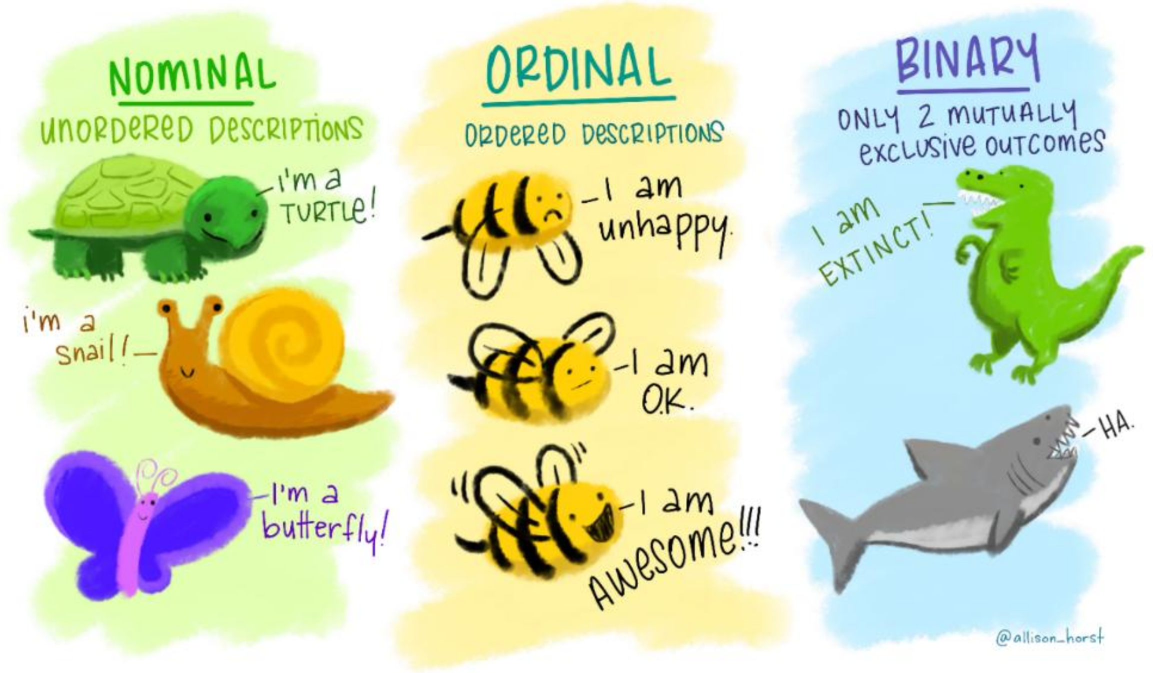
Categorical data continued and introduction to quantitative data analysis



Overview

Review of:

- Basics of R
- Categorical data concepts and R

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, Januray 25th at 11pm

```
library(SDS1000)
goto_homework(1)
```

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

Practice sessions this week

- Wednesday: 4-6 pm
- Thursday: 4-6 pm
- Friday: 10-12 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

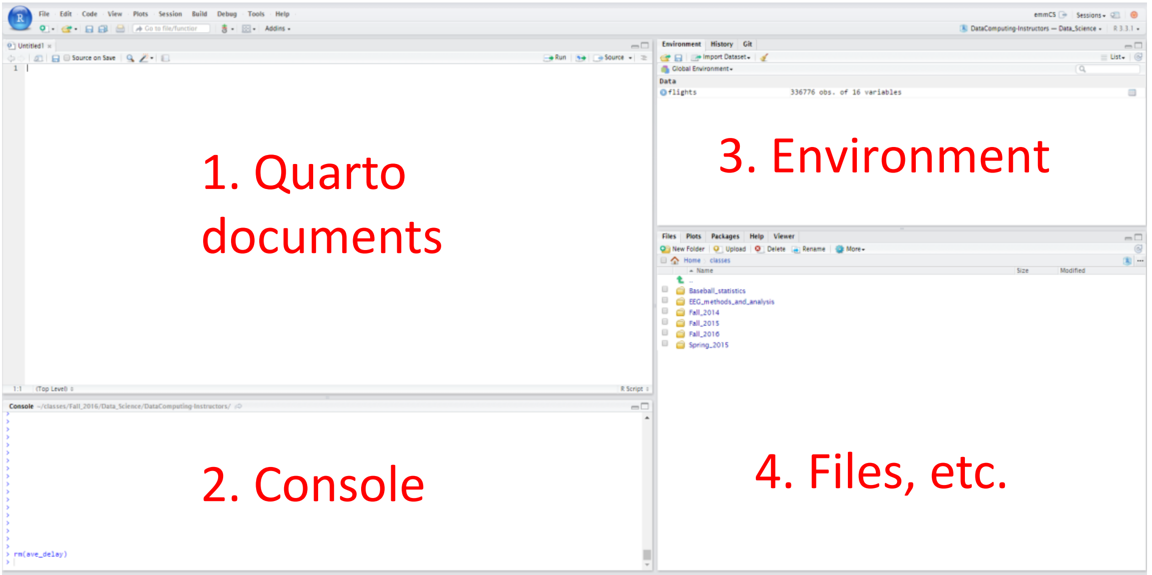
- Although don't print out hundreds of extra pages of numbers!

