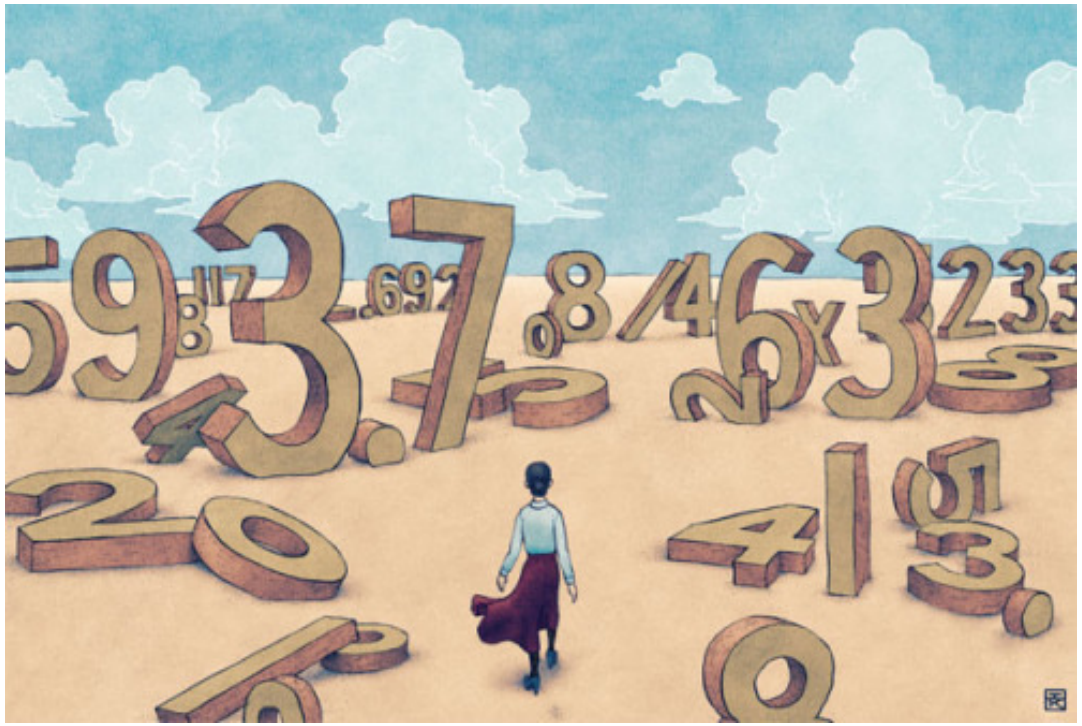
Quantitative data: shape and measures of central tendency



Overview

RMarkdown

Review of categorical data concepts and R

Quantitative data

Graphing the shape: histograms and outliers

Measures of the central tendency: mean and median

If there is time:

- Additional Lock5 practice questions/data analysis in R

Questions?


rmarkdown
www.rstudio.com

R Markdown

R Markdown (.Rmd files) documents allow you to combine written descriptions with R analysis code.

You can then ‘knit’ these documents to create nice looking report.

All homework in this class will be done using R Markdown.

R Markdown document structure

R Markdown documents have written sections and code sections.

Everything in R chunks is executed as code:

```
```{r}
    # this is a comment
    # the following code will be executed
    2 + 3
```
```

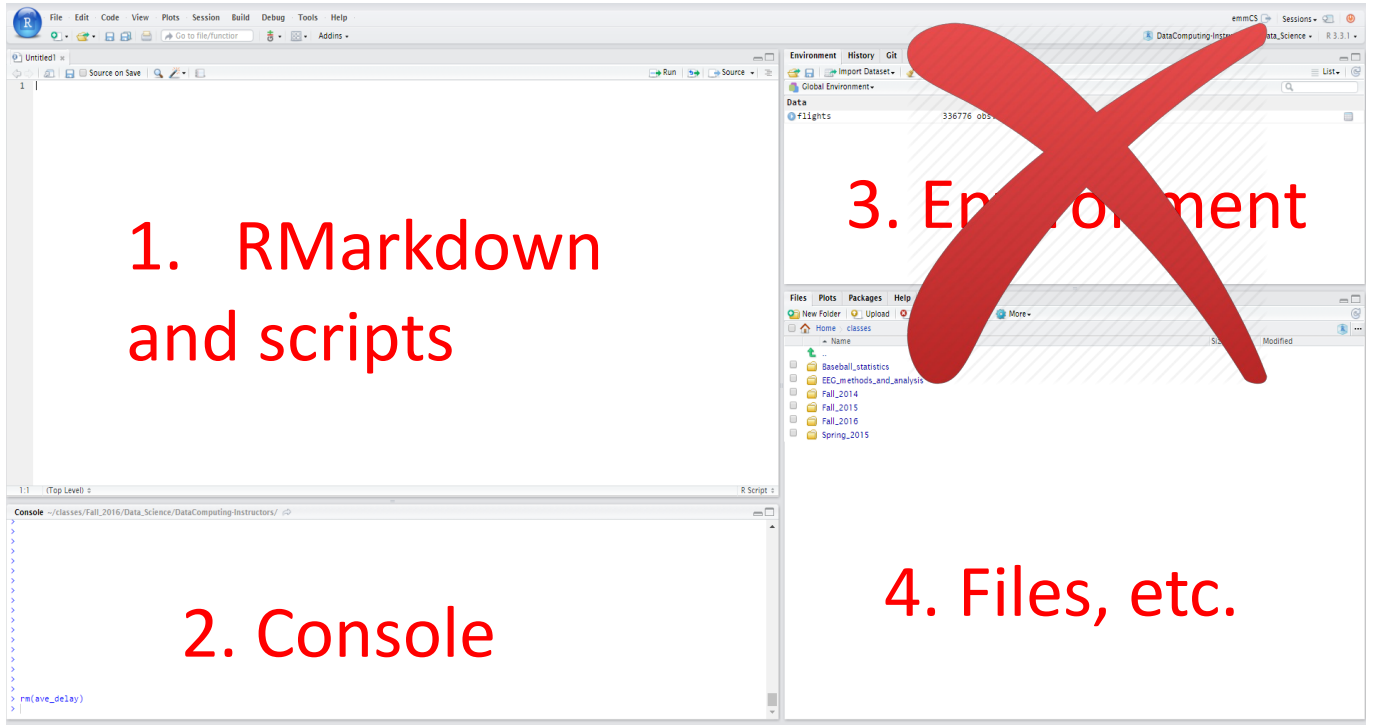
Everything outside R chunks appears as text.

R Markdown

Note: R Markdown documents **do not have access to variables in the global environment!**

Instead have their own environment.

Why is this a good thing???



R Markdown

Special LaTeX characters can be embedding in the text regions outside of the code chunks

Examples:

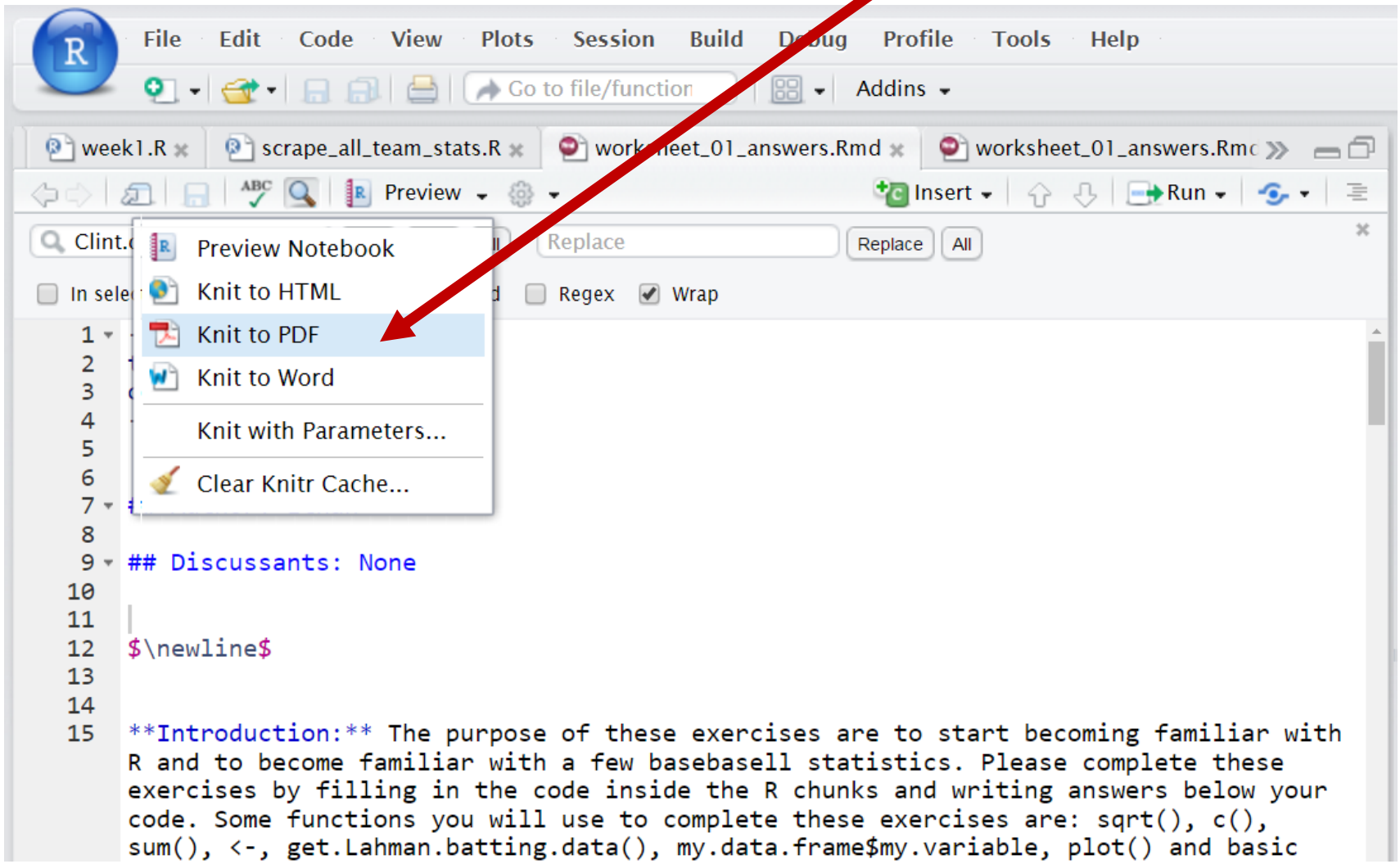
π

$\hat{p}$

$\hat{p}_{red}$

Knitting to a pdf

Turn in a pdf of your solutions to Gradescope



Avoid hard to debug code!

