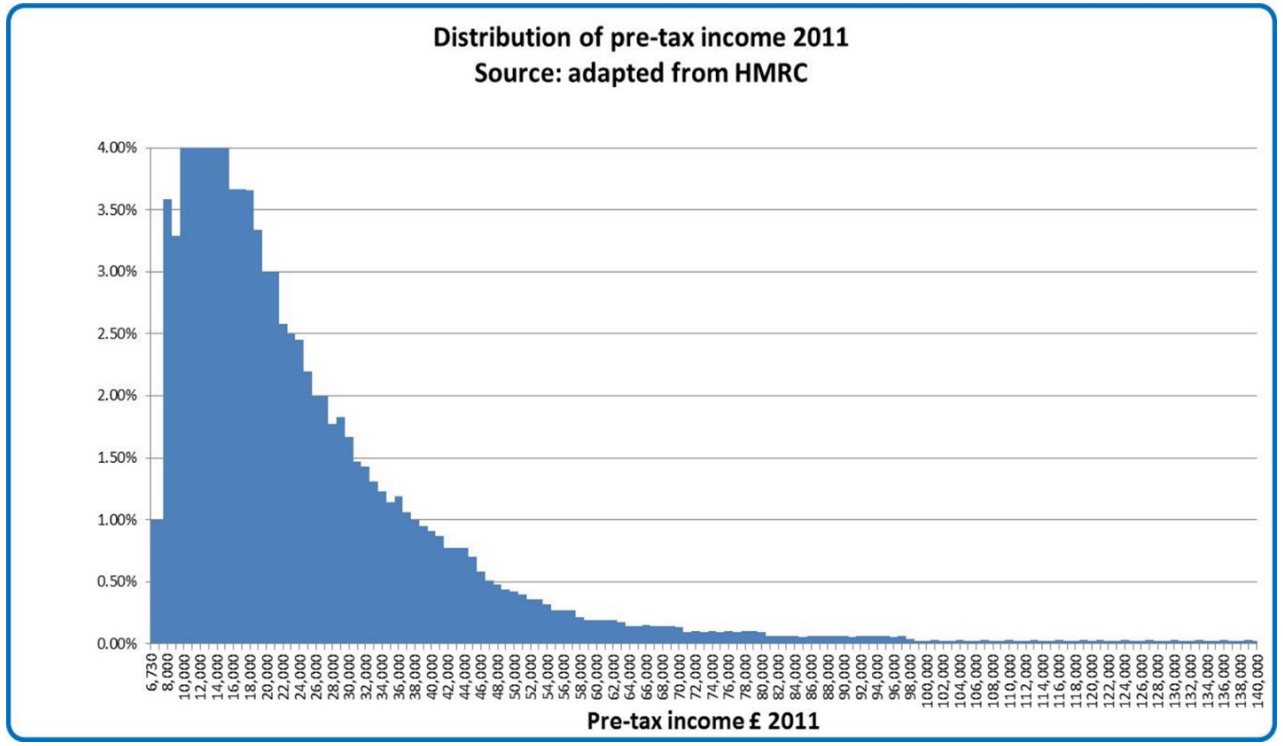


Can you think of a distribution that is right skewed?

