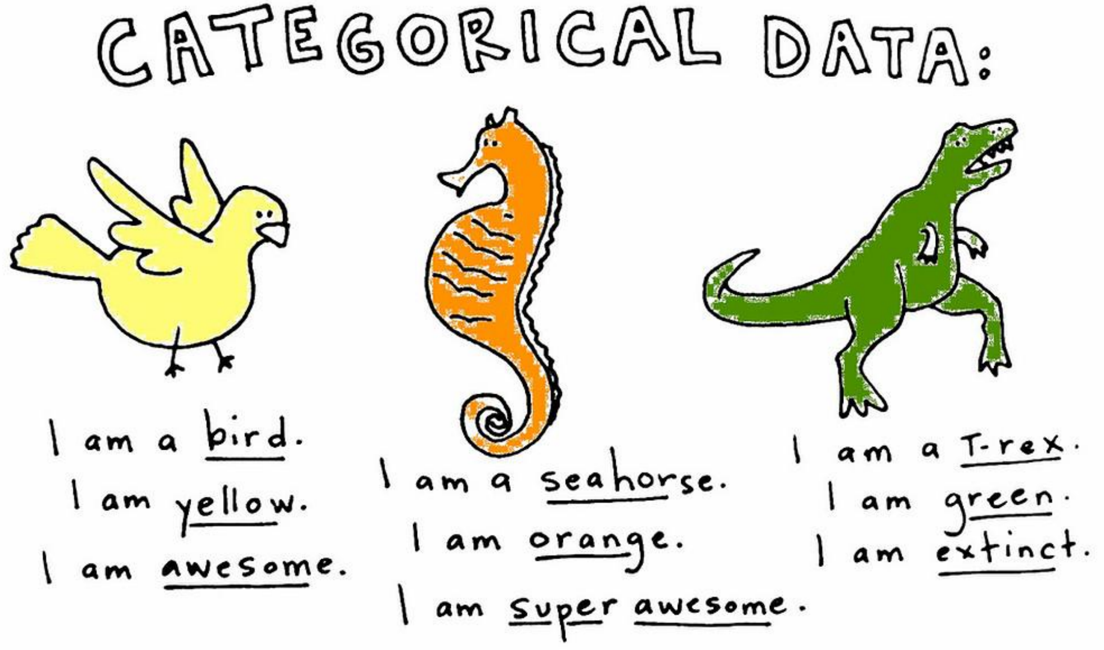
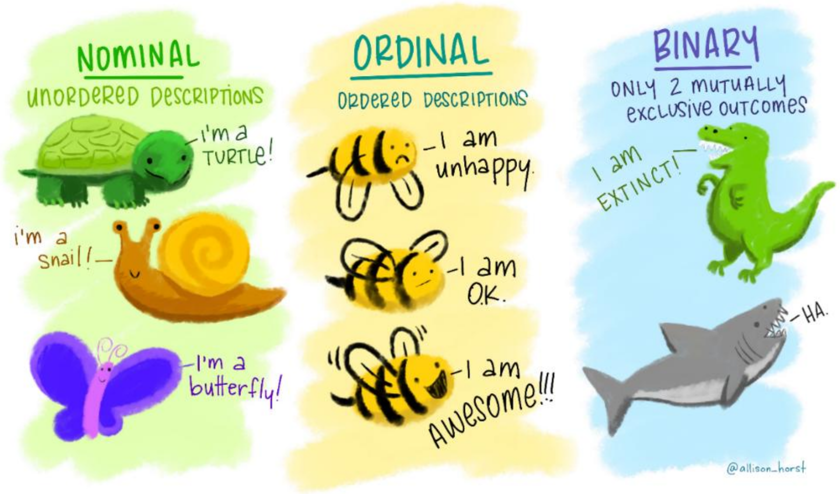
Categorical data continued and introduction to quantitative data analysis



Overview



Review of:

- Quarto documents and R
- Categorical data concepts and R

Brief discussion of analyzing two categorical variables

If there is time: quantitative data

Graphing the shape: histograms and outliers

Measures of the central tendency: mean and median

Announcement: homework 1

Homework 1 is due on Gradescope on Sunday, September 7th at 11pm

library(SDS1000)

goto_homework(1)

The TA office hours are on Canvas if you need help with the homework

Lynda's practice sessions

- Thursday: 3:00–5:00 PM
- Friday: 10:00 AM–12:00 PM

Announcement: homework 1

Instructions for how to submit homework on Gradescope are on Canvas

- Please mark all pages that answers correspond to on Gradescope!

Be sure to also "show your work" by printing out any values you report

- Although don't print out hundreds of access pages of numbers

Ask/answer questions on Ed Discussions, but don't give away the solutions!

Review: Quarto

Quarto

Quarto (.qmd files) allow you to embed written descriptions, R code and the output to create a reproducible research document!

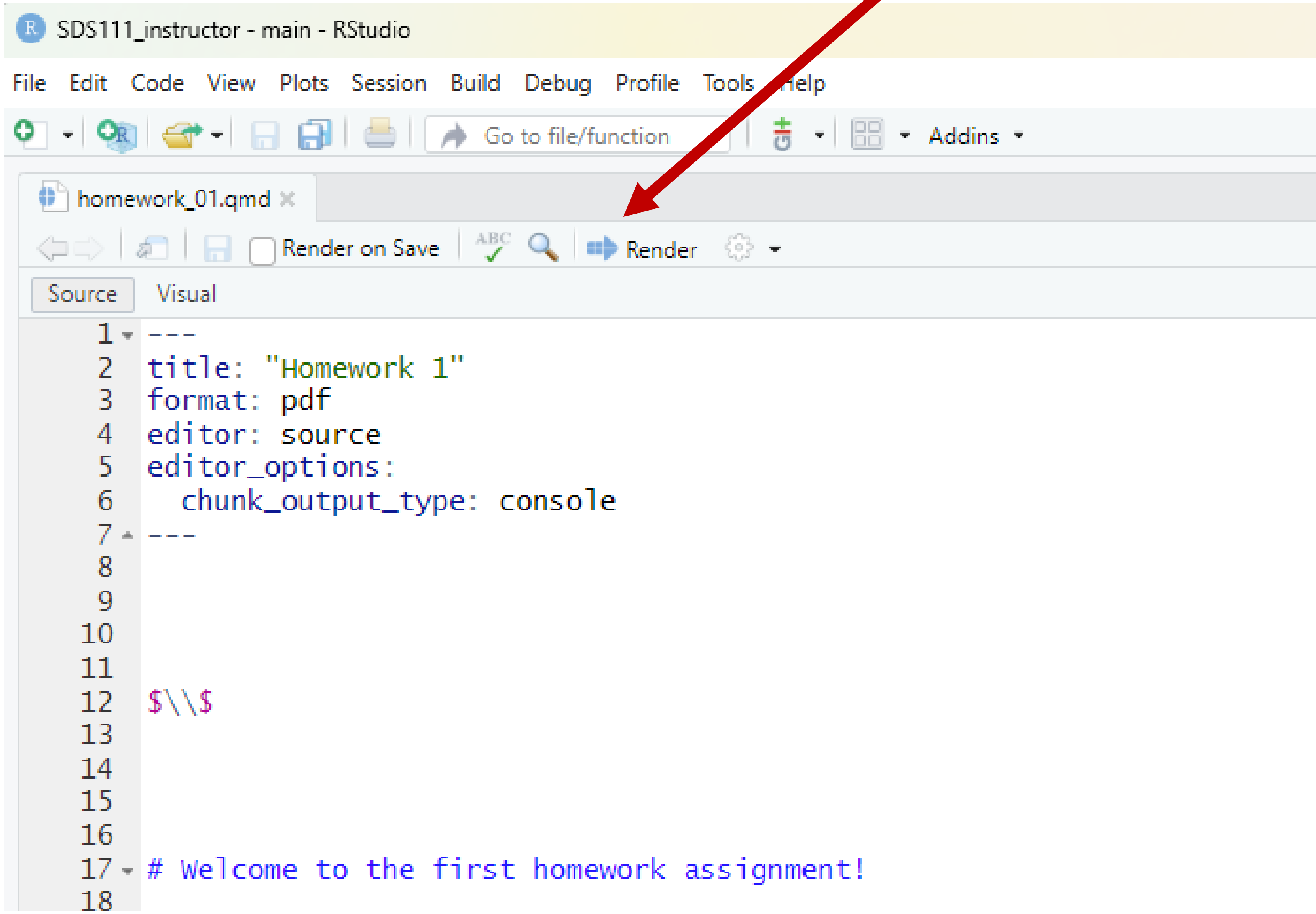
Everything in R chunks is executed as code:

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

Everything outside R chunks appears as text

Render to a pdf

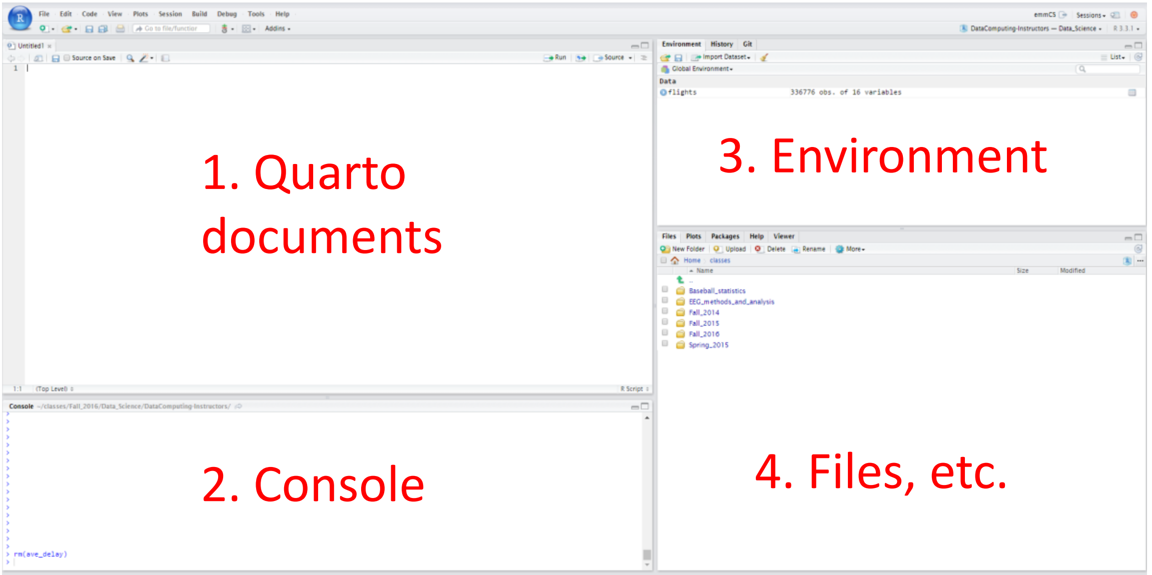
Renders to a pdf document
(which you will submit to Gradescope)



Quarto and the global environment

Note: When you render a Quarto document, your Quarto document does not have access to objects in the global environment

- i.e., it can't access any objects you created at the console



Why is this a good thing???

Takeway: All object you use in your Quarto document must be defined/created in the Quarto document

Formatting in Quarto

We can add formatting to text outside the code chunks

Examples:

	`## Level 2 header`	**Level 2 header**
	`**bold**`	**bold**
LaTeX	`$\pi$`	π
	`$x_{outcome}$`	$x_{outcome}$

To repeat: avoid hard to debug code!

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

If you document isn’t rendering:

- **For code chunks:** use the # symbol to comment out code until you can find the line of code that is giving the error message
- **Outside of code chunk:** cut out part of the document until it renders and then paste it back

Questions?

Quick review of R…

Review: R Basics

Arithmetic:

```
>   2 + 2
>   7 * 5
```

Assignment of values to ***objects***:

```
>  a <- 4
>  b <- 7
>  z  <- a + b
>  z
[1]  11
```

Review: Character strings and Booleans

```
> a <- 7
> s <- "s is a terrible name for an object"
> b <- TRUE

> class(a)
[1] numeric

> class(s)
[1] character
```

Review: Functions

Functions use parenthesis: functionName(x)

```
> sqrt(49)
> tolower("DATA is AWESOME!")
```

To get help

```
> ? sqrt
```

One can add comments to your code

```
> sqrt(49)    # this takes the square root of 49
```

Review: Vectors

Vectors are ordered sequences of numbers or letters

The c() function is used to create vectors

```
> v  <-  c(5, 232, 5, 543)
> s  <-  c("statistics", "data", "science", "fun")
```

One can access elements of a vector using square brackets []

```
> s[4]        # what will the answer be?
```

We can also apply functions to vectors

```
> length(v)      # this tells us how many elements there are in a vector
```

Data frames

Data frames contain structured data

Below is a data frame (from the homework) where people were asked their opinions about the Oxford comma

	respondent_id	care_oxford_comma	gender	age	household_income
1	3292953864	Some	Male	30-44	$50,000 - $99,999
2	3292950324	Not much	Male	30-44	$50,000 - $99,999
3	3292942669	Some	Male	30-44	NA
4	3292932796	Some	Male	18-29	NA
5	3292932522	Not much	NA	NA	NA

Data frames

Suppose our Oxford comma survey data was stored in an object called comma_survey

We can extract the columns of a data frame as vector objects using the $ symbol

```
gender <- comma_survey$gender
```

	respondent_id	care_oxford_comma	gender	age	household_income
1	3292953864	Some	Male	30-44	$50,000 - $99,999
2	3292950324	Not much	Male	30-44	$50,000 - $99,999
3	3292942669	Some	Male	30-44	NA
4	3292932796	Some	Male	18-29	NA
5	3292932522	Not much	NA	NA	NA

Questions?

Categorical variables

Motivation: The sprinkle business

ACME CORPORATION

PERFECT Corporation

ACME corporation believes that if they had the correct ratio (proportion) of red sprinkles that PERFECT corporation uses, their sales will increase

Where do samples/data come from?

To assess the proportion of sprinkles that PERFECT corporation uses, AMCE sampled 100 of PERFECT corporation's sprinkles

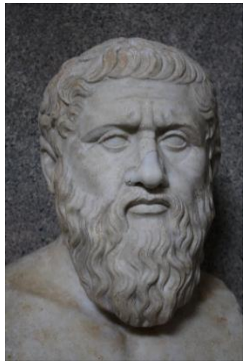
- The ***sample size*** is 100 (n = 100)

1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

Population parameters vs. sample statistics

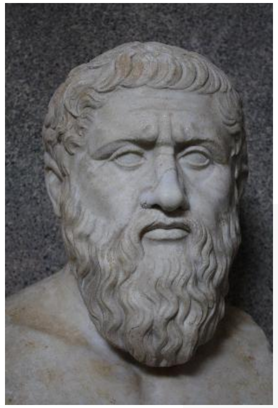
A **statistic** is a number that is computed from ***data in a sample***

A **parameter** is a number that describes some aspect of a ***population***

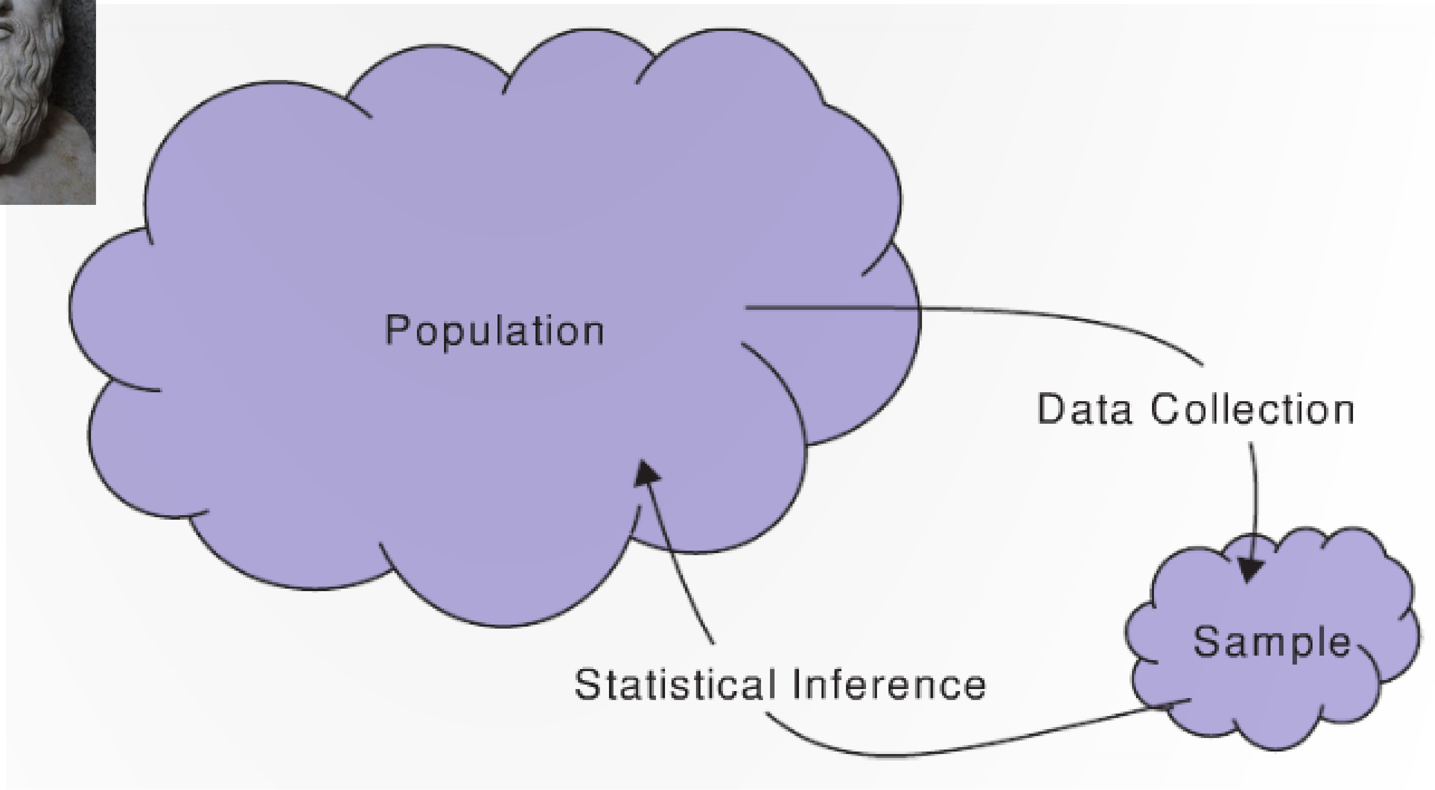


?