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

Note about 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

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
```

Questions?

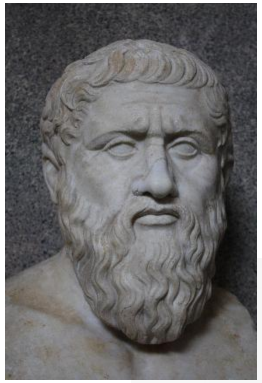
Review

Categorical variables

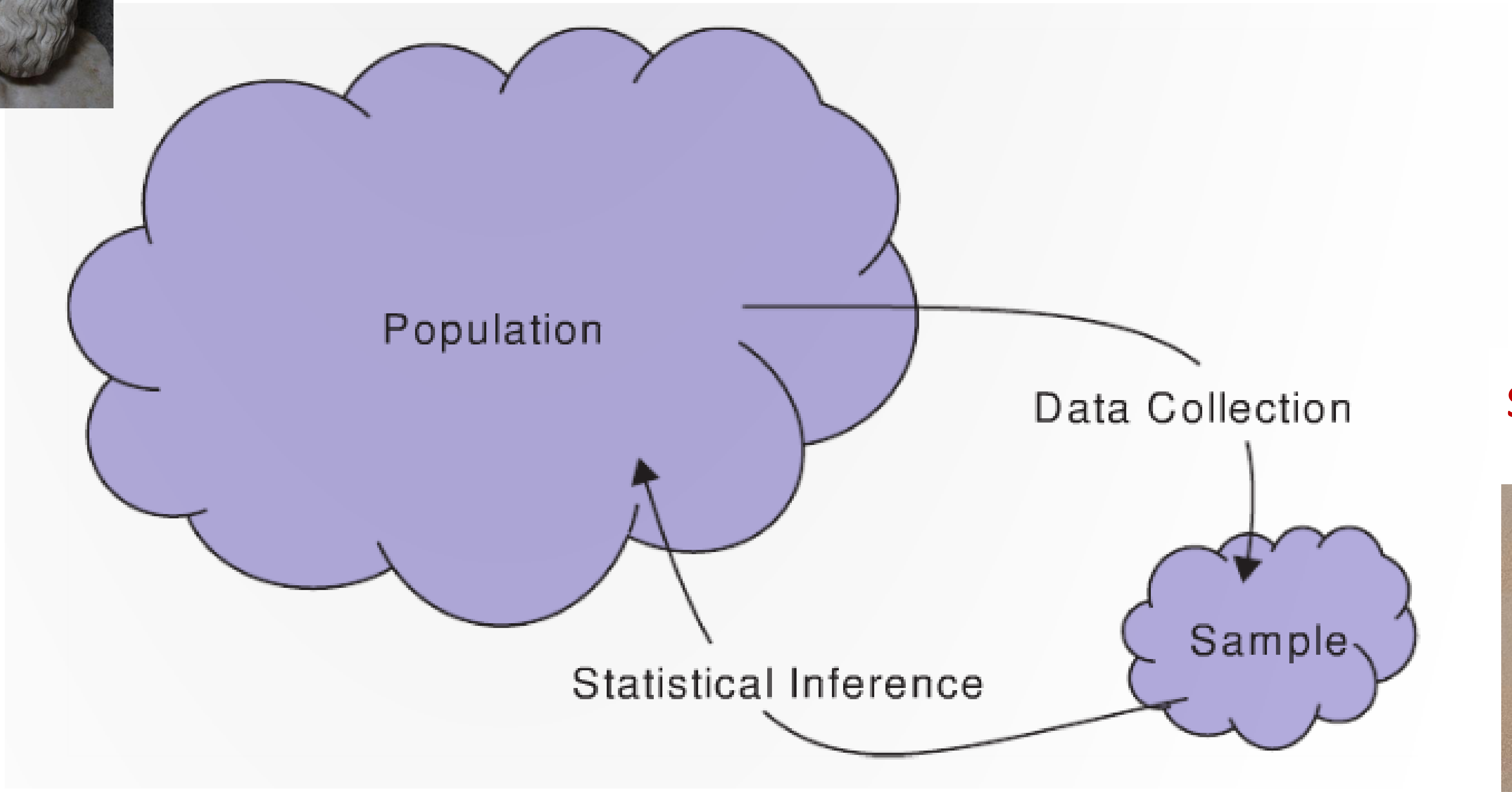
Quiz: Art time!

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

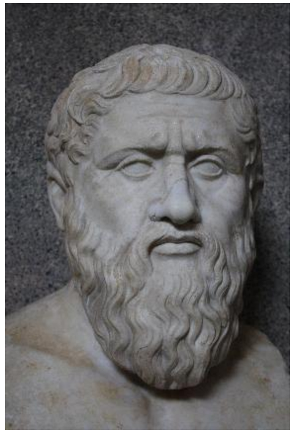