Only change a few lines at a time and then knit your document to make sure everything is working!

Comment out parts of the code that isn’t working (using the # symbol) until you can find the line of code that is giving the error message

Homework 1

Homework 1 is due at 11pm on Sunday January 28th

To get the homework use:

```
> SDS100::download_homework(1)
```

Run the first chunk at the top of the homework to get the needed data

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

Upload a pdf with your answers to Gradescope

Quick Review

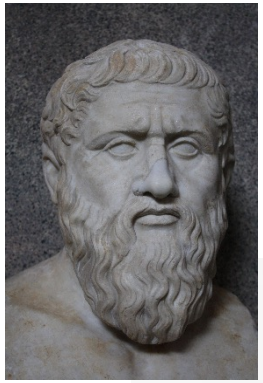
Categorical variables

Art show!

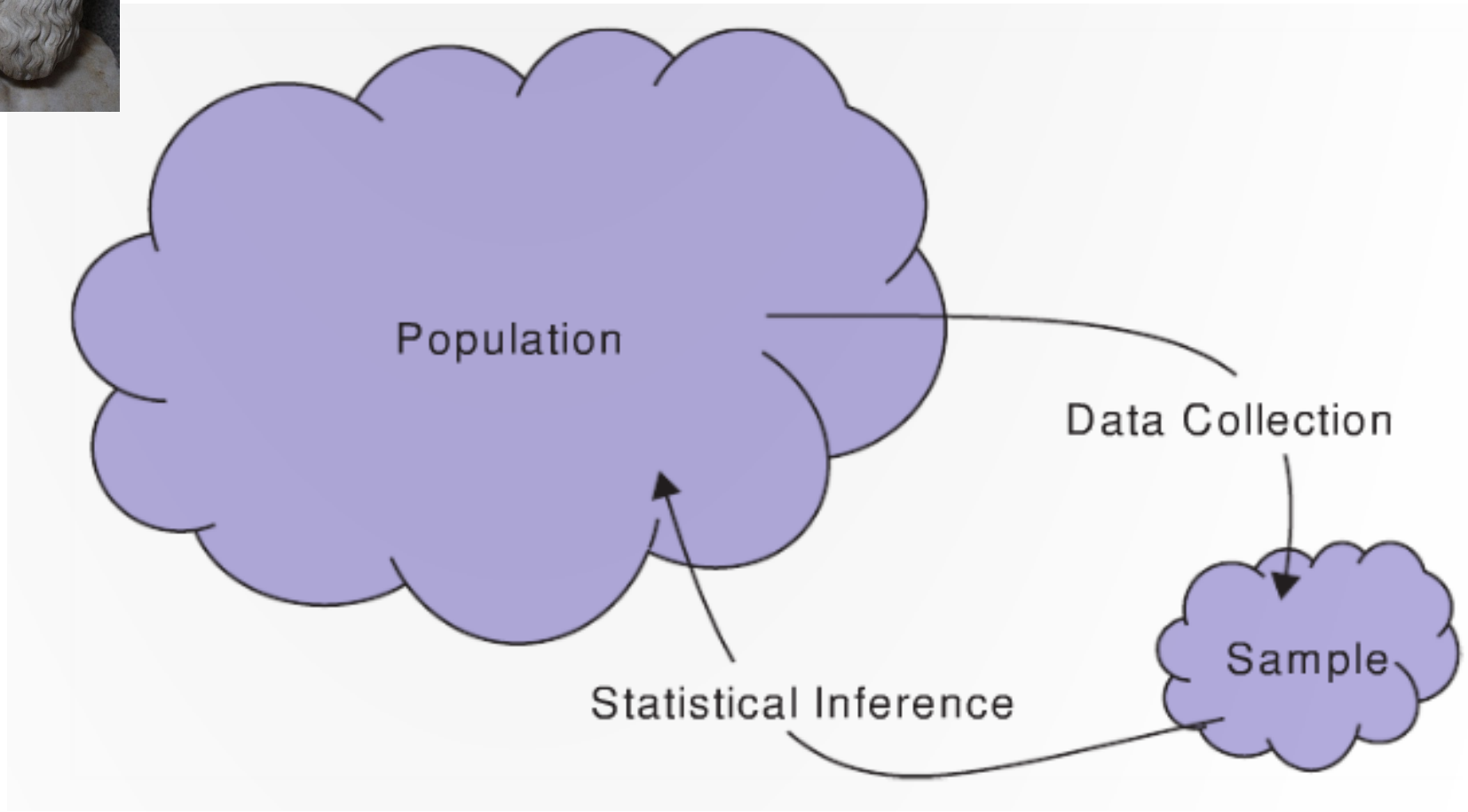
Let's see your drawings of:

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

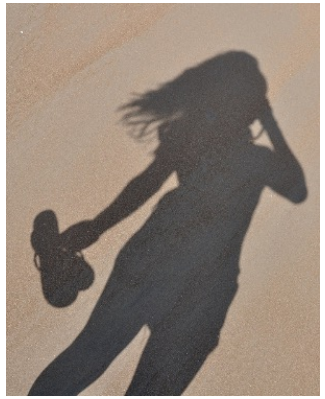
What was the best drawing?