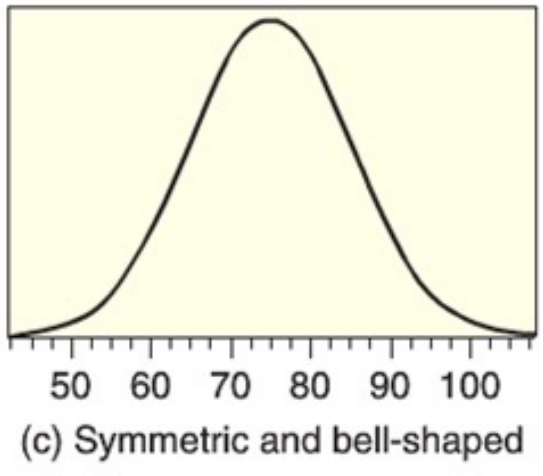


(a) Skewed to the right

Income distribution

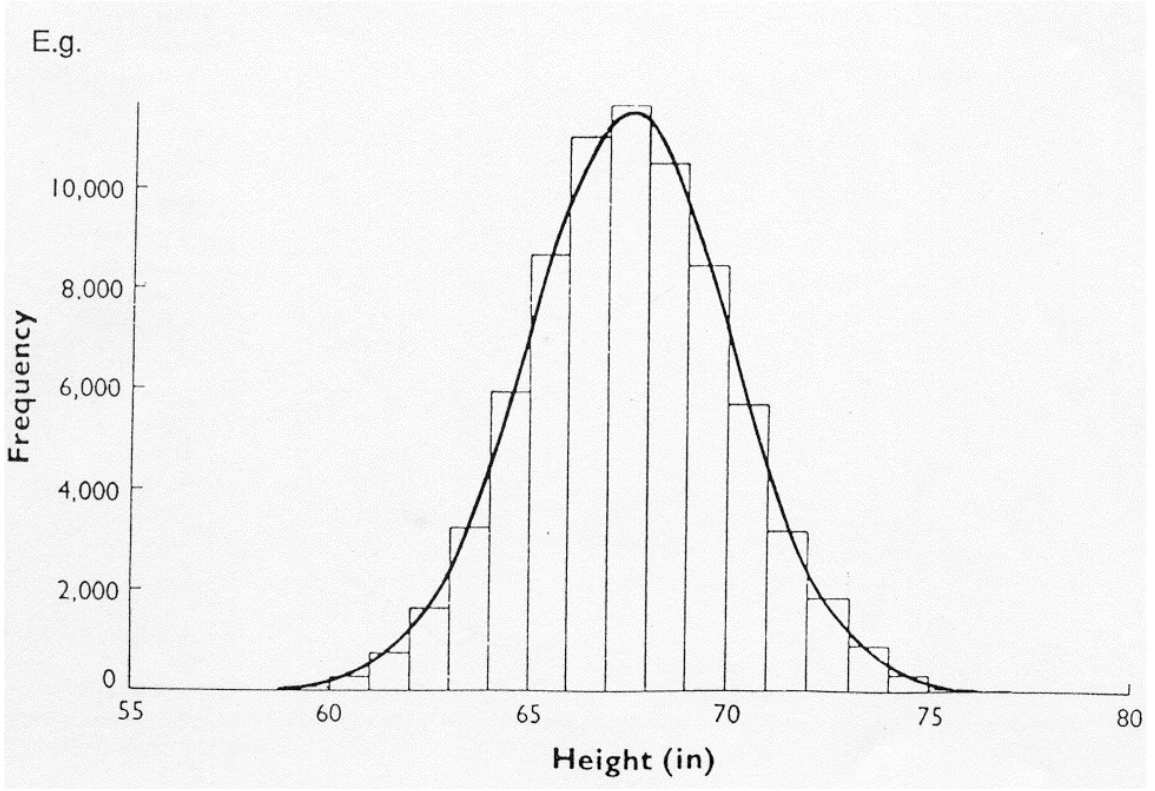


Can you think of a distribution that is symmetric and bell-shaped?