parameter: π



statistic: $\hat{p}$

Underlying concepts: the P's and the S's



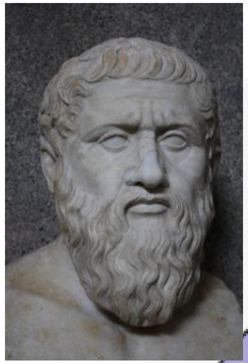
P-Truth

- Population or process
- Parameter
- Plato (Greek symbols)

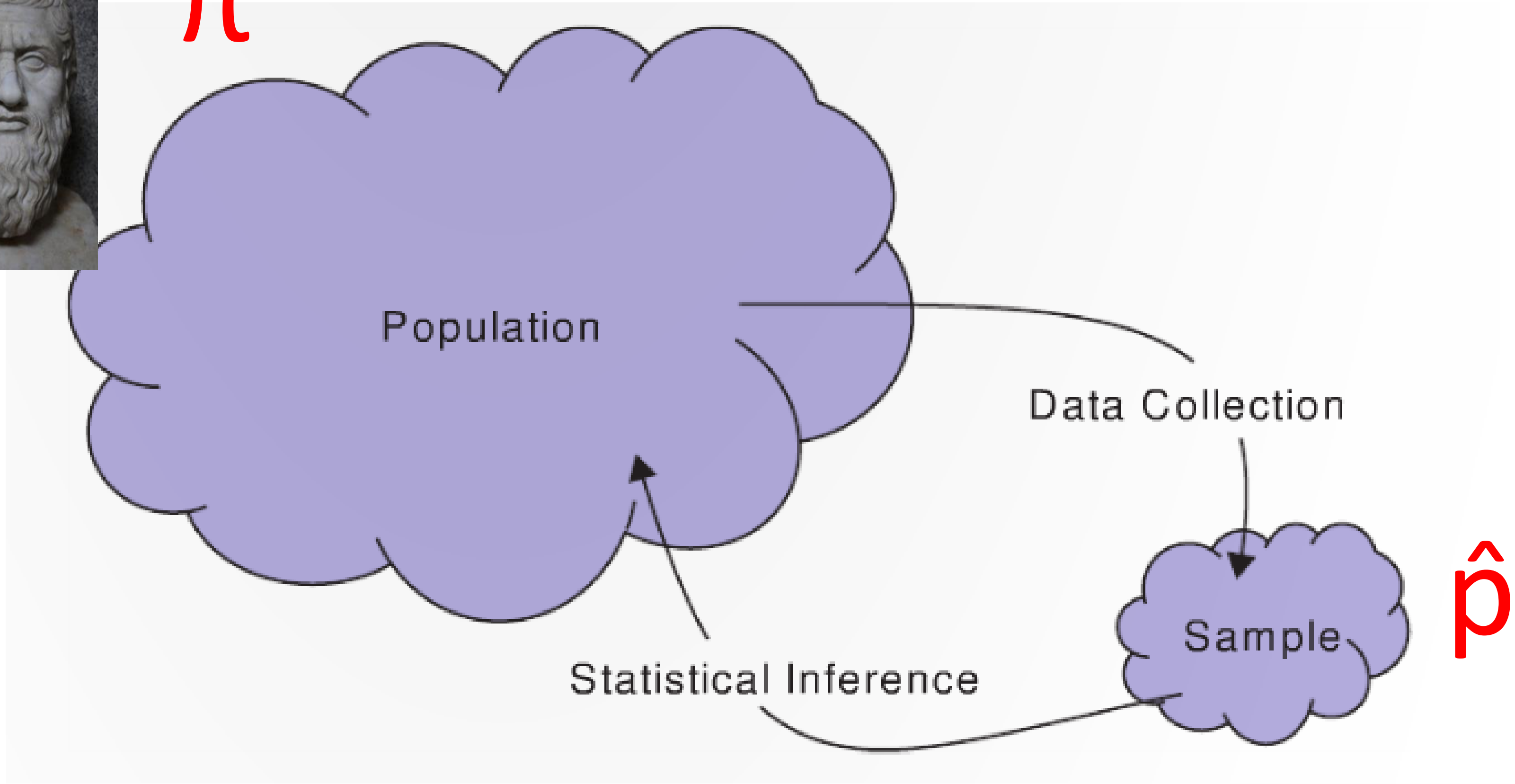
S-shadows

- Sample
- Statistic
- Shadow (Latin symbols)

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

Suppose we have a vector of sprinkle colors:

```
cat_vec  <-  c("red", "white", "red", …)
```

We can create a frequency table using:

```
my_table  <-  table(cat_vec)
```

Vector-like object that has the counts for each color

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

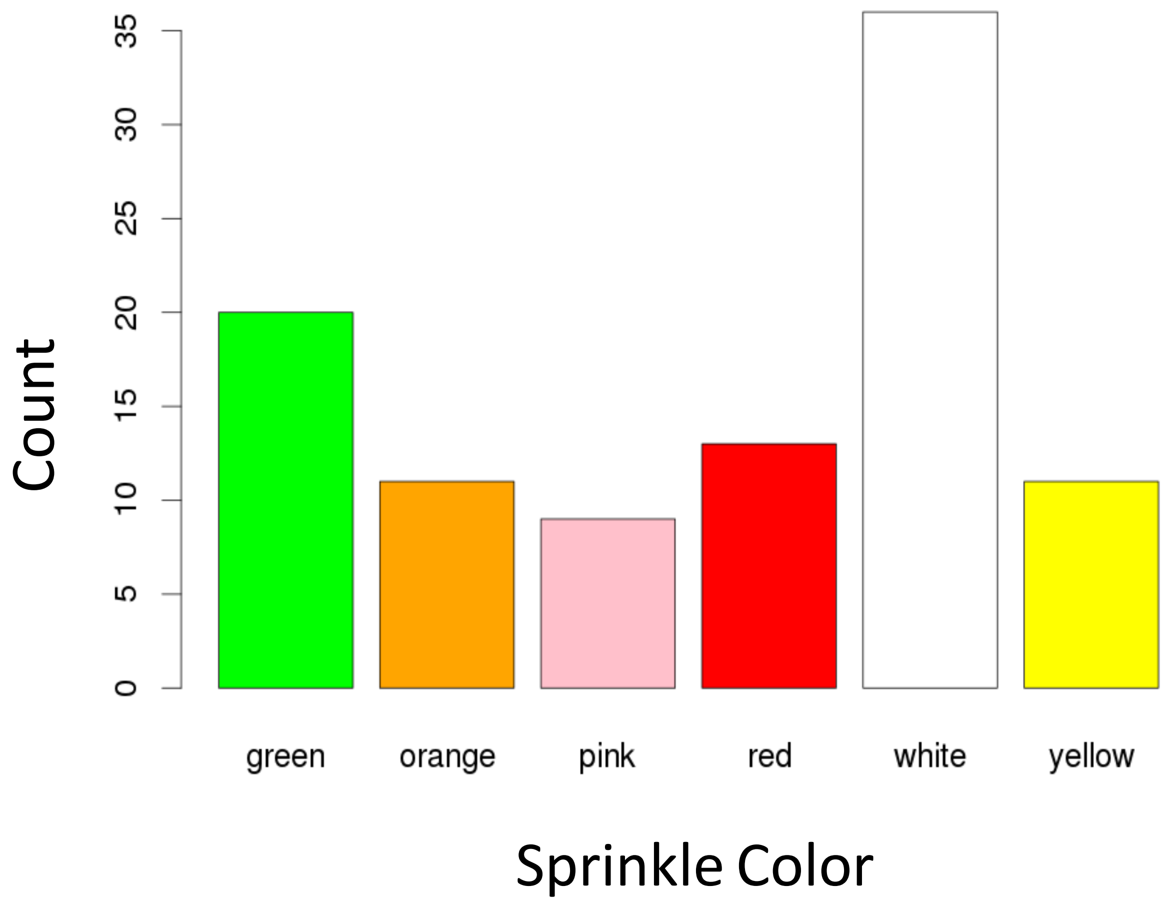
In R: `prop.table(my_table)`

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)