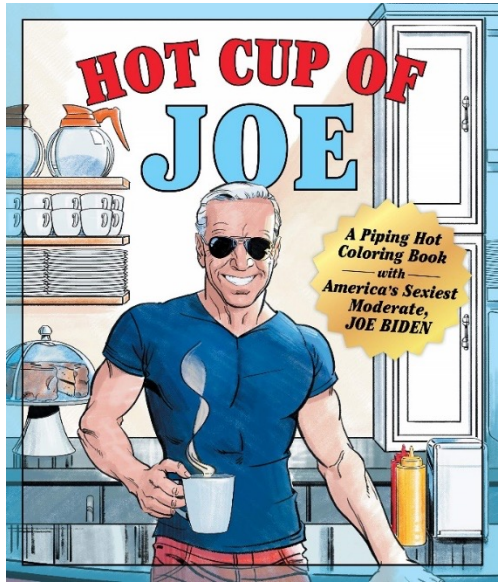
parameter: π



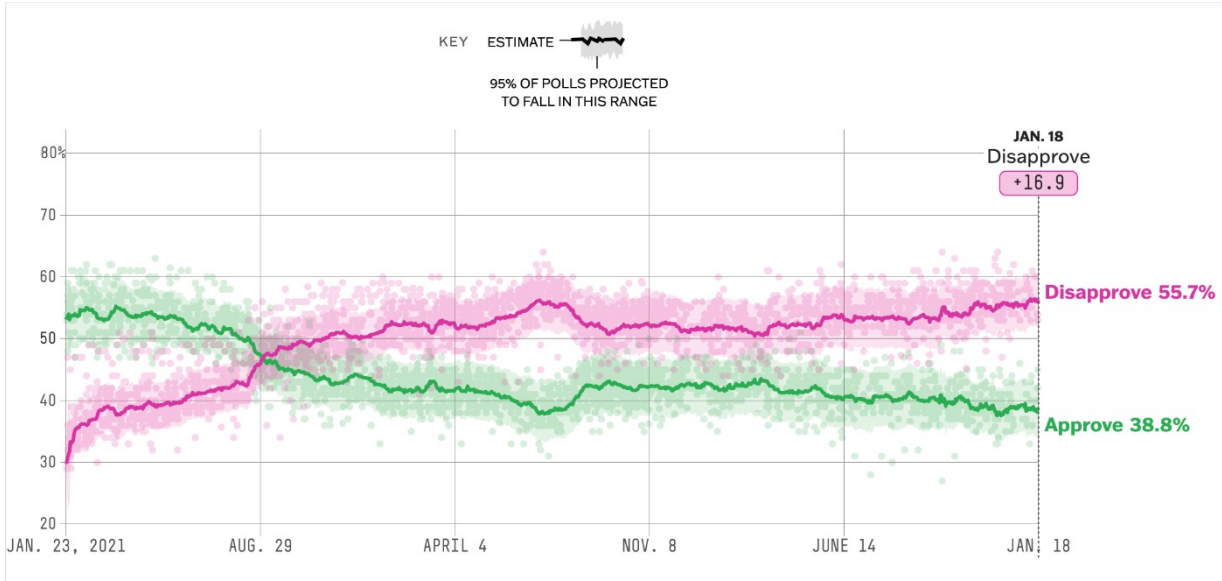
statistic: $\hat{p}$



Example: Biden's approval rating



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



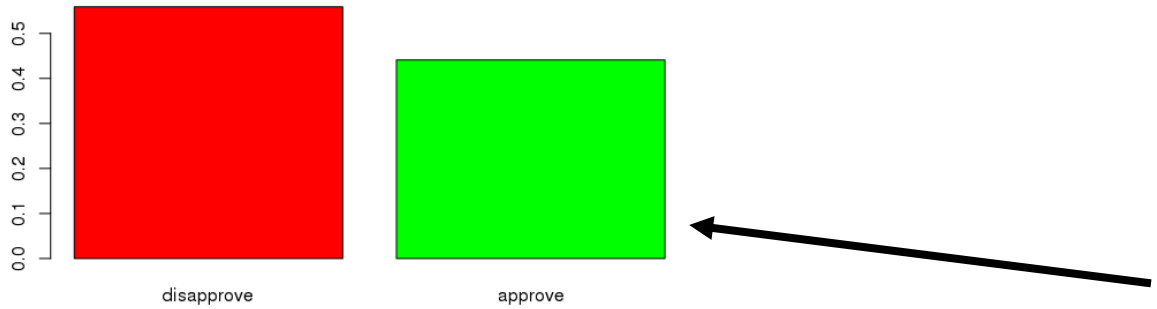

```
# get Biden's approval rating from 1,000 simulated voters
> library(SDS100)
> approval_sample <- get_approval_sample(1000)
```

Example: Biden's approval rating

```
> approval_sample <- get_approval_sample(1000)
```

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

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



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

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

Questions?

Quantitative variables

Descriptive statistics for one quantitative variable

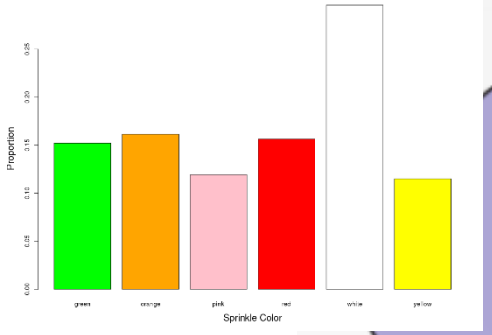
We will be looking at:

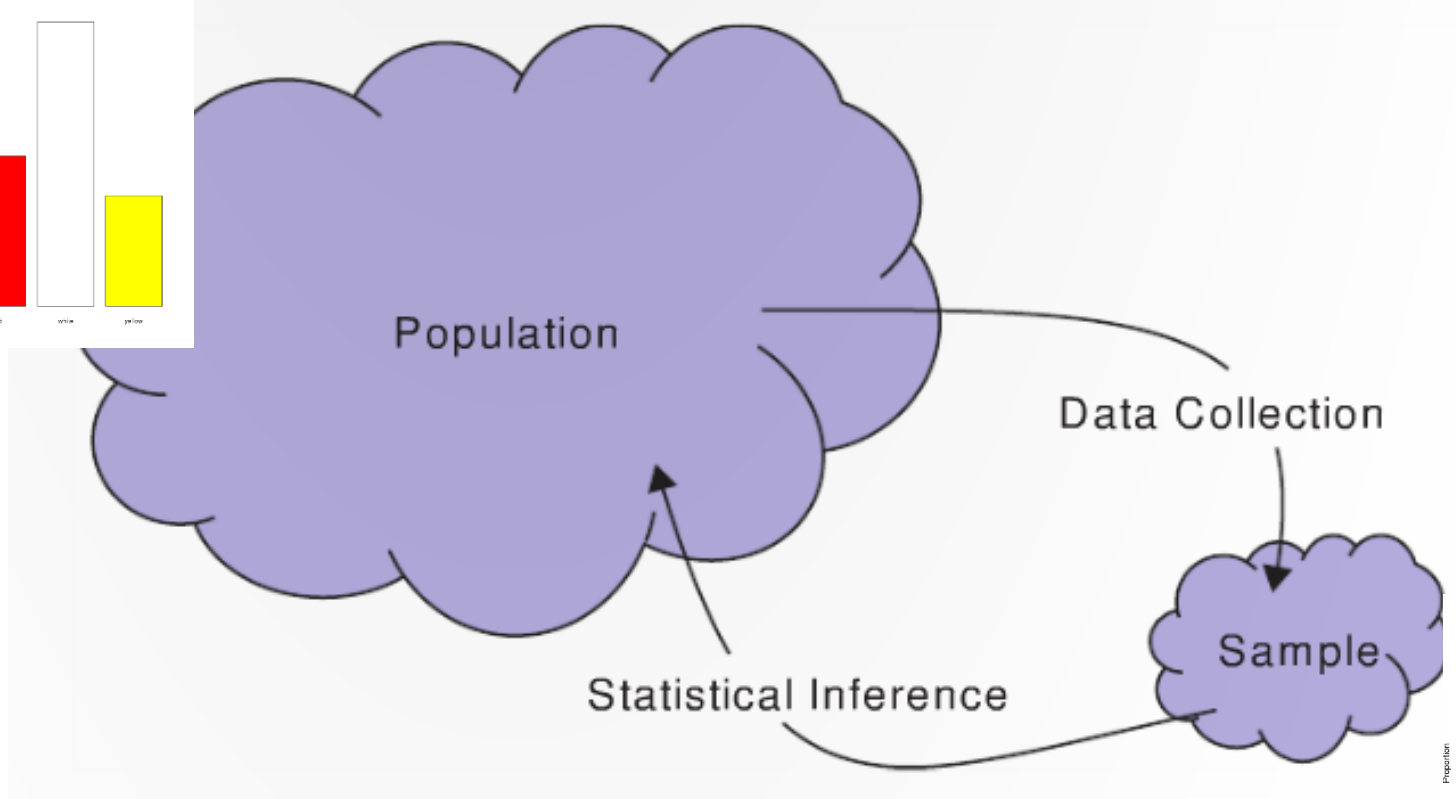
- What is the general ‘shape’ of the data
- Where are the values centered
- How do the data vary

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

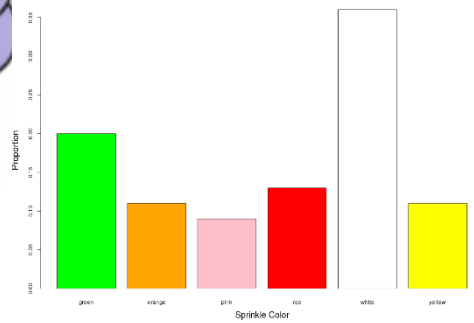
For categorical data we had…

Categorical Distribution (π)