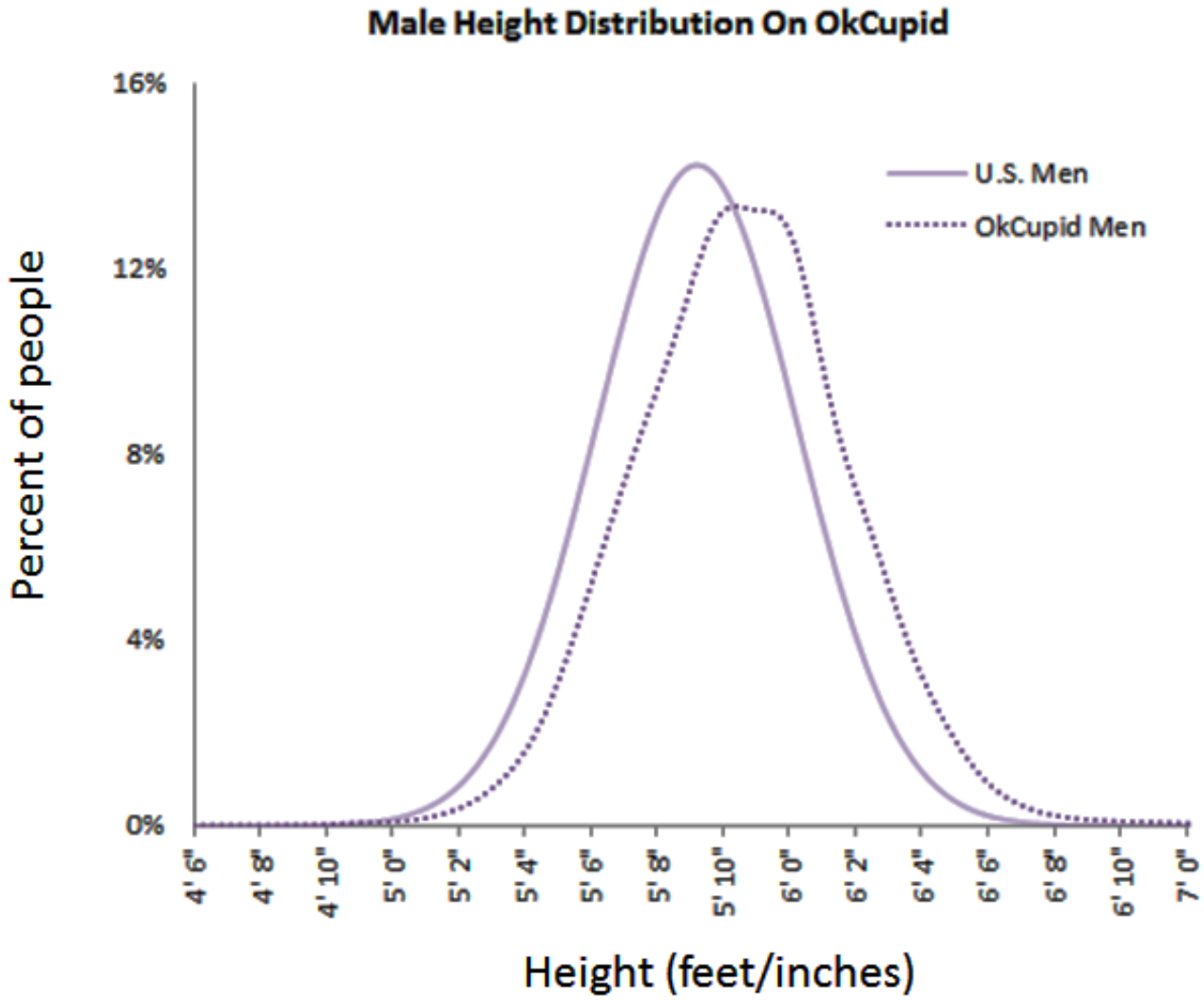


(c) Symmetric and bell-shaped

Young adult male heights (Martin, 1949)

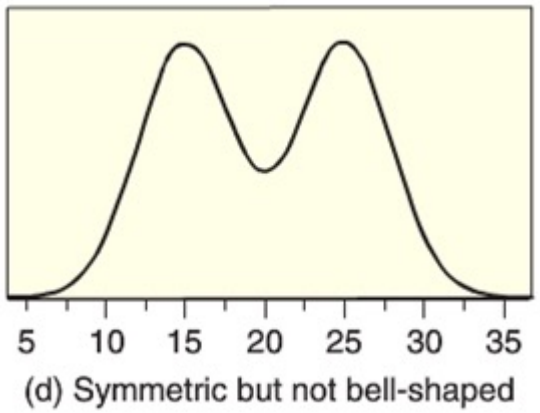


Men on OkCupid are taller!



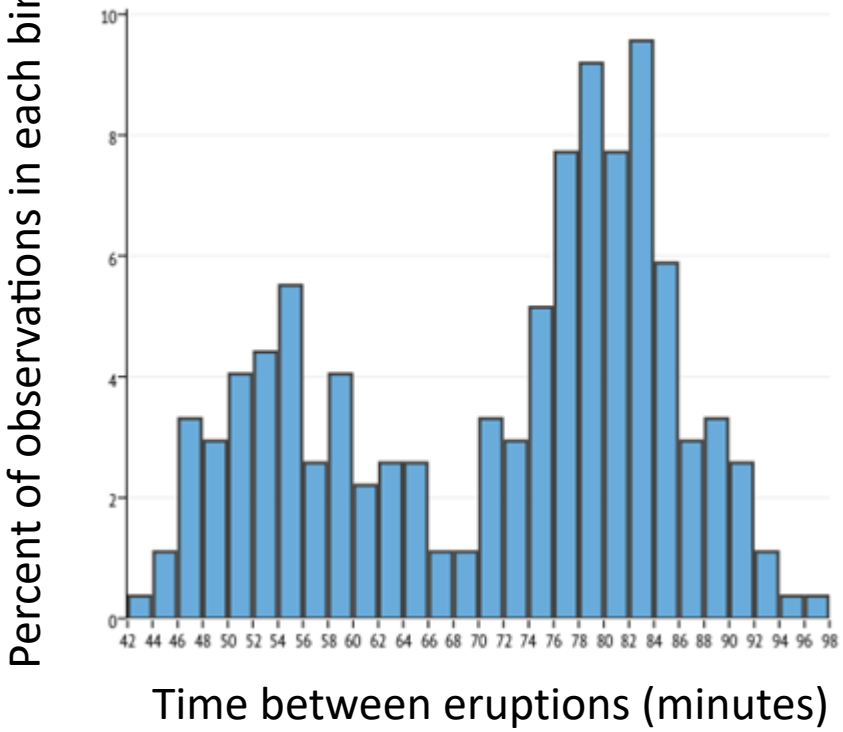
Bias?