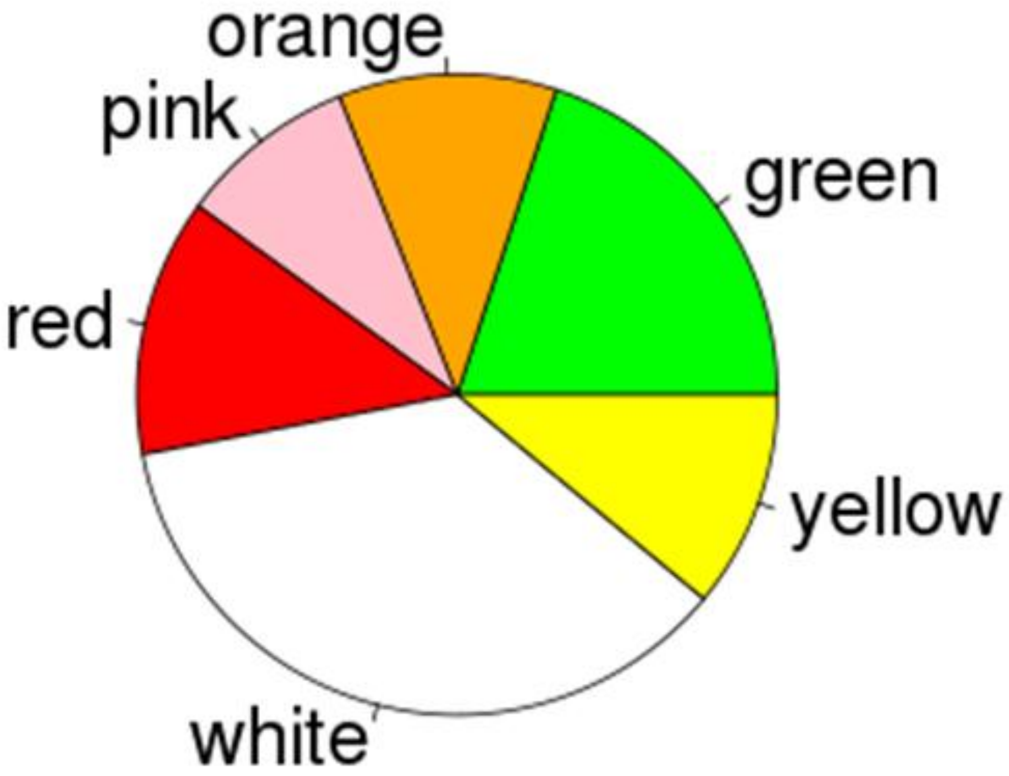


Visualizing categorical data: The pie chart

A pie chart plots the proportion of items in each category

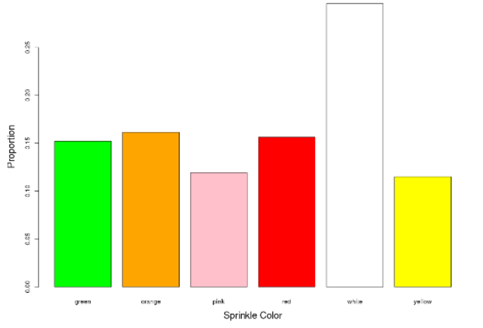
The area of each segment corresponds to the proportion of items in that segment

In R: `pie(my_table)`

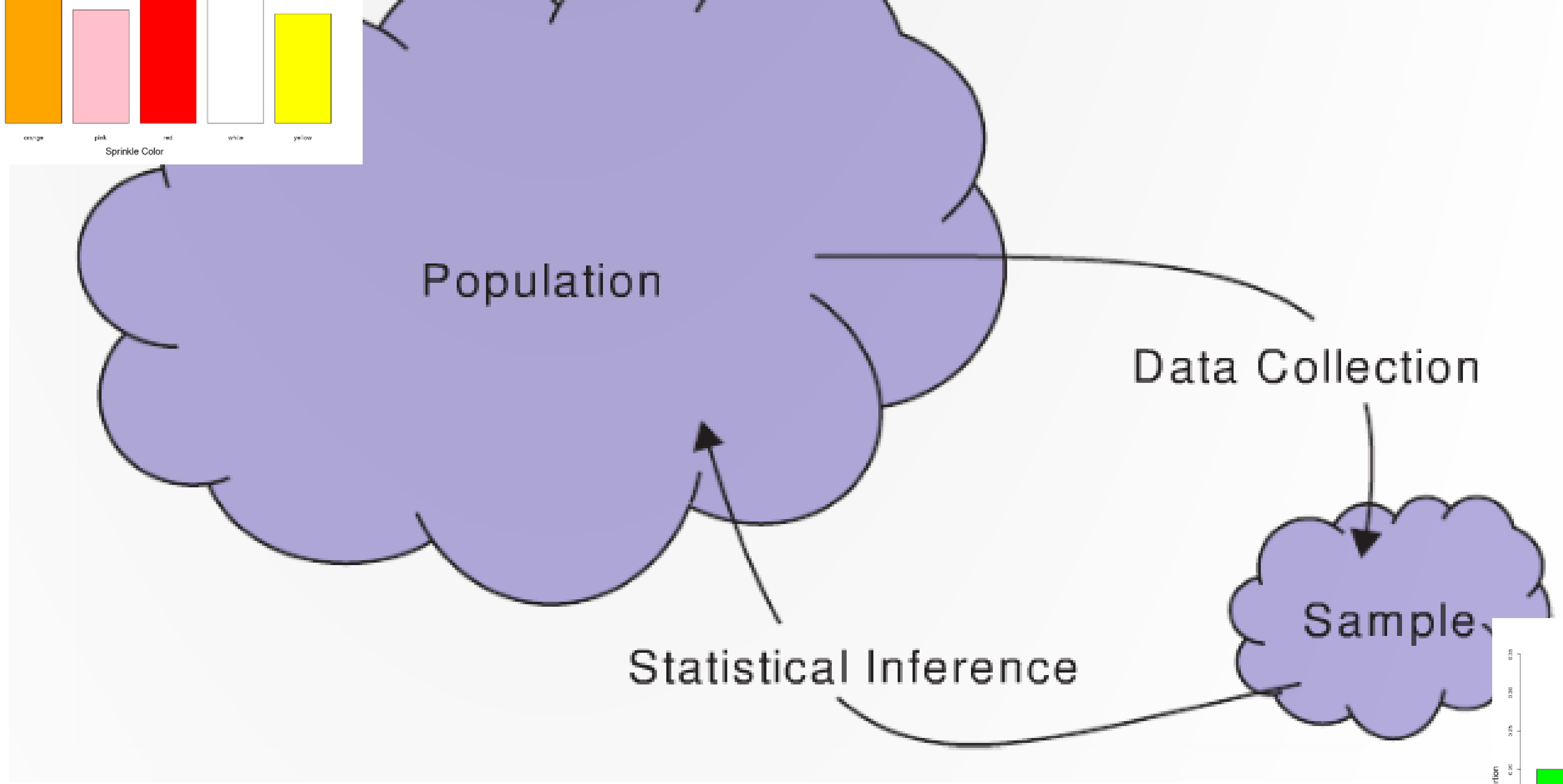