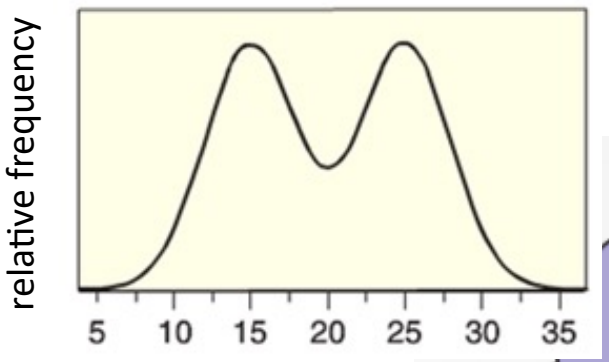


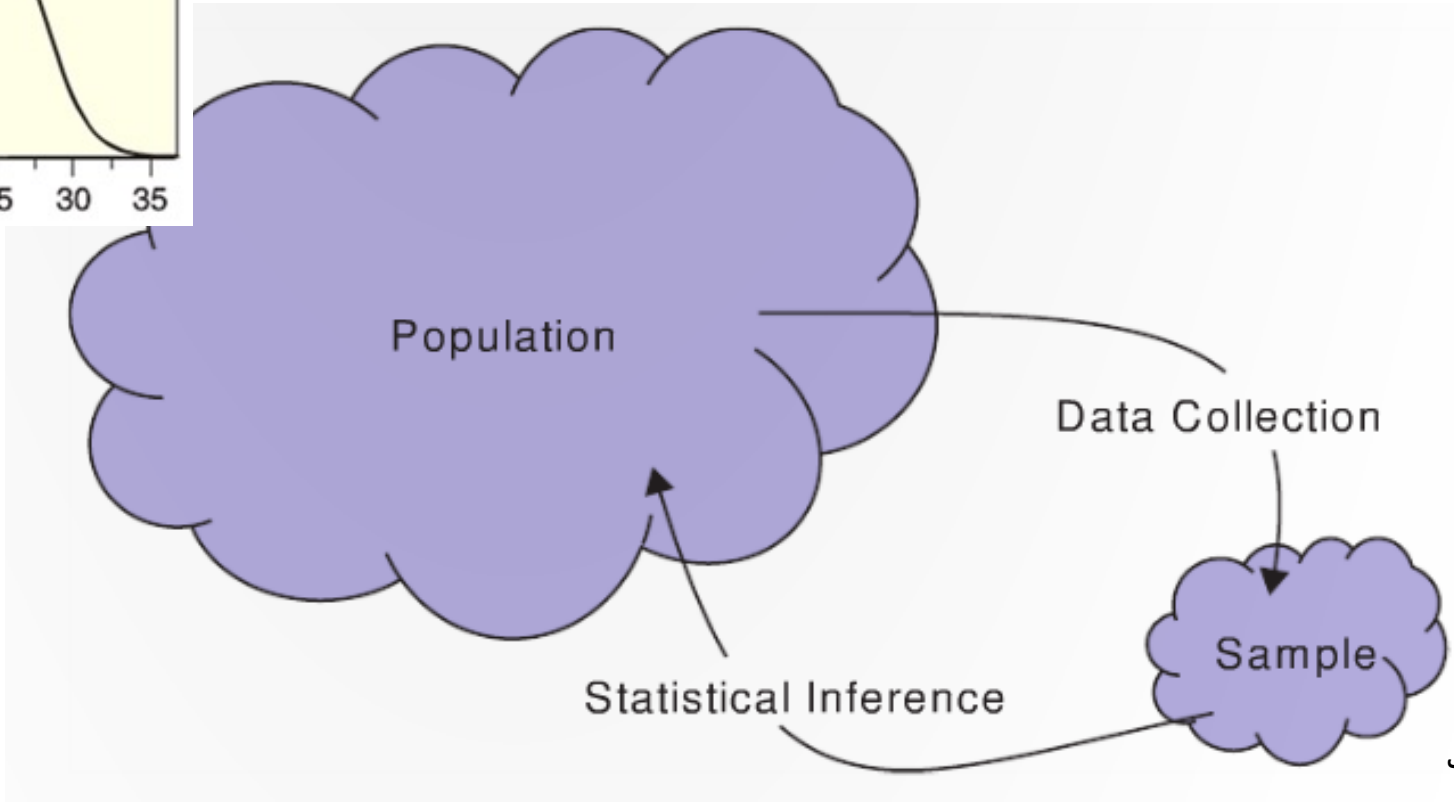


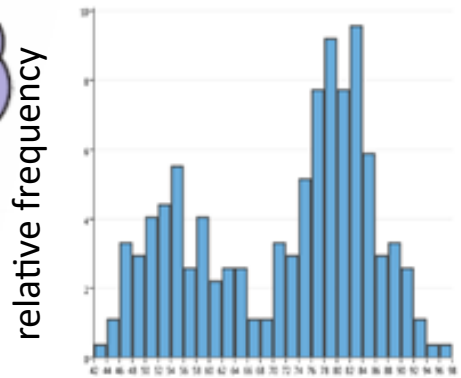
Bar chart ($\hat{p}$)



Population distributions and sample histograms





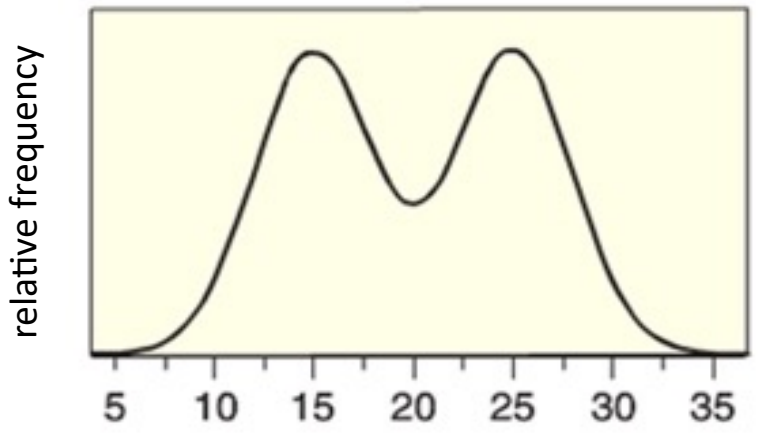


Histograms

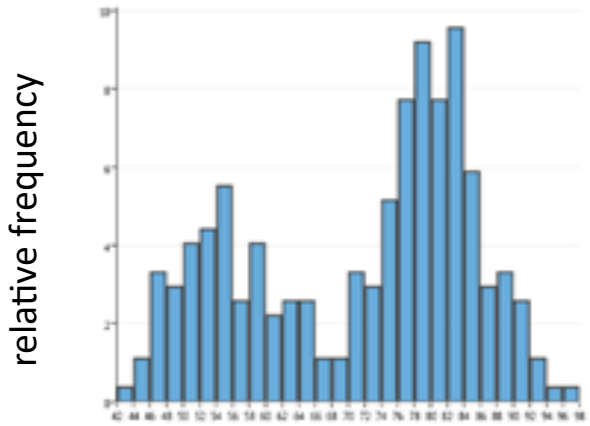
Histograms are a way of visualizing a sample of quantitative data

- 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



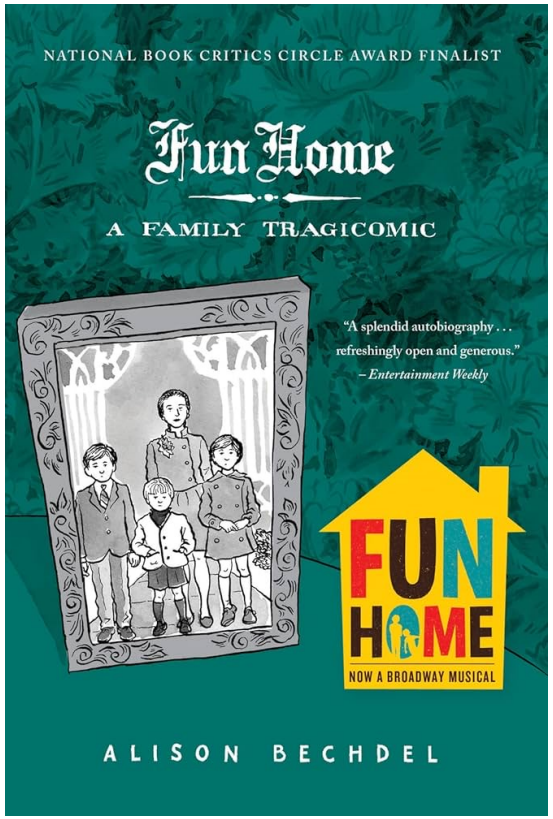
Bechdel data


```
> library("fivethirtyeight")
> View(bechdel)
```

year	title	binary	budget_2013	domgross_2013	intgross_2013
2013	21 & Over	FAIL	13000000	25682380	42195766
2012	Dredd 3D	PASS	45658735	13611086	41467257
2013	12 Years a Slave	FAIL	20000000	53107035	158607035
2013	2 Guns	FAIL	61000000	75612460	132493015
2013	42	FAIL	40000000	95020213	95020213
2013	47 Ronin	FAIL	225000000	38362475	145803842
2013	A Good Day to Die Hard	FAIL	92000000	67349198	304249198

adjusted for inflation to 2013 dollars

Side note: fun home



Bechdel data

year	title	binary	budget_2013	domgross_2013	intgross_2013
2013	21 & Over	FAIL	13000000	25682380	42195766
2012	Dredd 3D	PASS	45658735	13611086	41467257
2013	12 Years a Slave	FAIL	20000000	53107035	158607035
2013	2 Guns	FAIL	61000000	75612460	132493015
2013	42	FAIL	40000000	95020213	95020213
2013	47 Ronin	FAIL	225000000	38362475	145803842
2013	A Good Day to Die Hard	FAIL	92000000	67349198	304249198

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

```
# whether a movie passed is a categorical variable
> pass_data <- bechdel$binary
> pass_table <- table(pass_data)
> barplot(pass_table)
```

Bechdel data

Let's look at the gross domestic earnings of movies (in 2013 inflation adjusted dollars), which is a quantitative variable

```
# pull a vector of gross domestic earnings
    > domgross <-  bechdel$domgross_2013
```

Histograms

In histograms we:

1. Create a number of interval ranges (bins)

2. We count the number of points that fall in each interval

3. We create a bar chart with the counts in each bin

In R:

```
> hist(domgross, breaks = 100)
```

Histograms – movie revenue

Histogram of movie revenue in 10's of million $:

- 43.83, 72.30, 76.42, 42.73, …

To create a histogram we create a set of intervals

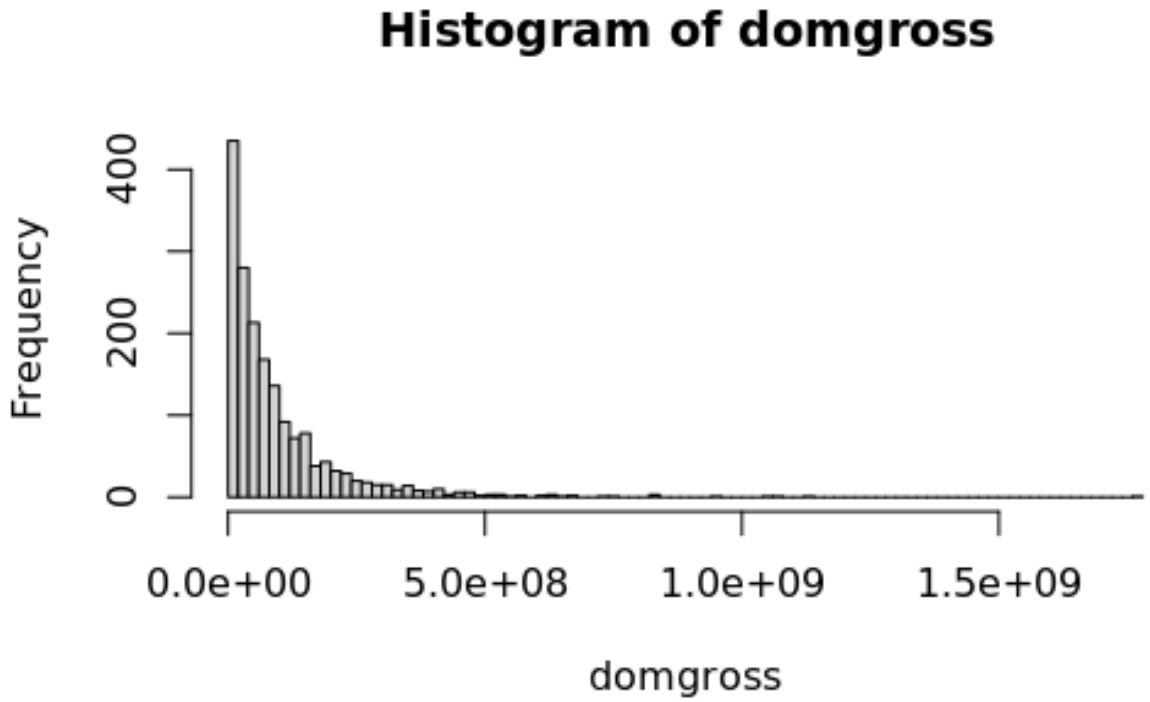
- 35-40, 40-45, 45-50, … 75-80, 80-85

We count the number of points that fall in each interval

We create a bar chart with the counts in each bin

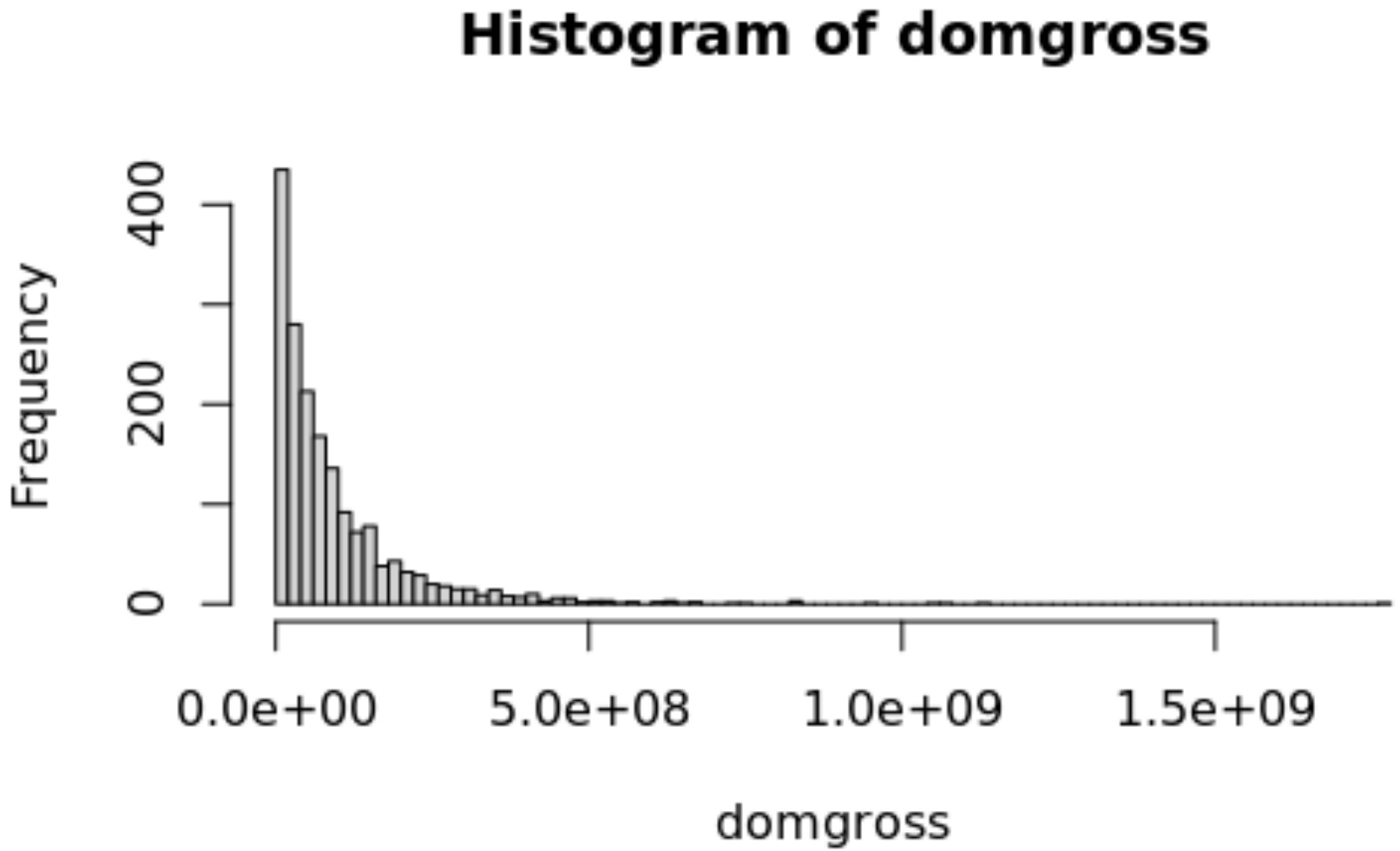
Histograms – movie revenue

Movie revenue	Frequency Count
(0 – 20]	435
(20-40]	144
(40 – 60]	136
(60 – 80]	107
(80 – 100]	106
(120– 140]	94
(140 – 160]	74
(160 – 180]	72
(180 – 200]	63
(200 – 220]	12

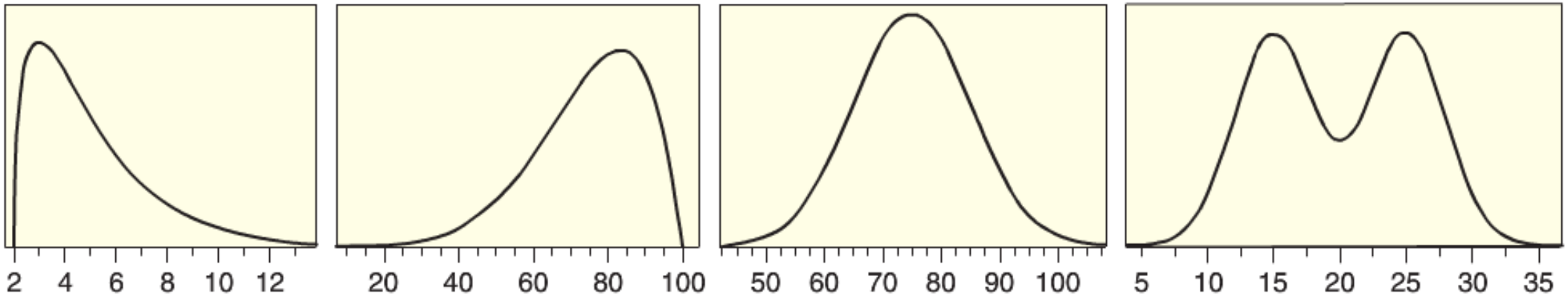


R: `hist(v)`

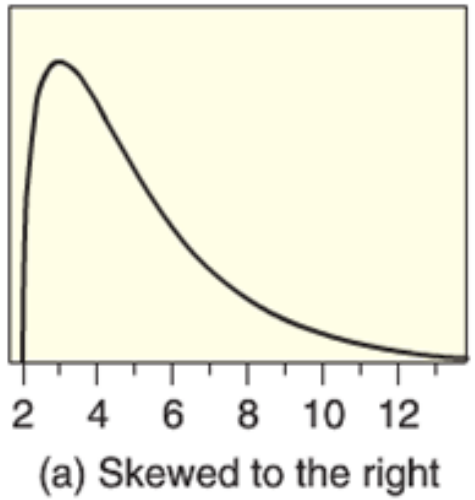
Histograms


Histogram of domgross
Frequency
0
200
400
0.0e+00
5.0e+08
1.0e+09
1.5e+09
domgross