Parameters and statistics



Parameters



statistics

Proportions

For a *single* ***categorical variable***, the main ***statistic*** of interest is the *proportion* in each category

- E.g., the proportion of red sprinkles

$$\text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

Example proportion of red sprinkles

The sample

- orange, red, green, white, white, white, …, pink

The proportion for a ***sample*** is denoted $\hat{\mathbf{p}}$ (pronounced “p-hat”)

- $\hat{p}_{red}$ = 13/100 = 0.13

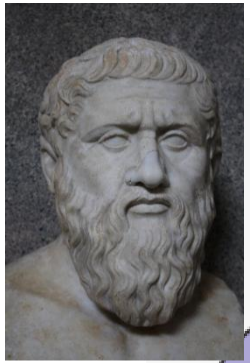
The proportion for a ***population*** is denoted $\boldsymbol{\pi}$ (the book uses p)

- π_{red} proportion if we had measured all sprinkles in the population

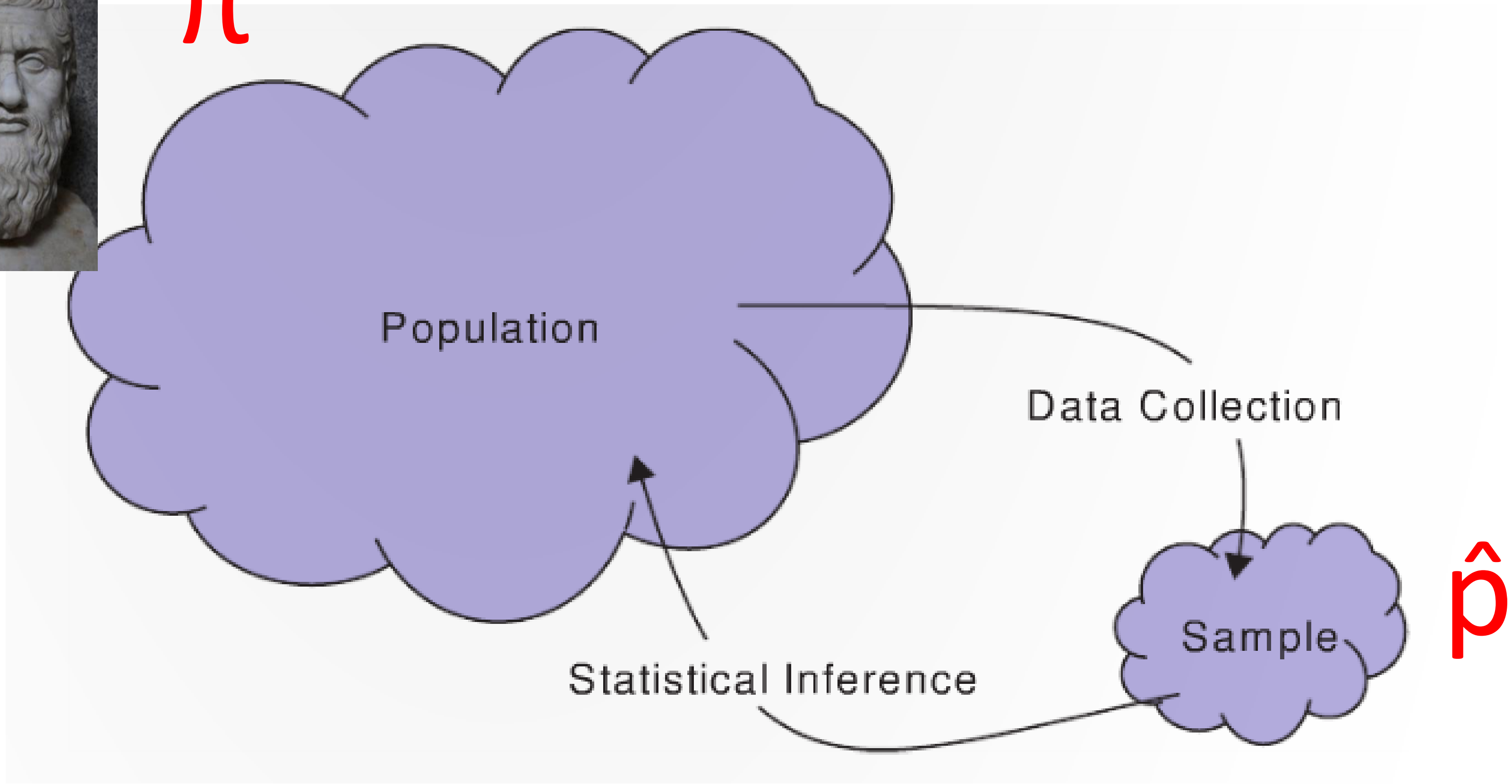
$\hat{p}$ is a ***point estimate*** of π

- i.e., $\hat{p}$ our best guess of what π is

Sample vs. Population proportion



π



$\hat{p}$

Different samples yield different values for the statistic

$\hat{p}_{s1\_red} = 0.13$

$\hat{p}_{s2\text{-}red} = 0.11$

$\hat{p}_{s3\text{-}red} = 0.15$

Calculating counts on a categorical variable

The count of how many items are in each category can be summarized in a ***frequency table***

Color	**green**	**orange**	**pink**	**red**	**white**	**yellow**		**Total**
Count	20	11	9	13	36	11		100

In R: my_table <- table(my_vector)

Calculating proportions (relative frequencies)

We can convert a frequency table into a ***relative frequency table*** by dividing each cell by the total number of items

Color	**green**	**orange**	**pink**	**red**	**white**	**yellow**		**Total**
Count	.20	.11	.09	.13	.36	.11		1

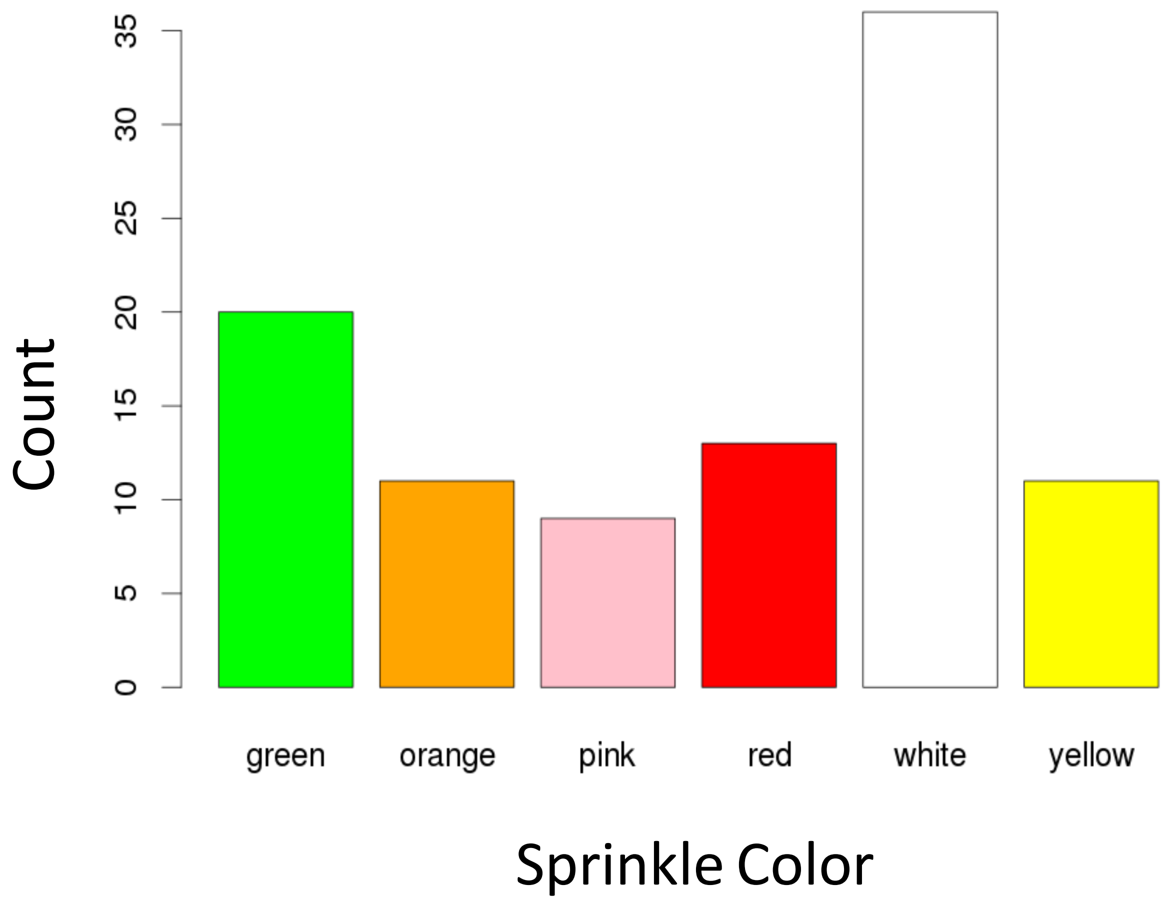
In R: prop.table(my_table)