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

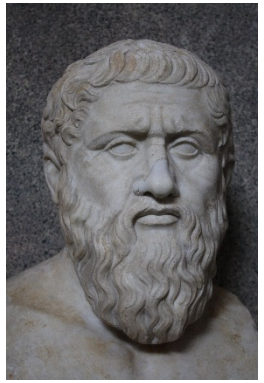


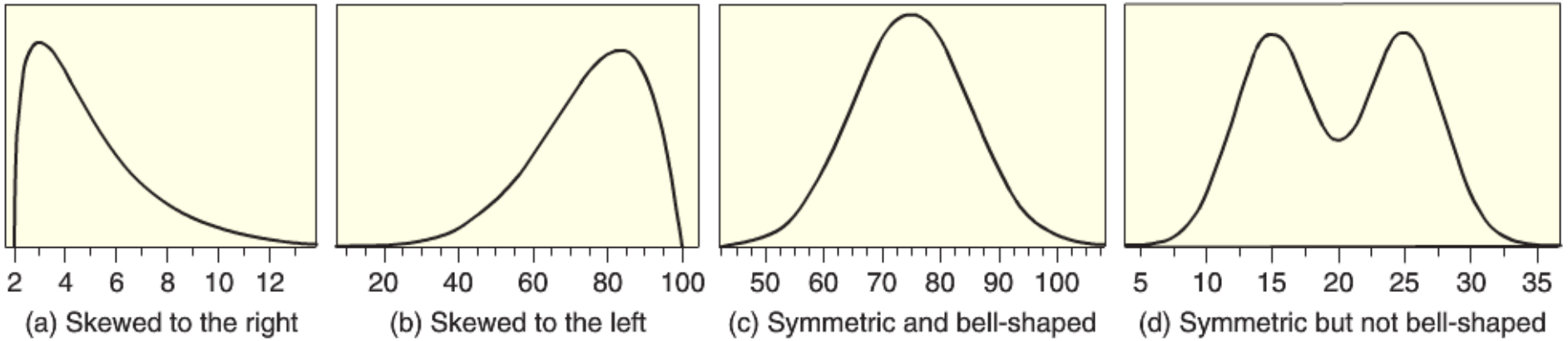
(d) Symmetric but not bell-shaped

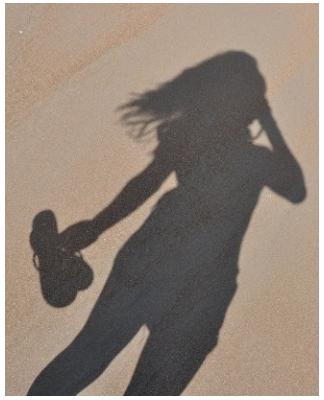
Old Faithful eruption times



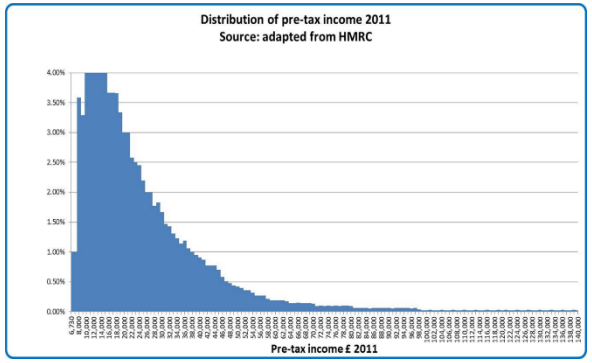
Plato and shadows: distributions and histograms