Summary: Sample and Population proportion

π_{red}

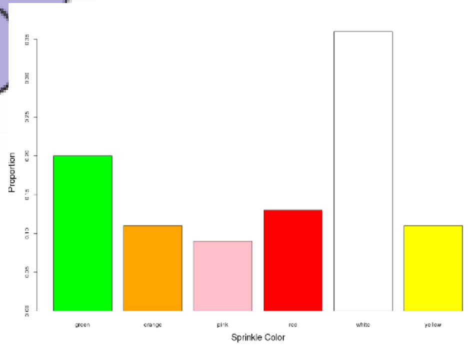


Categorical distribution



$\hat{p}_{red}$

Bar chart



Example of categorical data: Presidential approval ratings

Attend the practice sessions to try this example!

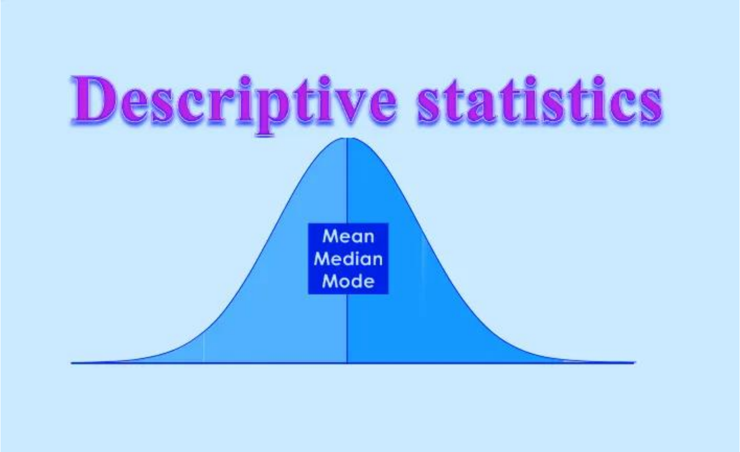
Questions?

Sampling virtual sprinkles