Visualizing categorical data: The bar plot

A bar plot shows the number of items in each category

The height of each bar corresponds to the number of items in a given category

In R: barplot(my_table)

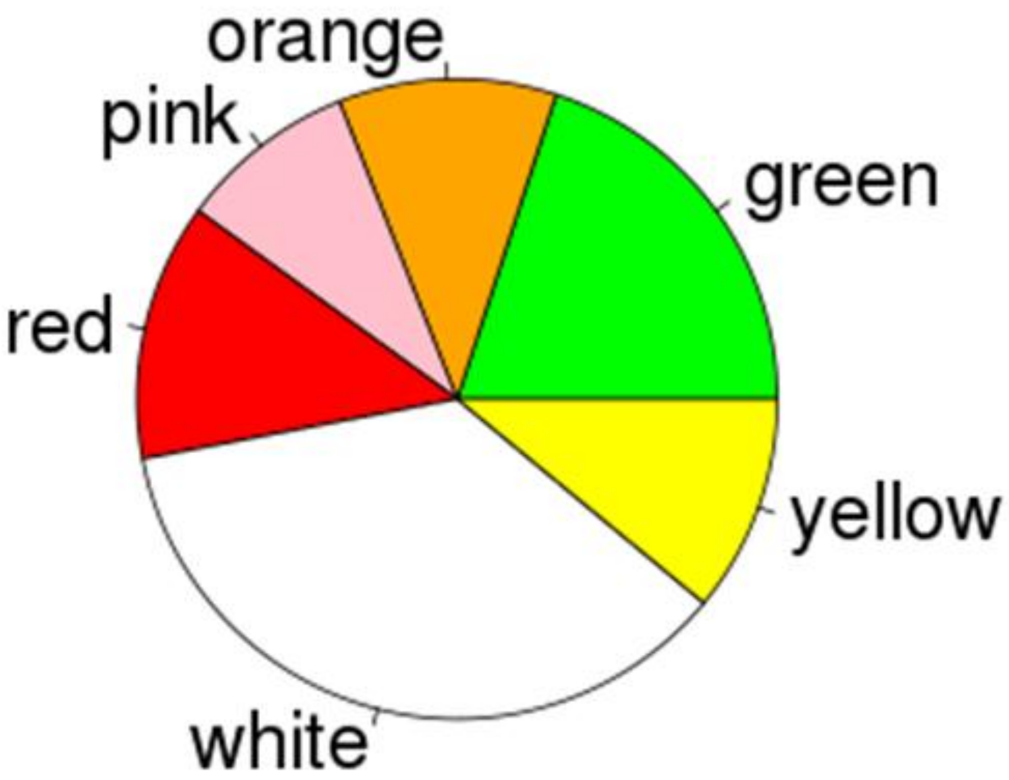


Visualizing categorical data: The pie chart

A pie chart plots the proportion of items in each category

The area of each segment corresponds to the proportion of items in that segment

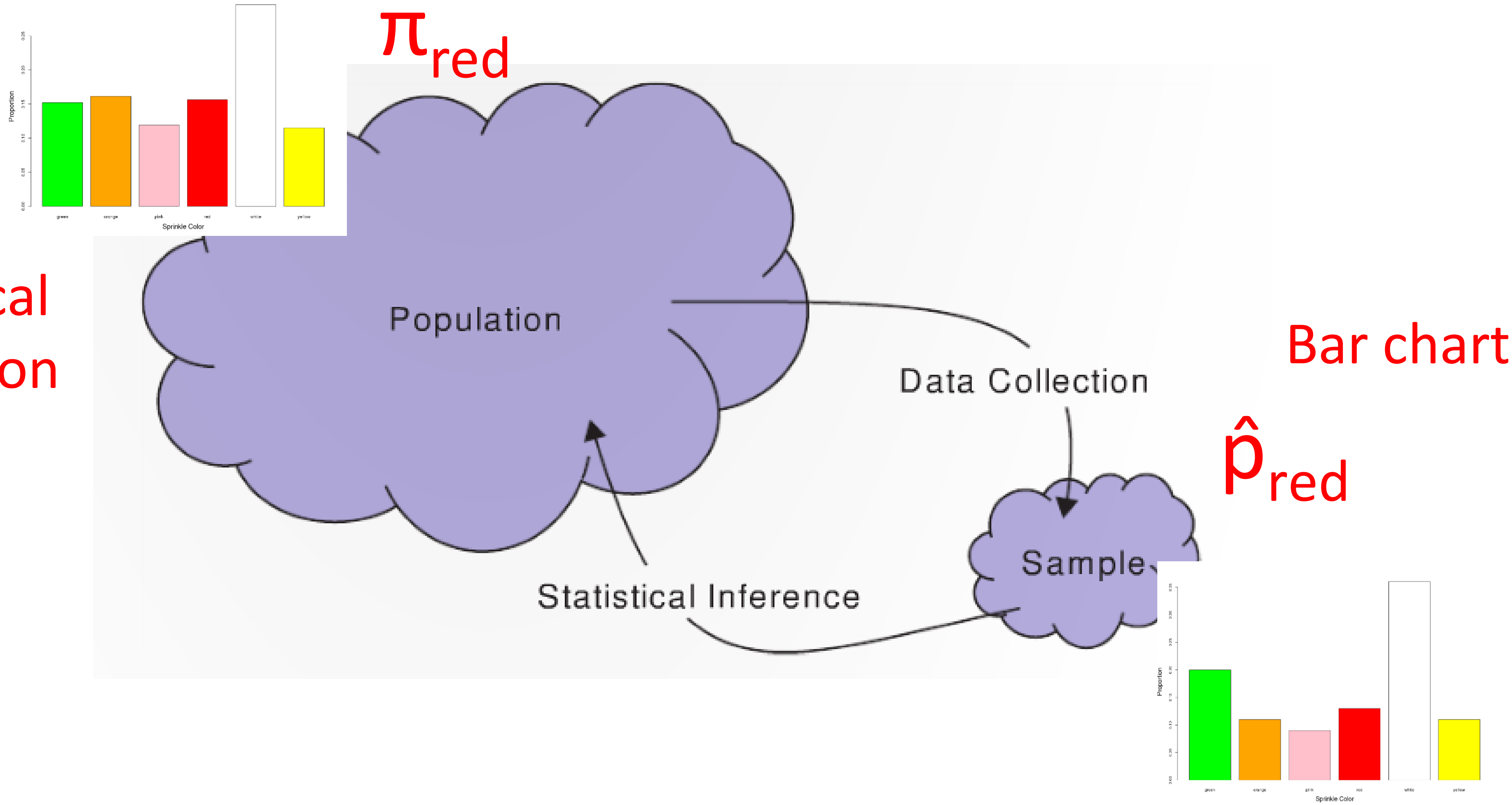
In R: pie(my_table)



Summary: Sample and Population proportion

π_{red}

Categorical distribution



Bar chart

$\hat{p}_{red}$

Example of categorical data: Presidential approval ratings

Attend the practice sessions to try this example!