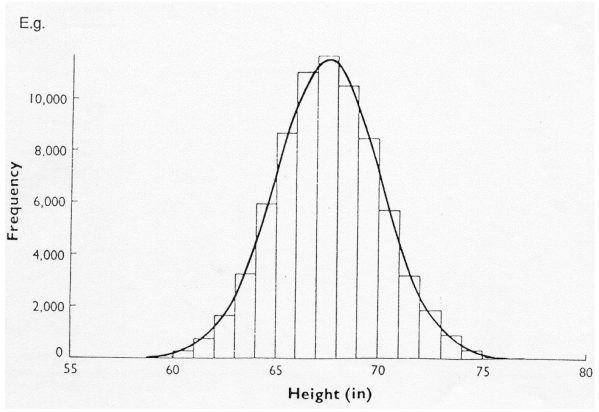


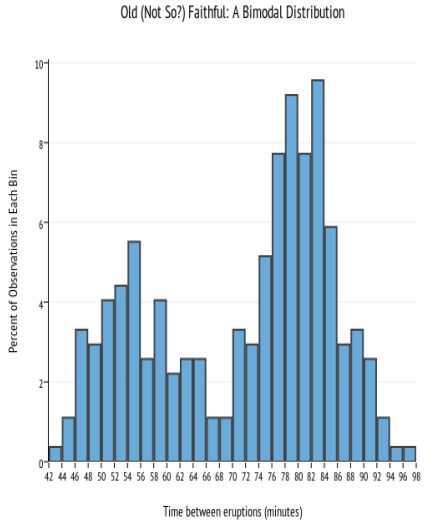


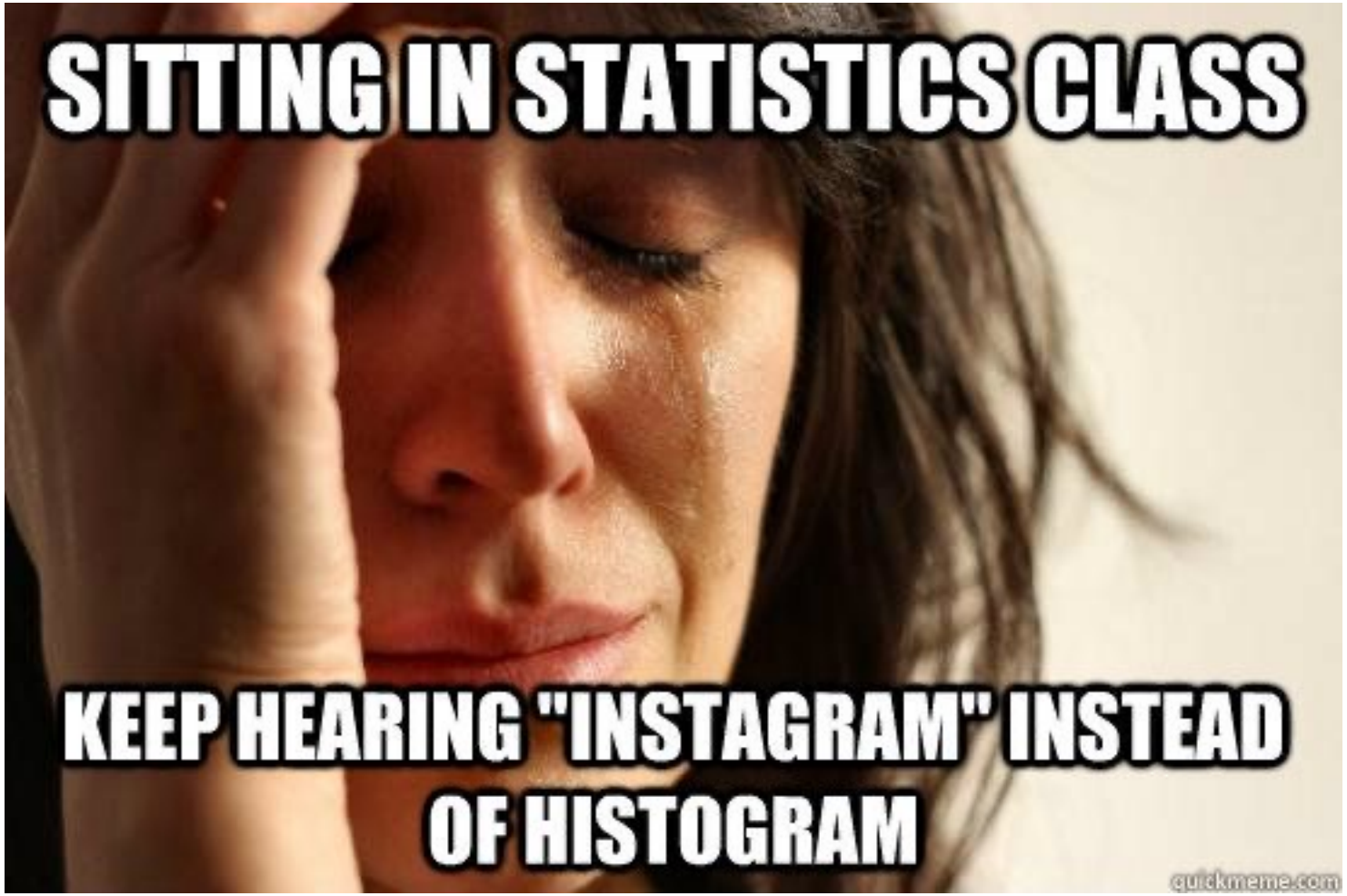


Income distribution



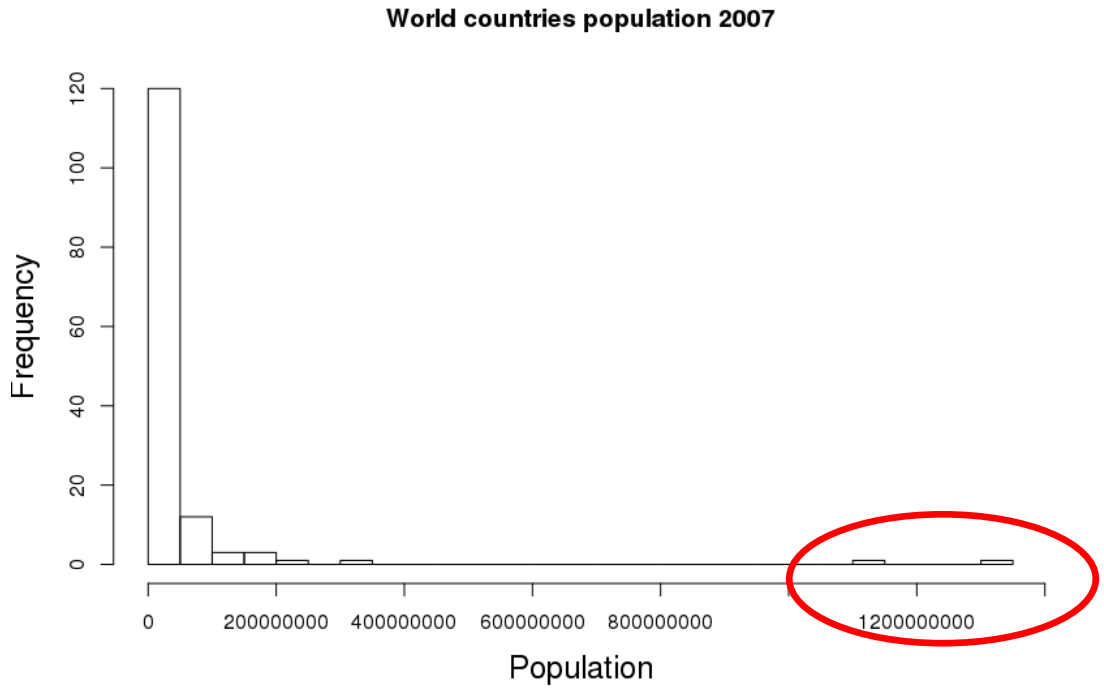





SITTING IN STATISTICS CLASS
KEEP HEARING "INSTAGRAM" INSTEAD OF HISTOGRAM
quickmeme.com

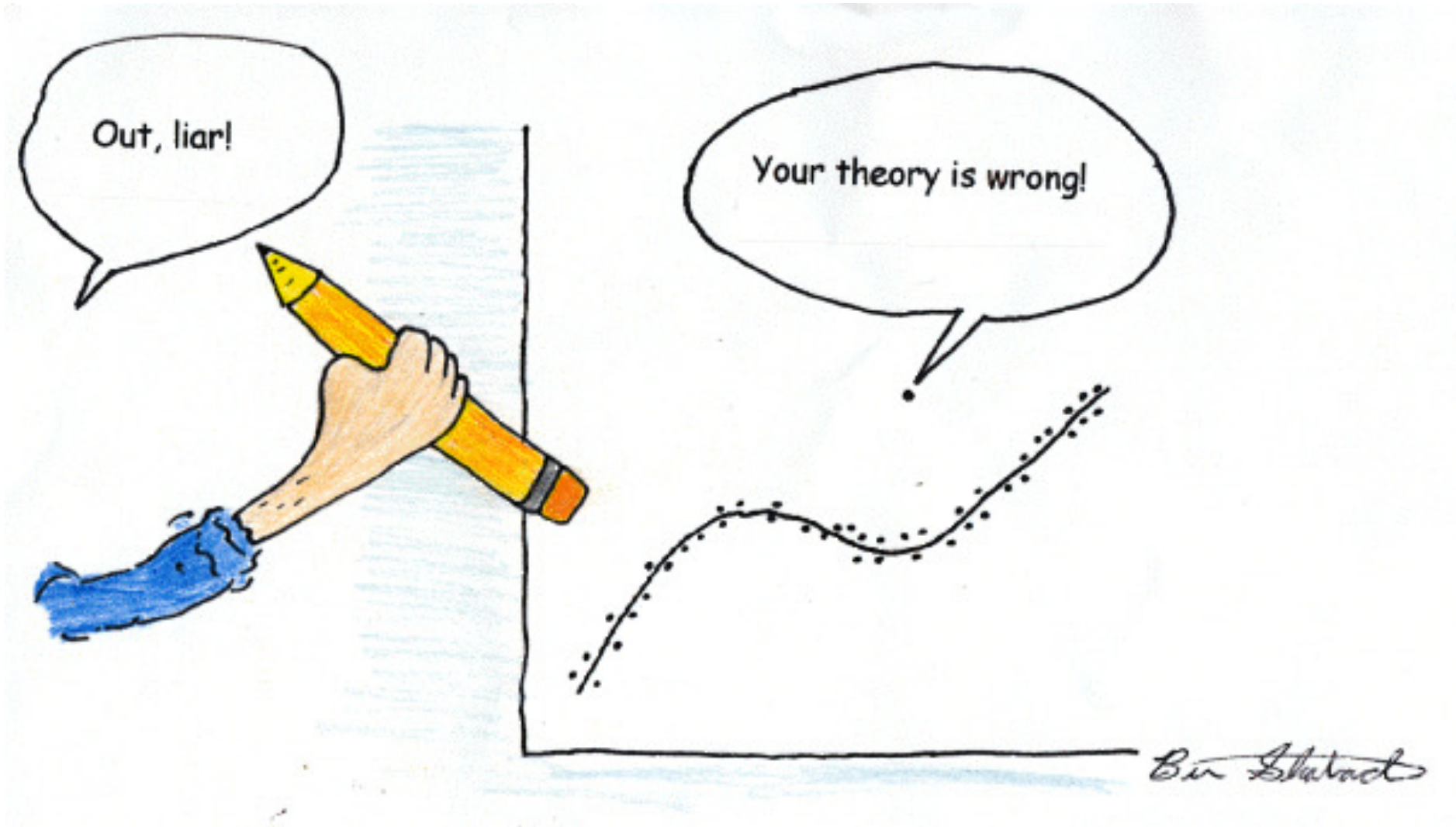
Outliers