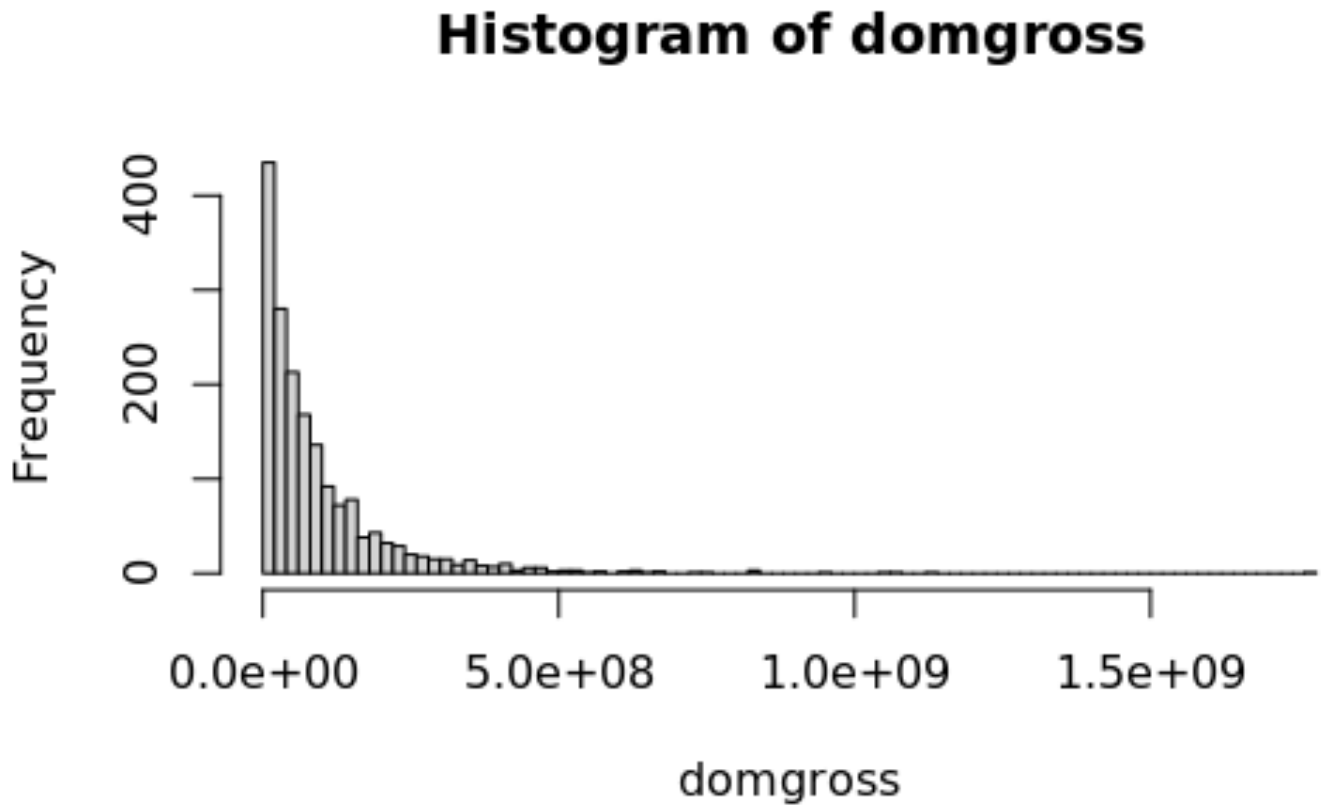
Common shapes for distributions



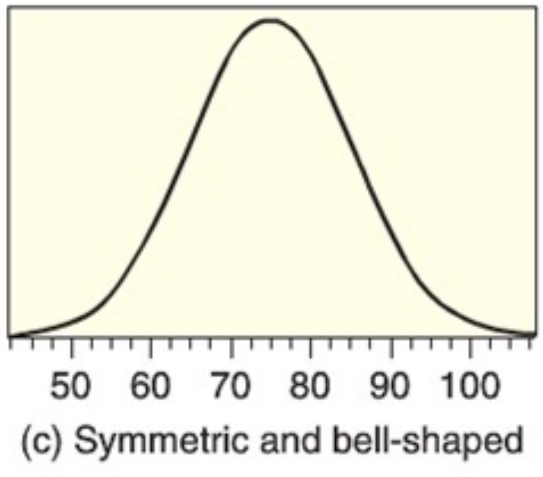
Can you think of a distribution that is right skewed?



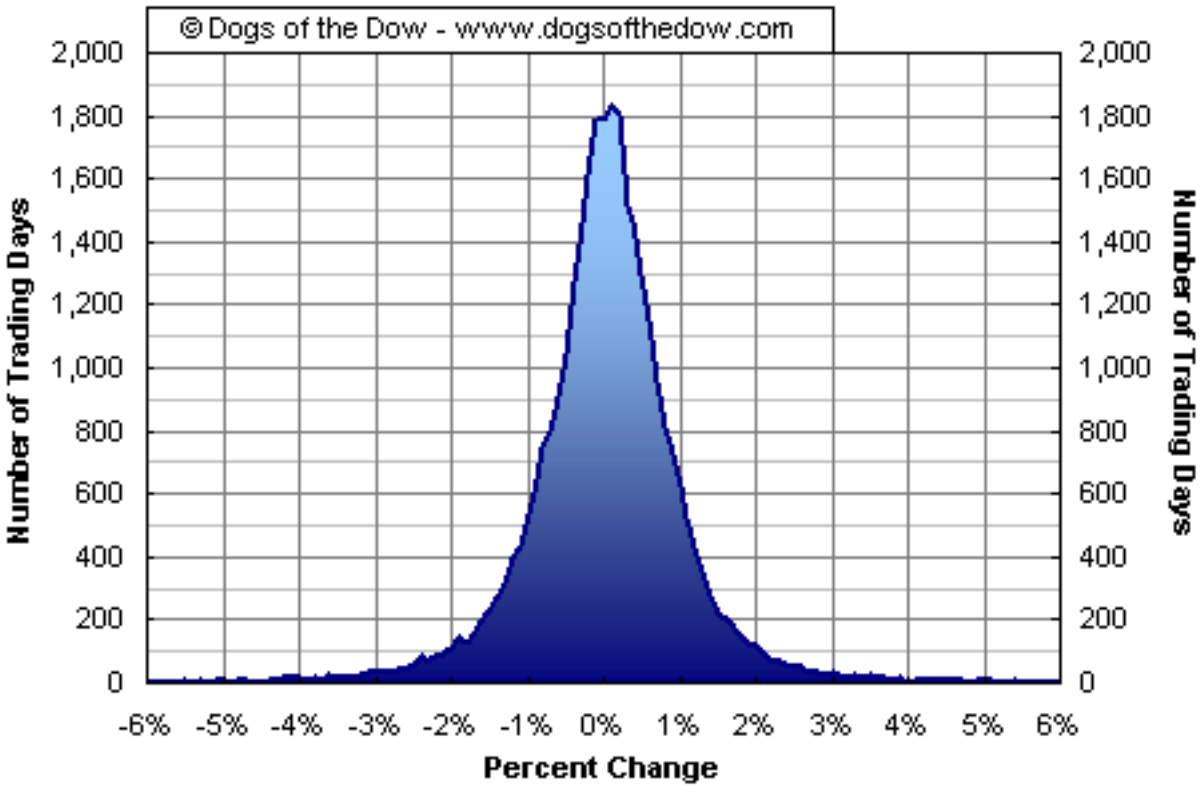
(a) Skewed to the right



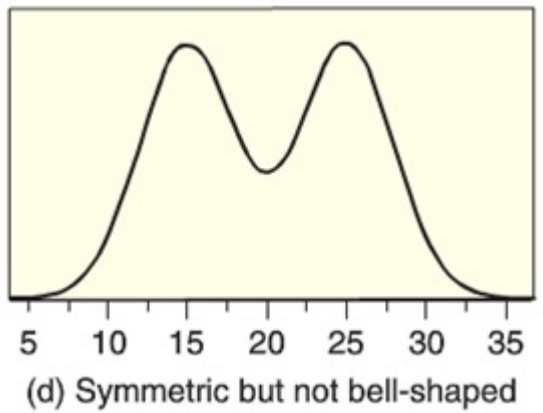
Can you think of a distribution that is symmetric and bell-shaped?