Let’s sample virtual sprinkles in R…

Sampling virtual sprinkles

```
library(SDS100)

sprinkle_sample <- get_sprinkle_sample(100)

sprinkle_count_table <- table(sprinkle_sample)
sprinkle_prop_table <- prop.table(sprinkle_count_table)

barplot(sprinkle_count_table)
pie(sprinkle_count_table)
```

Two categorical variables

Two categorical variables

Sometimes we have measured two categorical variables for each case, and we want to investigate if there is a relationship between the levels of these categorical variables

- E.g., Suppose we have measure sprinkle **color and size**, and we want to investigate whether there is a relationship between these variables

	color	size
1	orange	large
2	green	large
3	white	medium
4	green	small
5	red	large

A **two-way table** shows the relationship between two categorical variables

- The category levels for one of the variables (factors) are listed down the rows
- The category levels for the other variable (factor) are listed across the columns
- Each cell in the table counts the number of cases that are in both the row and column categories

Size

Color

	Large	Medium	Small	Total
Green	5	7	8	20
Orange	3	2	8	13
Pink	3	3	2	8
Red	10	3	7	20
White	10	14	7	31
Yellow	2	3	3	8
Total	33	32	35	100

In R: `table(vector1, vector2)`

Two categorical variables

Sometimes we are interested in the proportion of one variable, given the other variable is a fixed value

- E.g., the proportion of large sprinkles that are red: $\hat{p}_{red|large}$

We can calculate these values by looking at the proportion in the relevant column or row

- $\hat{p}_{red|large}$ = 10/33 = 0.303

Note: In general: $\hat{p}_{A|B} = \hat{p}_{B|A}$

- $\hat{p}_{large|red}$ = 10/20 = 0.5

Size

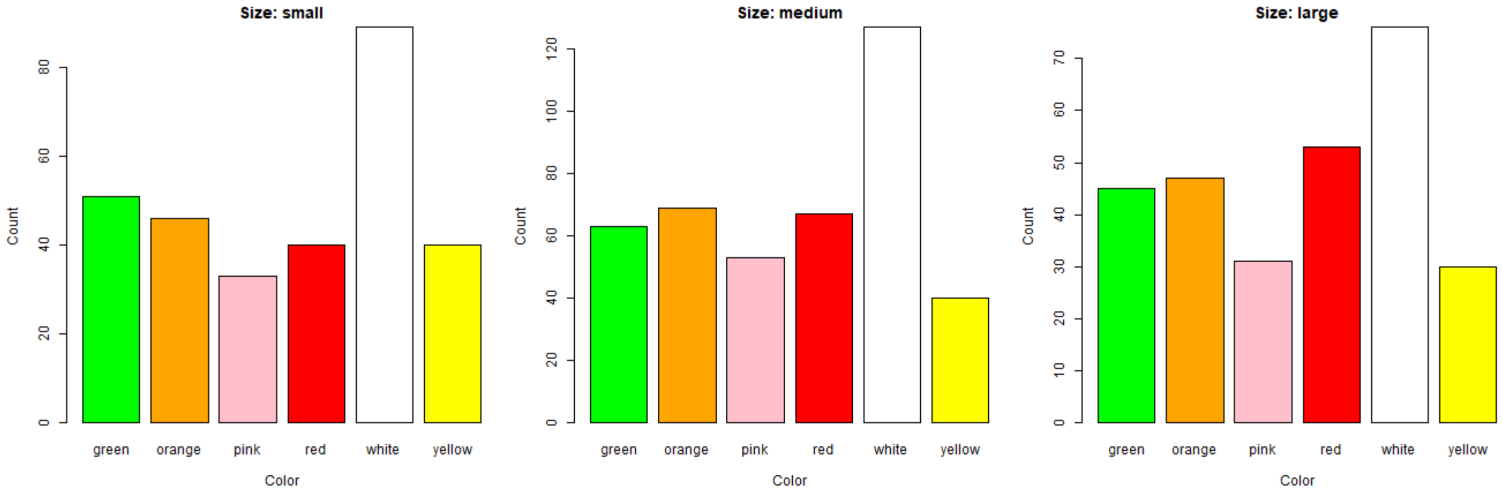
Color

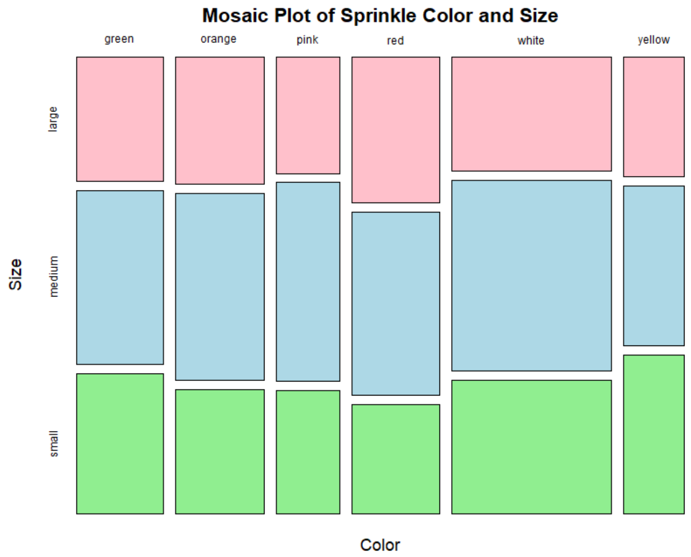
	Large	**Medium**	**Small**	**Total**
Green	5	7	8	20
Orange	3	2	8	13
Pink	3	3	2	8
Red	10	3	7	20
White	10	14	7	31
Yellow	2	3	3	8
Total	33	32	35	100

Brief mention: Visualizing two categorical variables

Faceted bar plot: A series of bar plots split into panels (“facets”) by a categorical variable, making it easier to compare patterns across groups

Mosaic plot: A graphical display of contingency tables where the area of each tile is proportional to the cell frequency, showing relationships between categorical variables





Let’s try it in R!

We will use the data from the class survey with the variables:

- The month you were born in
- Whether you were older or younger than other students in your class

Summary of concepts

1. A **statistic** is a number that is computed from ***data in a sample***
- The number of items in a sample is called the ***sample size*** and is usually denoted with the symbol n

2. A **parameter** is a number that describes some aspect of a ***population***

3. A point estimate is using a value of a statistic as a guess for the value of a parameter

4. When calculating proportions:
- The proportion statistic is denoted $\hat{\mathbf{p}}$
- The population proportion is denoted $\boldsymbol{\pi}$
- Thus $\hat{p}$ is a ***point estimate*** of π

5. Proportions can be summarized in a **relative frequency table** and can be visualized using **bar plots** and **pie charts**

6. **Two-way tables** can be used to summarize data from two categorical variables

Summary of R

```
# a vector of character strings (or factors)
my_sample <- c("orange", "red", "green", "white", " white", ... )

# creating a table using the table() function
my_table <- table(my_sample)

# creating a frequency table using the prop.table() function
prop.table(my_table)

# creating bar and pie charts
barplot(my_table)
pie(my_table)
```

Quantitative variables

Descriptive statistics for one quantitative variable

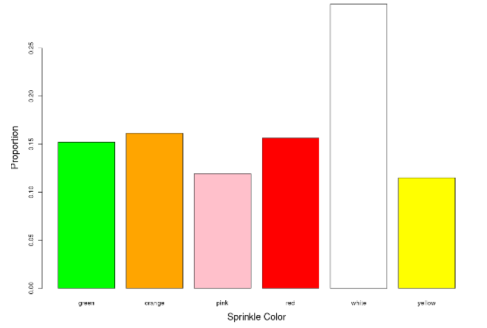
We will be looking at:

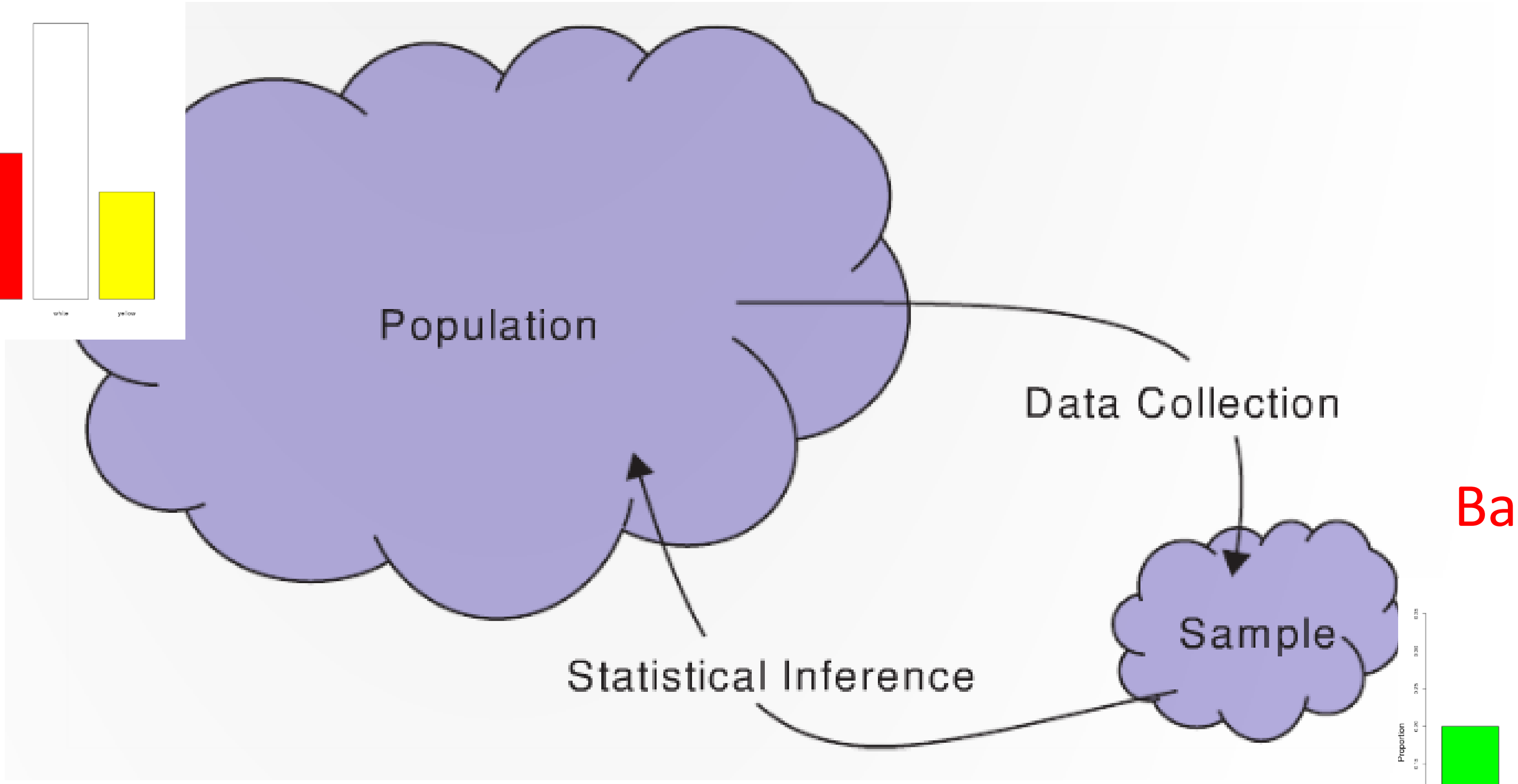
- What is the general ‘shape’ of the data
- Where are the values centered
- How do the data vary