An **outlier** is an observed value that is notably distinct from the other values in a dataset by being much smaller or larger than the rest of the data.



Outliers can potentially have a large influence on the statistics you calculate

- One should examine outliers in more detail to understand what is causing them


Out, liar!
Your theory is wrong!

Descriptive statistics for the center of a distribution

Graphs are useful for visualizing data to get a sense of what of what the data look like

We can also summarize data numerically

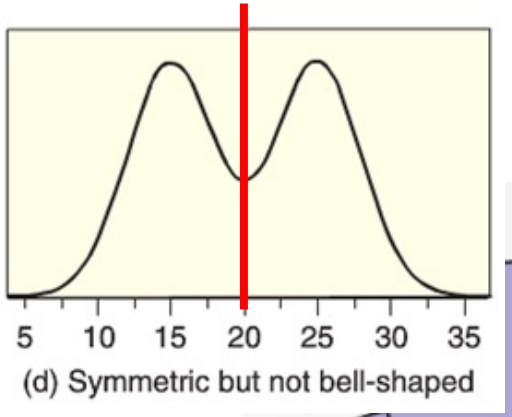
Question: what is a numerical summary of a sample of data called?

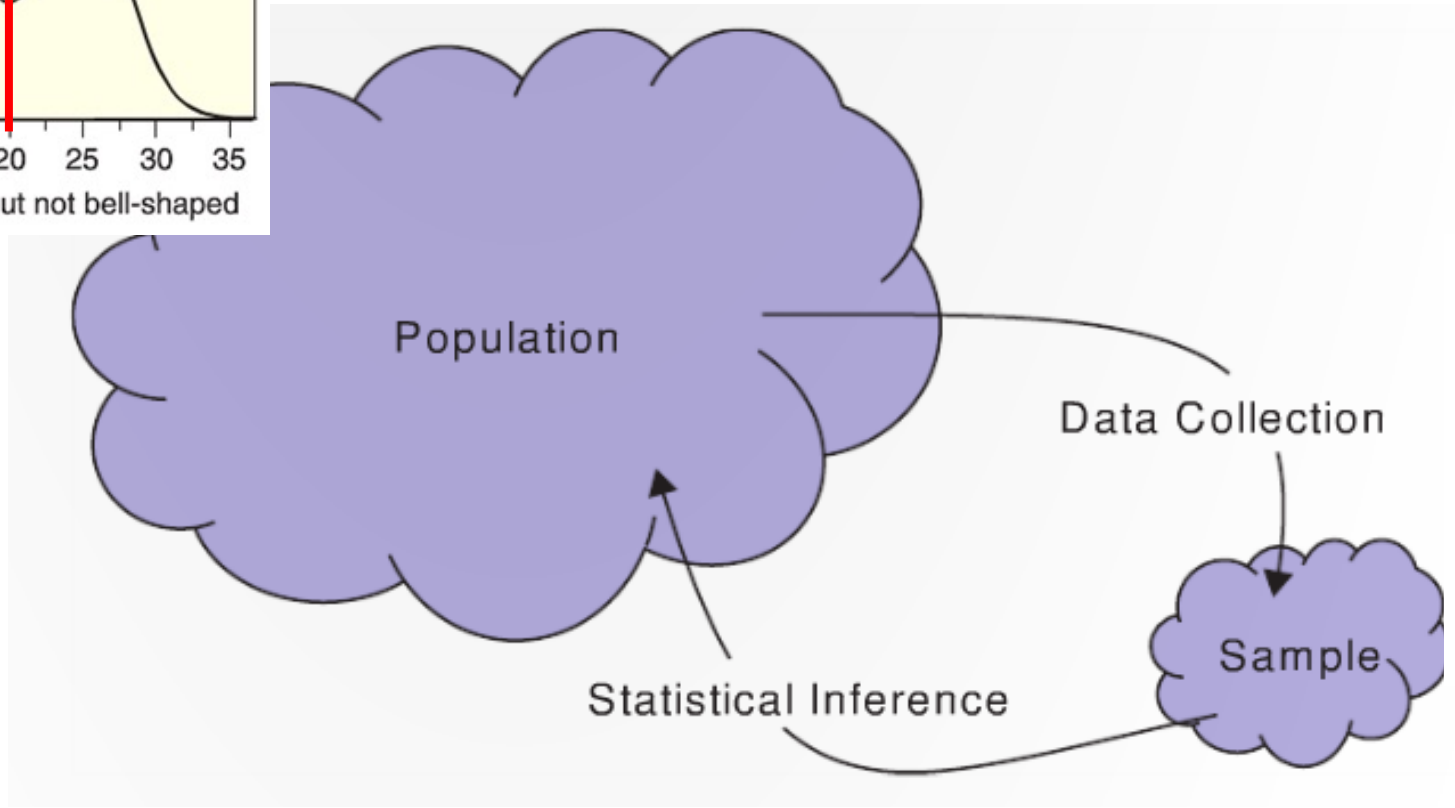
A: a statistic!