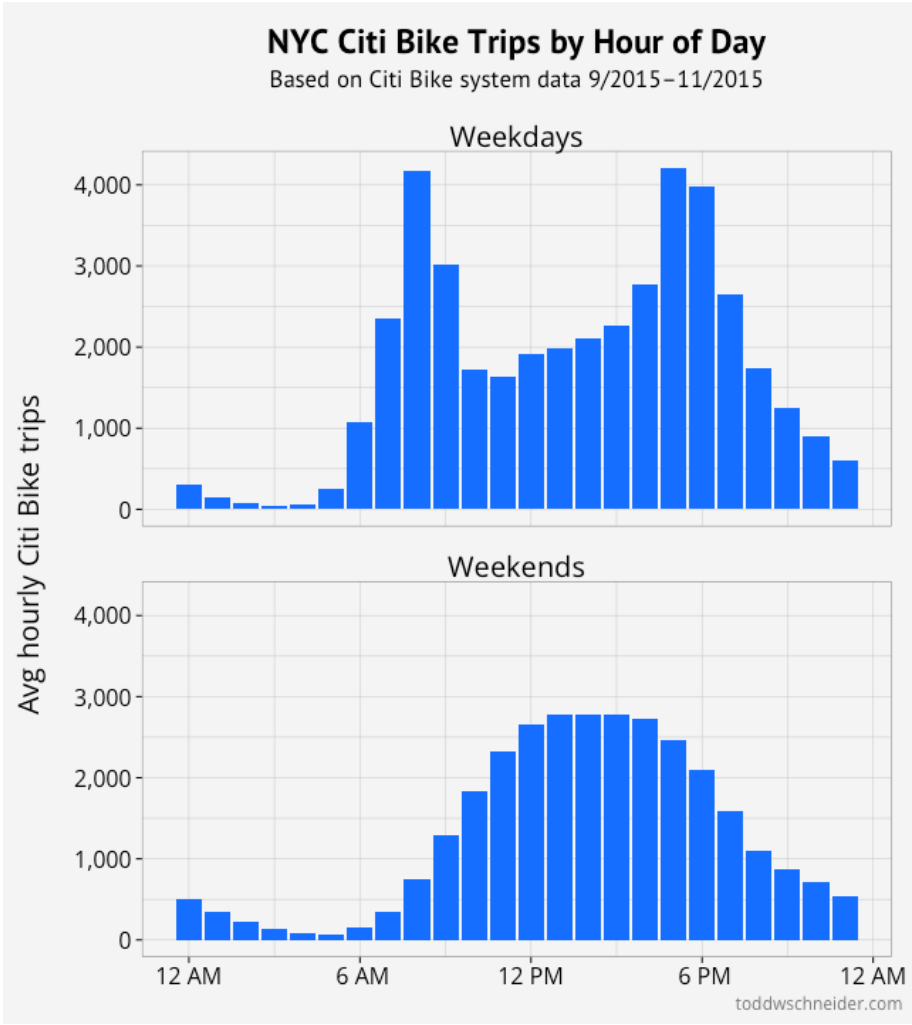
(c) Symmetric and bell-shaped



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

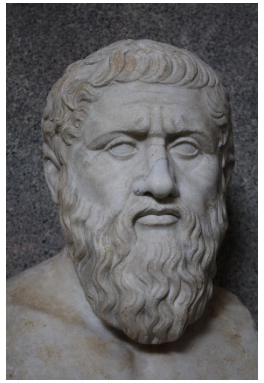


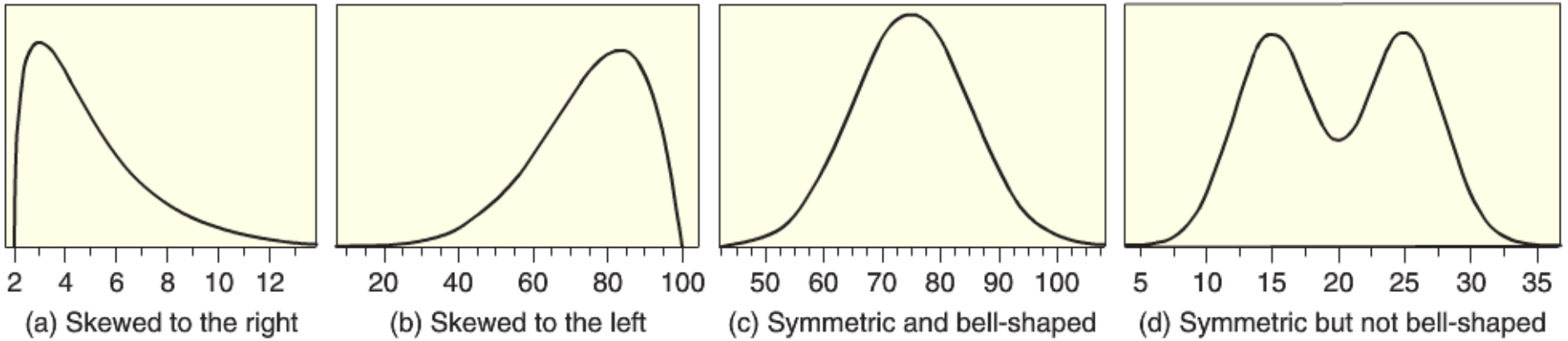
(d) Symmetric but not bell-shaped

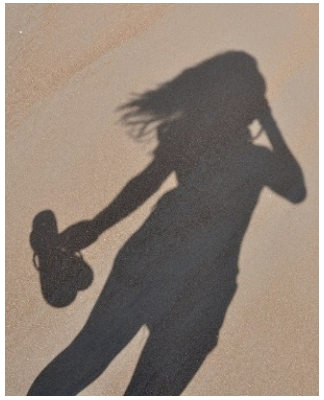


[Citi bike analysis]

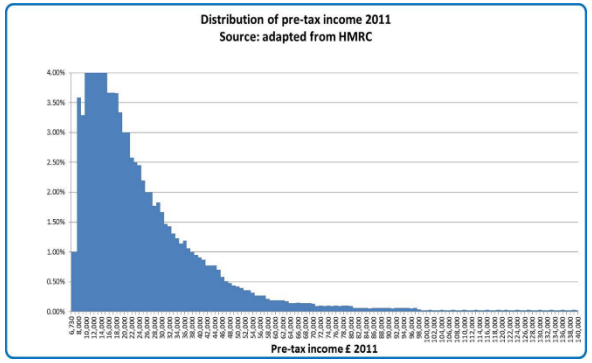
Plato and shadows: distributions and histograms