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

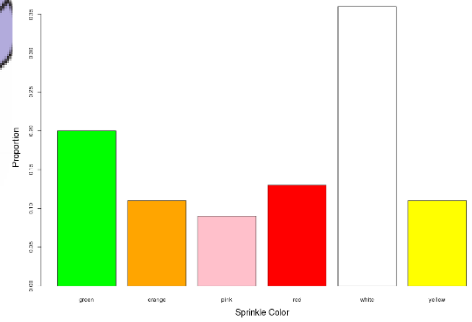
For categorical data we had...

Categorical Distribution (π)

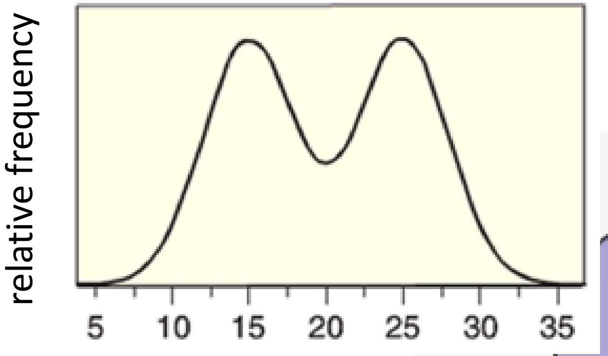


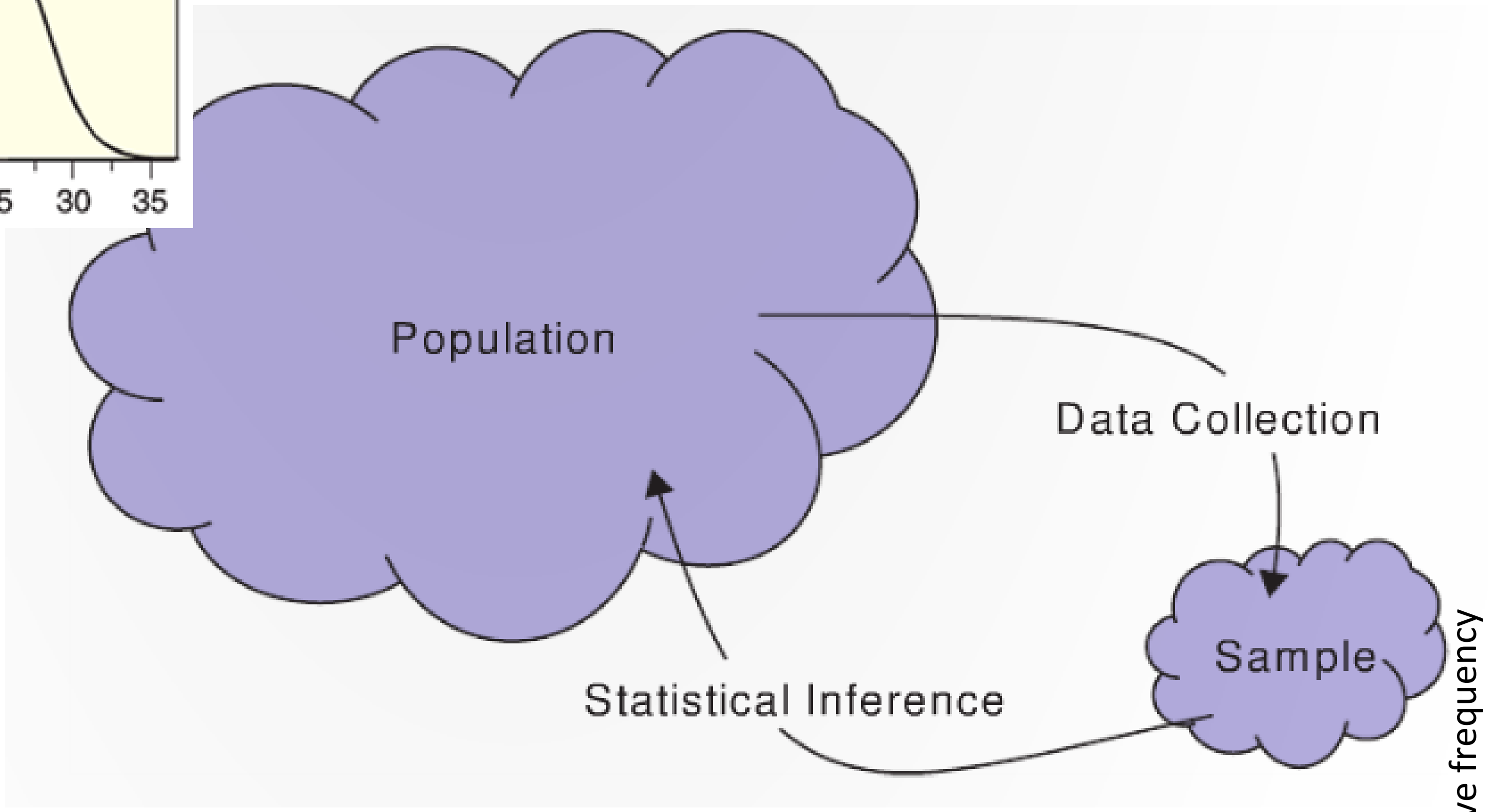


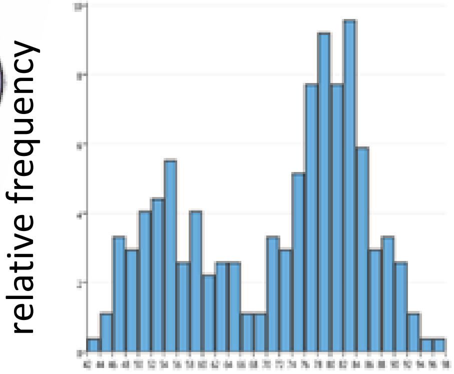
Bar chart ($\hat{p}$)



Population distributions and sample histograms





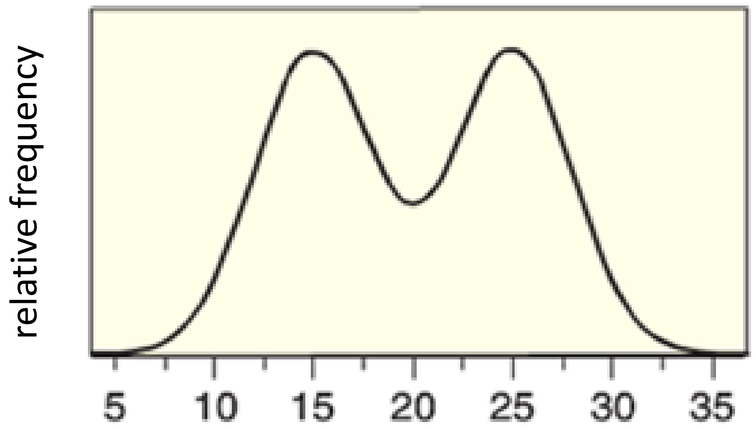


Histograms

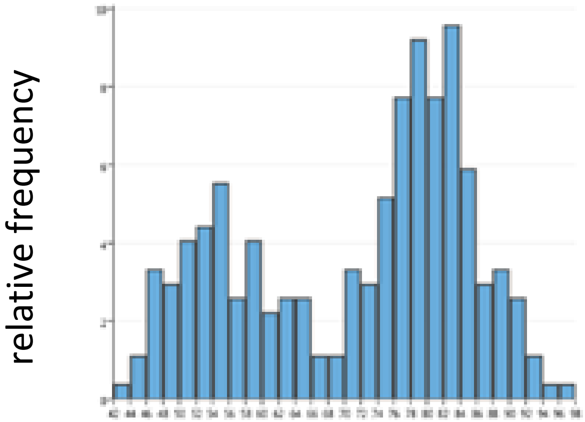
Histograms are a way of visualizing a sample of quantitative data