Two important statistics that can be used to describe the center of the data are the **mean** and the **median**

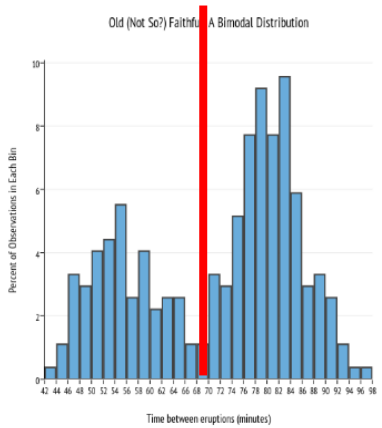
Sample and population mean

μ





$\bar{x}$



The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
R: mean(x)
R: mean(x, na.rm = TRUE)
```

Give the proper notation: μ vs. $\bar{x}$?

We measure the height of 50 randomly chosen Yale students

We measure the height of all Yale students

Can you calculate the mean of the countries life expectancy in R?

```
> life_expectancy <-  gapminder_2007$lifeExp
> mean(life_expectancy)
```

The median

The **median** of a data set of size n is

- If n is odd: The middle value of the sorted data
- If n is even: The average of the middle two values of the sorted data

The median splits the data in half

R: `median(v)`
`median(v, na.rm = TRUE)`

Resistance

We say that a statistics is **resistant** if it is relatively unaffected by extreme values (outliers).

The median is resistant when the mean is not

Example:

Mean US salary = $72,641

Median US salary = $51,939

Summary of concepts

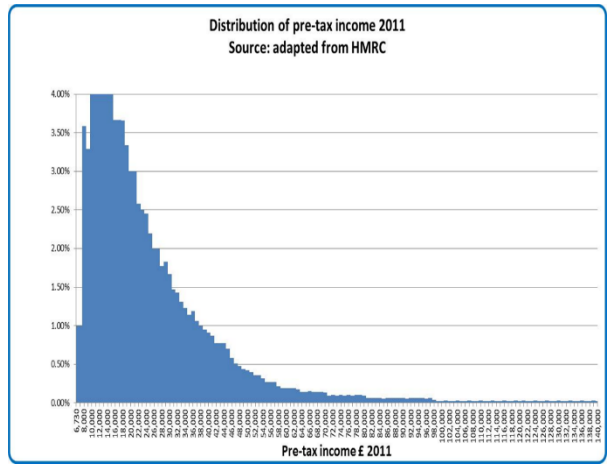
1. A ***probability distribution*** shows the ***relative likelihood*** that we will get a data point in the population with a particular value
    - (for a more precise definition take a class in probability)

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

3. The **mean** is one measure of central tendency
    - Sample mean is denoted $\bar{x}$ (statistic)
    - Population mean is denoted μ (parameter)

4. The **median** is another measure of central tendency
    - The median is resistant to outliers while the mean is not

Income distribution



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)

We can calculate measures of central tendency using:

> mean(my_vec)
> median(my_vec)

Practice at home

Lock5 questions:

- Proportions
  - warmups: 2.1, 2.3, 2.5, 2.7, 2.9 (1st and 2nd edition)
  - 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)
  - 2.43, 2.45 (2nd edition 2.53, 2.55)
  - 2.47, 2.49 (2nd edition 2.57, 2.59)

Experiment with the Gapminder data frame and extended Biden approval ratings:

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

Homework 1

Homework 1 is due at 11pm on Sunday

Use Ed Discussions for any questions that come up, and/or attend class office hours

Upload a pdf with your answers to Gradescope