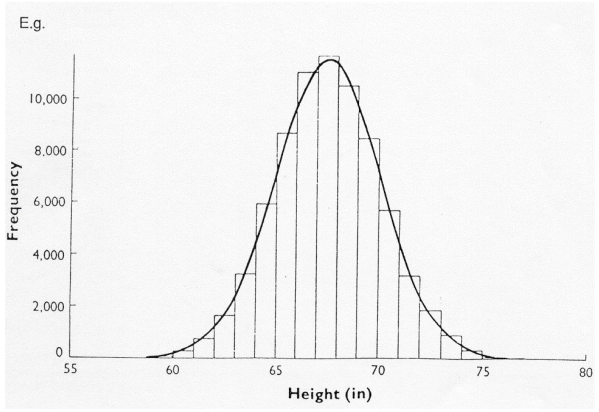


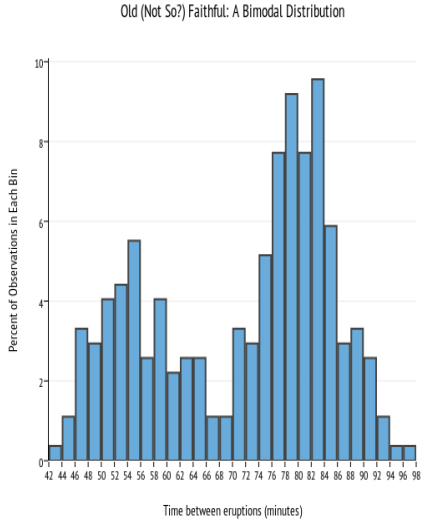




Income distribution

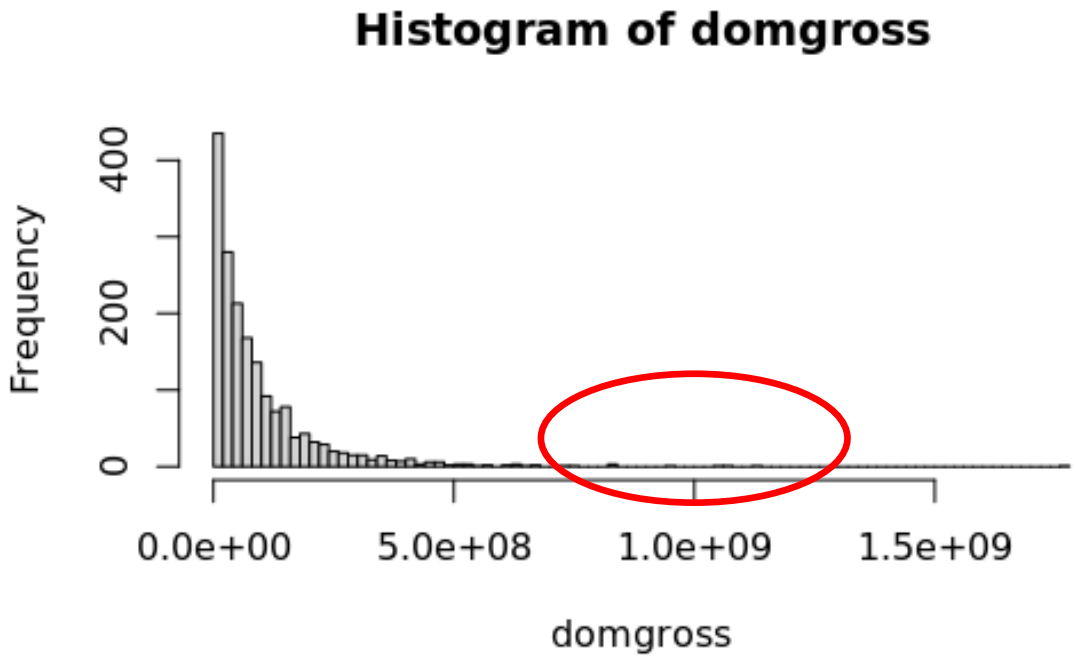






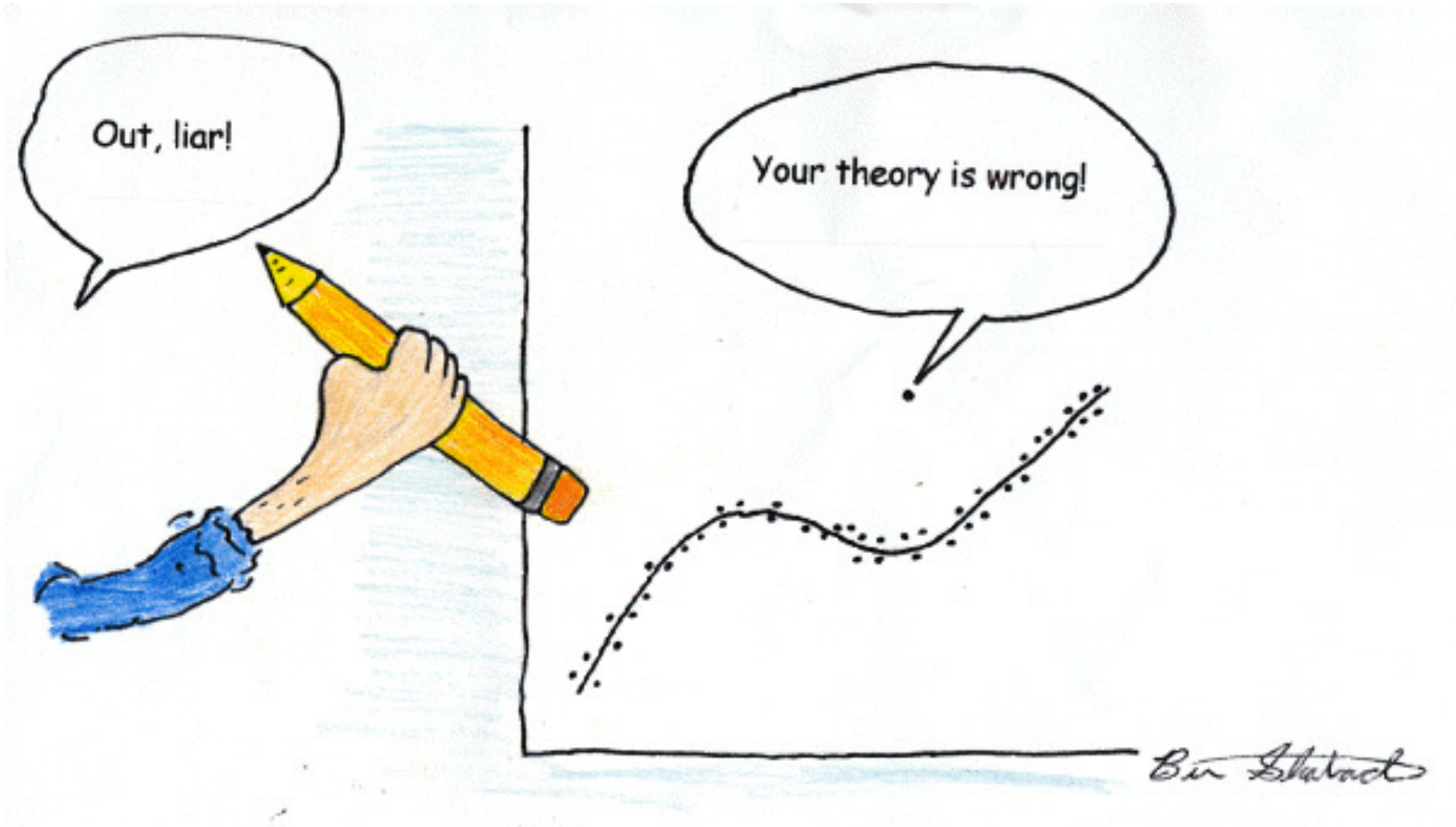
Outliers

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!

Questions?

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

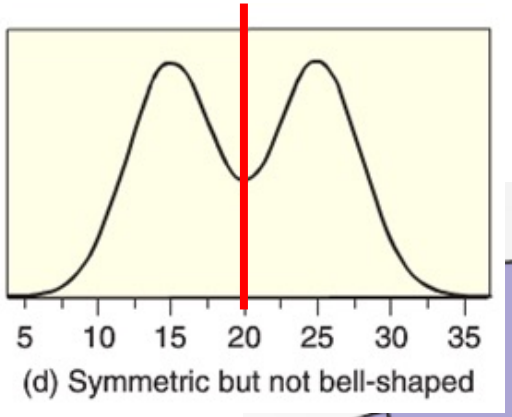
Question: what is a numerical summary of a sample of data called?

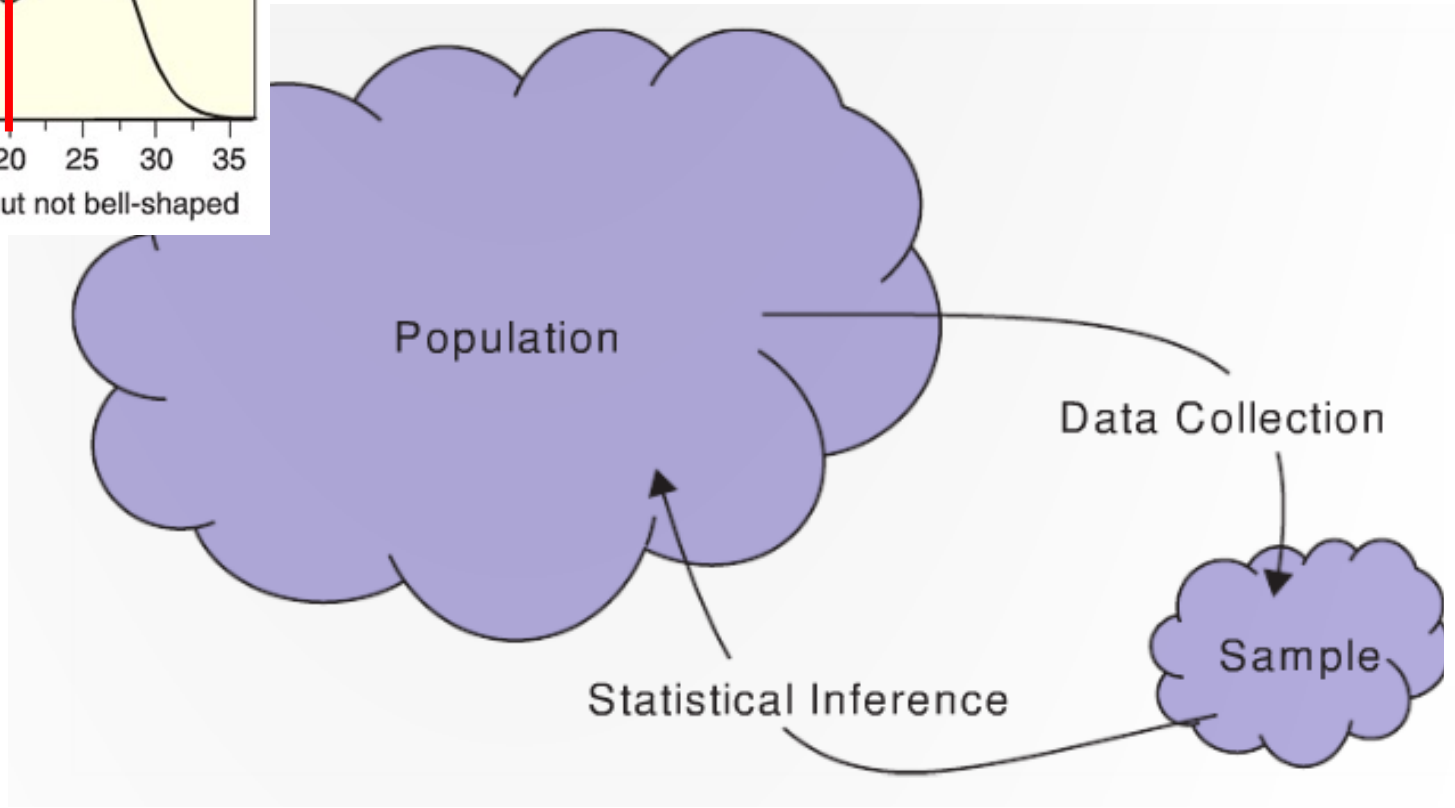
A: a statistic!

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

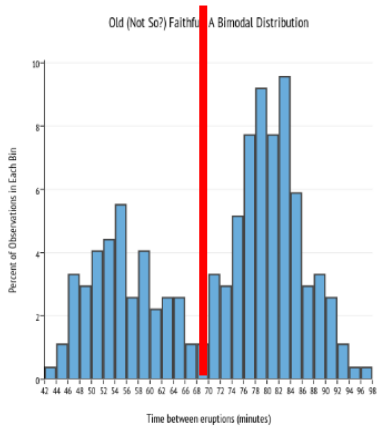
Sample and population mean

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. $\bar{x}$?

The mean body temperature of humans?

The mean amout of rainfall in January in San Francisco?

Can you calculate the mean gross domestic earnings of movies in R?

> mean(domgross)

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

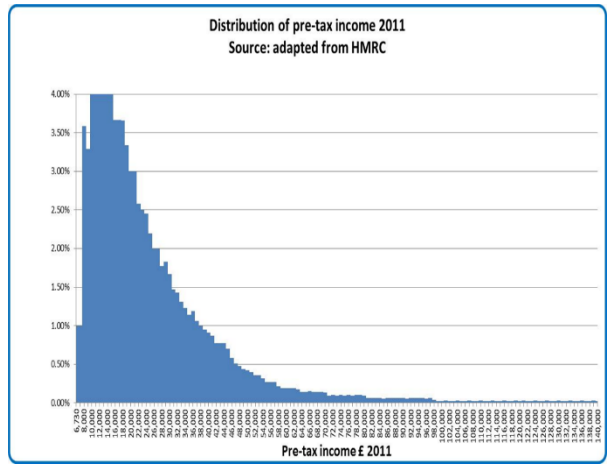
1. A ***probability distribution*** shows the ***relative likelihood*** that we will get a data point in the population with a particular value
    - (for a more precise definition take a class in probability)

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)

Homework 1

Homework 1 is due at 11pm on Sunday January 28th

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

Upload a pdf with your answers to Gradescope

Bonus practice questions

Practice questions

What is the mean and median of these values?

- 8, 12, 3, 18, 15
- 15, 22, 12, 28, 58, 18, 25, 18

Give the correct notation:

- The average number of calories eaten in one day is 2,386 calories for a sample of 100 participants
- The average number of tv sets owned per household for all households in the US is 2.6

Practice questions

Data about countries in the world can accesses in the Lock5Data package

- install.packages("Lock5Data")
- library(Lock5Data)
- View(AllCountries)

Create a histogram of life expectancies for all countries and…

- Describe the shape of the histogram
- From looking at the histogram, estimate the mean and median
  - Which will be larger?
- Check your answers using the mean() and median() functions

Practice questions

Let's look at the Lock5Data StudentSurvey data

- View(StudentSurvey)
- male_data <- subset(StudentSurvey, Sex == "M")$Exercise
- female_data <- subset(StudentSurvey, Sex == "F")$Exercise

From this data calculate:

- $\bar{x}_f$, the mean number of hours spent exercises by the females
- $\bar{x}_m$, the mean number of hours spent exercises by the males
- Compute the difference $\bar{x}_m - \bar{x}_f$, and interpret it in context