- 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

```
> 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

<u>Questions:</u>

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: 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

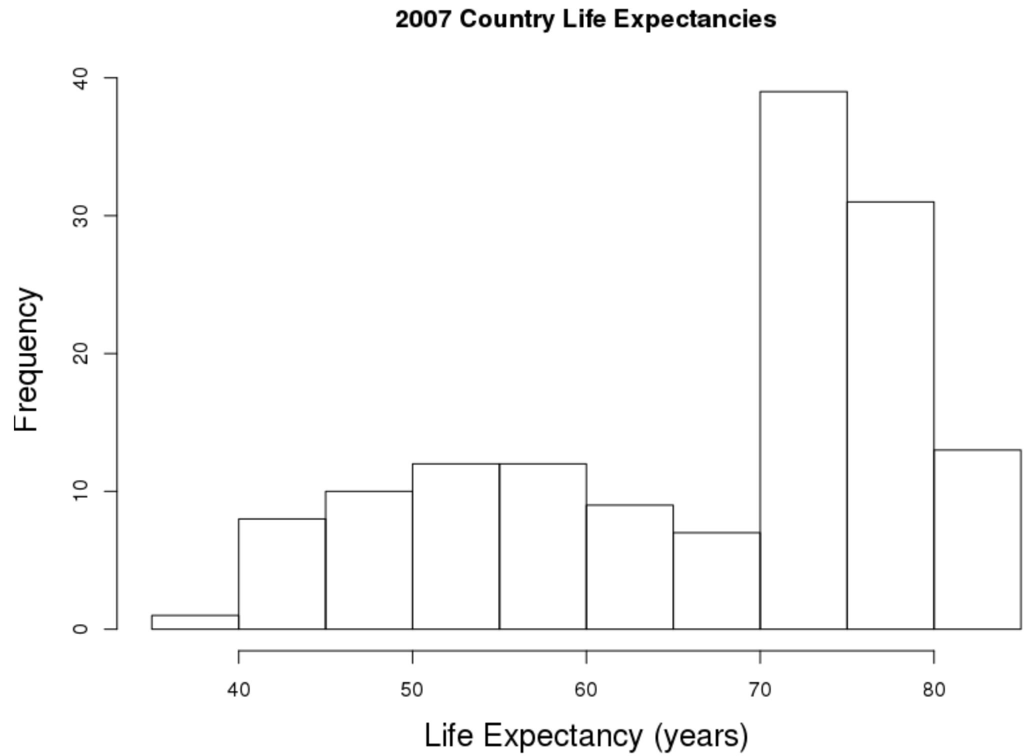
- 35-40, 40-45, 45-50, … 75-80, 80-85

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



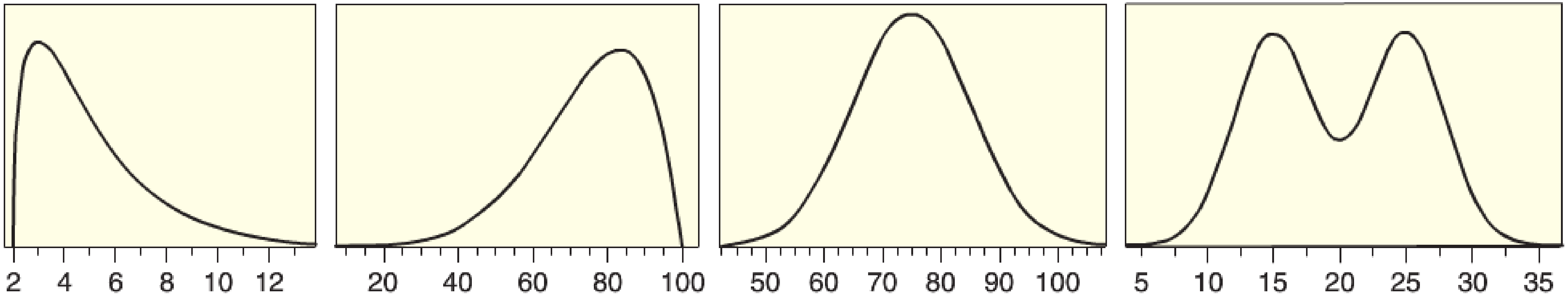
R: `hist(v)`

Gapminder: life expectancy in different countries

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

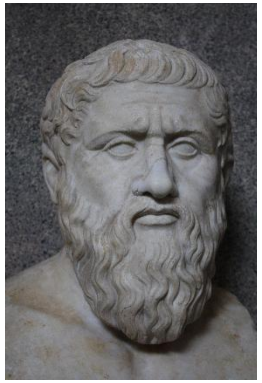
```
# pull a vector of life expectancies from the data frame
> life_expectancy <-  gapminder_2007$lifeExp
> hist(life_expectancy)
```

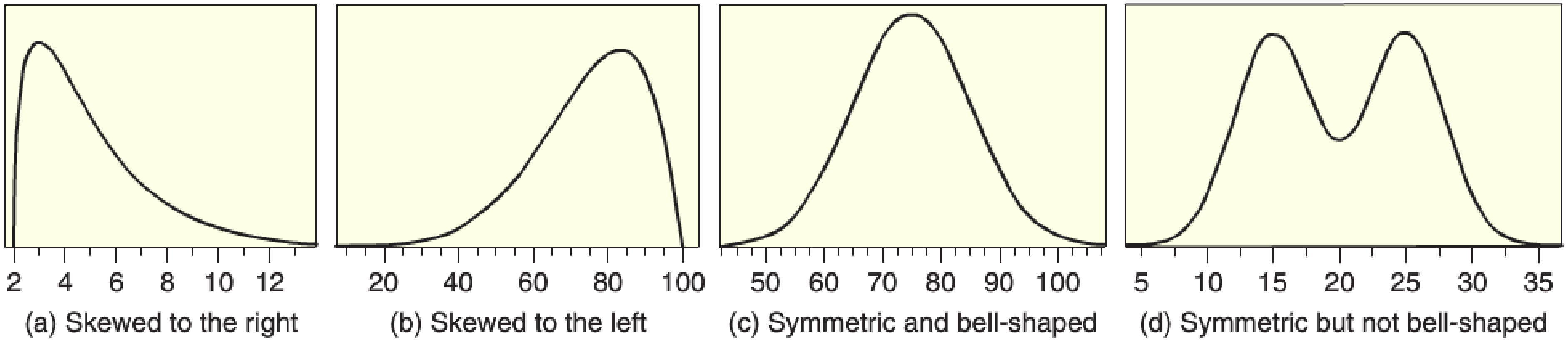
Common shapes for distributions



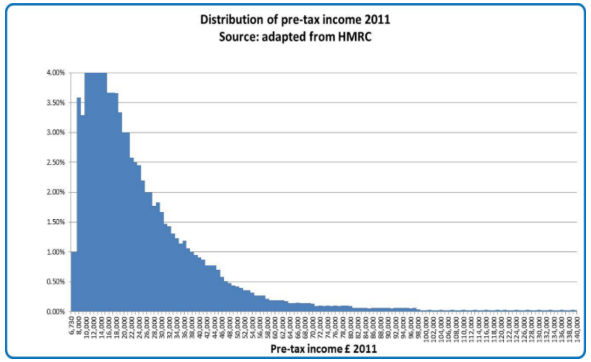
(a) Skewed to the right (b) Skewed to the left (c) Symmetric and bell-shaped (d) Symmetric but not bell-shaped

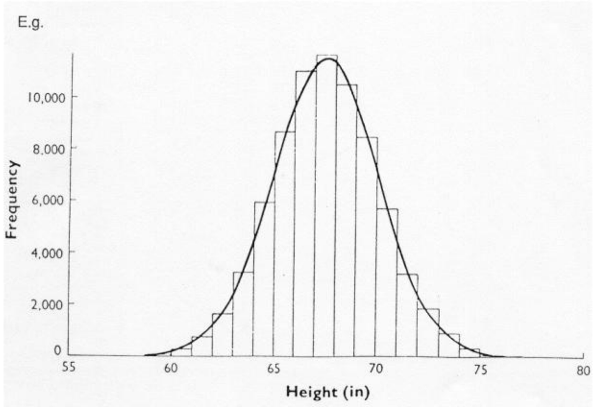
Plato and shadows: distributions and histograms

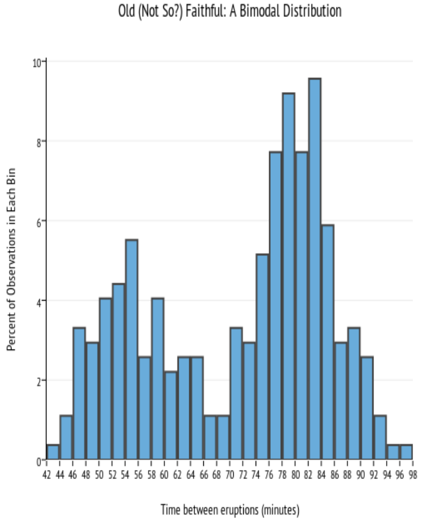




Income distribution

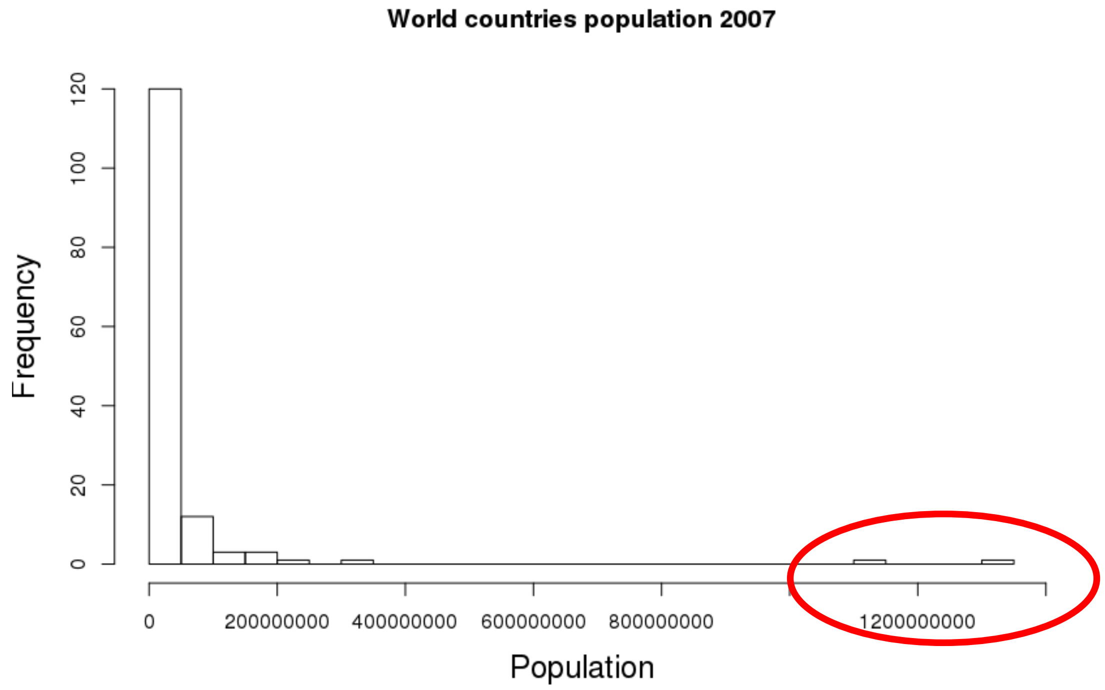






Outliers

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

Descriptive statistics for the center of a distribution

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

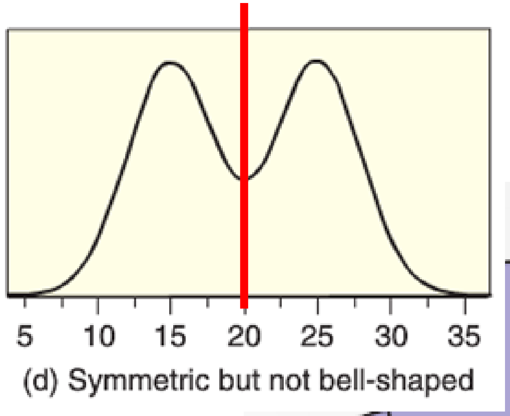
Question: what is a numerical summary of a sample of data called?

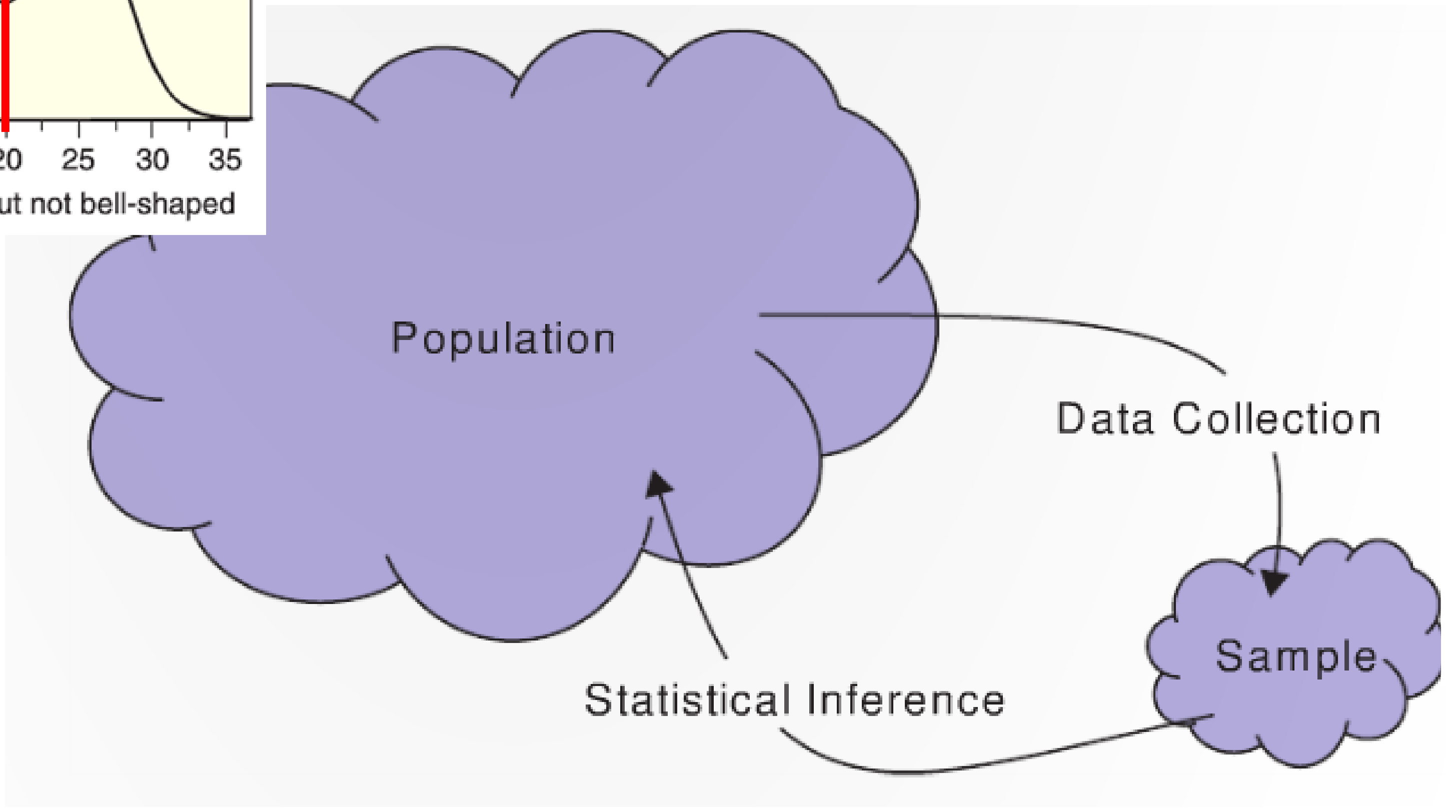
A: a statistic!

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

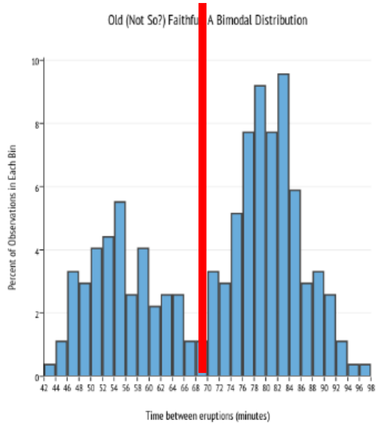
Sample and population mean

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

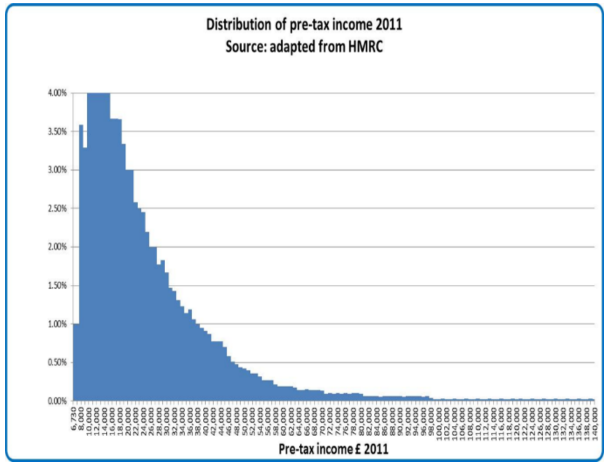
1. A ***probability distribution*** shows the ***relative likelihood*** that we will get a data point in the population with a particular value
    - (for a more precise definition take a class in probability)

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)
```