```
library(SDS100)          # load class package

sprinkle_sample <- get_sprinkle_sample(100)       # get a sample of sprinkles

sprinkle_count_table <- table(sprinkle_sample)              # frequency table
sprinkle_prop_table <- prop.table(sprinkle_count_table)     # relative frequency table

prop_red <-  get_proportion(sprinkle_sample, "red")      # proportion of red sprinkles

barplot(sprinkle_count_table)      # bar plot
pie(sprinkle_count_table)          # pie chart
```

Quantitative variables

Descriptive statistics for one quantitative variable



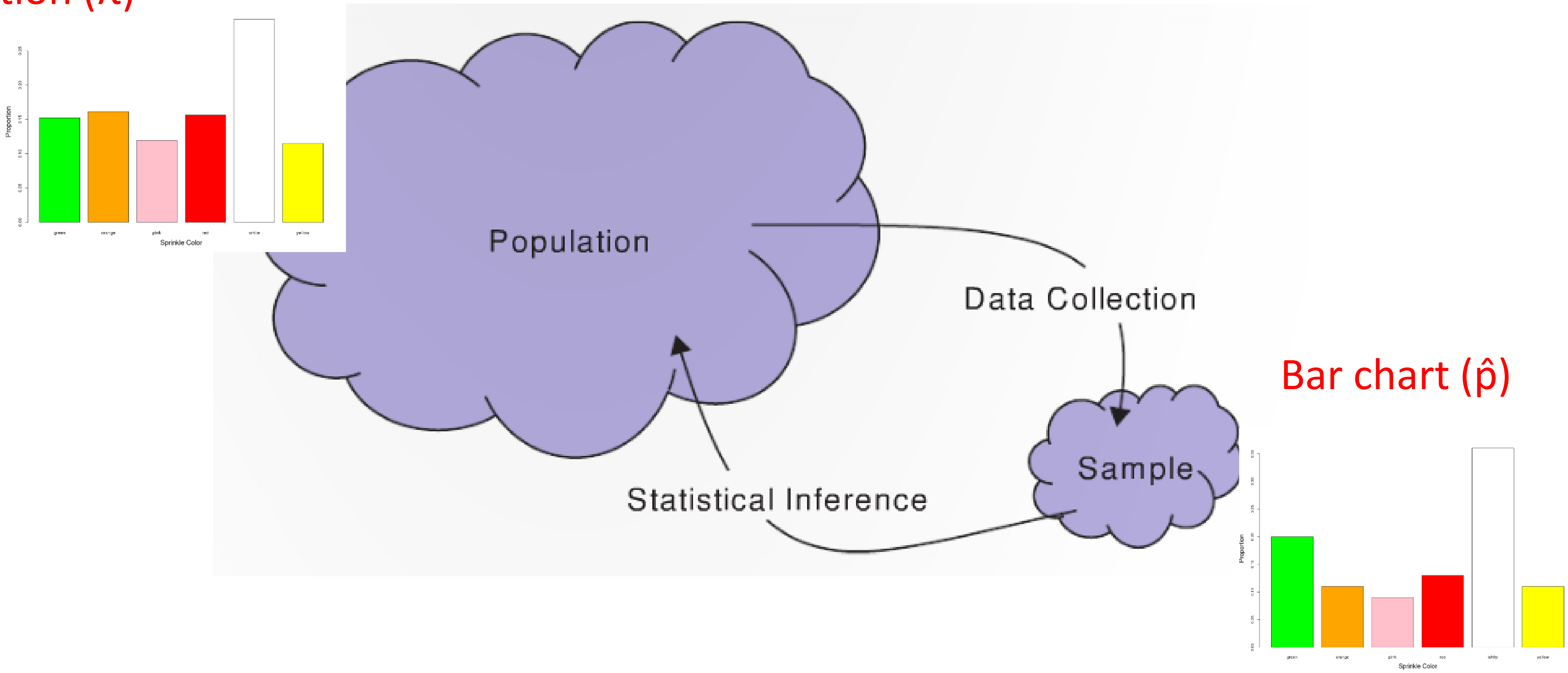
We will be looking at:

- 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***

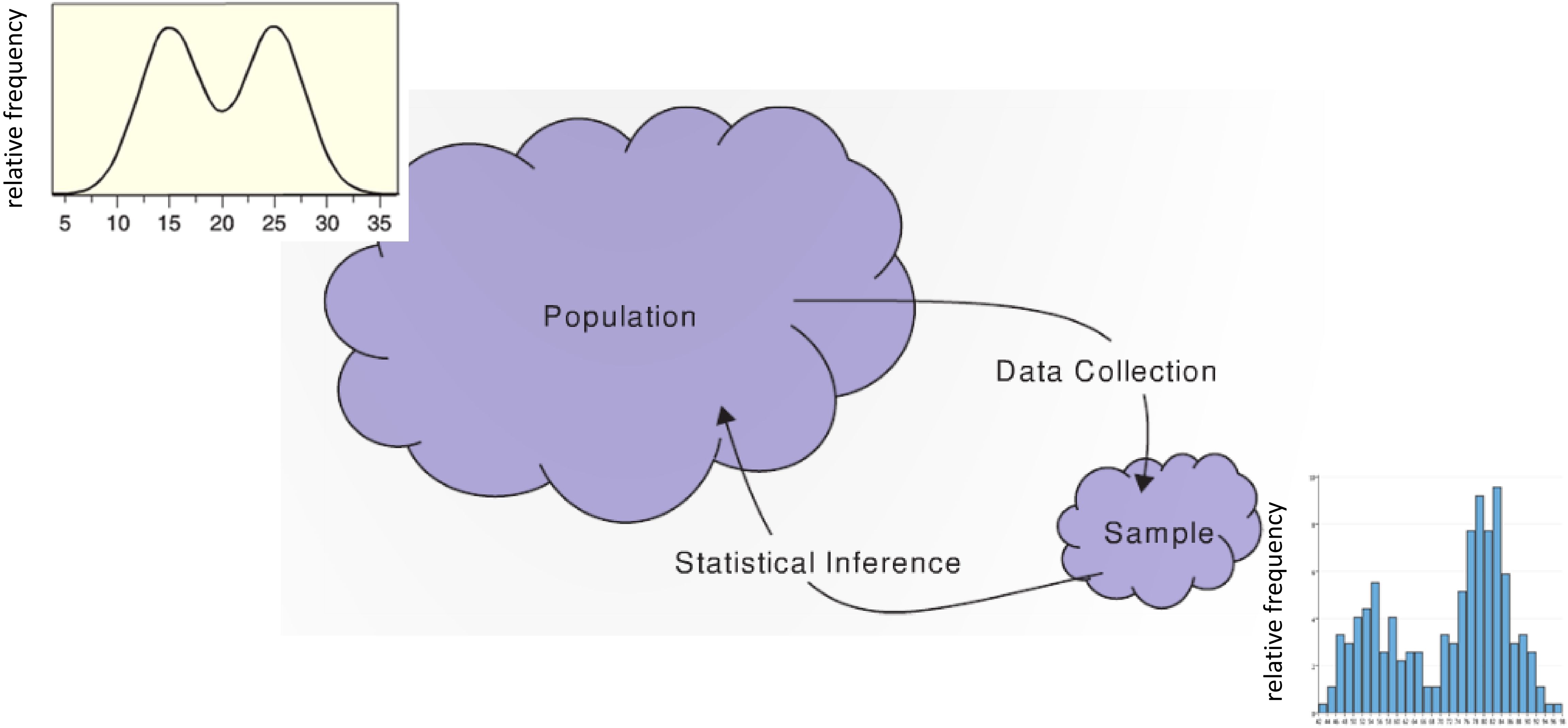
For categorical data we had…

Categorical Distribution (π)



Bar chart ($\hat{p}$)

Population distributions and sample histograms



Gapminder data

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 (column) in the structured 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

Gapminder data

The gapminder_2007 data frame contains information about countries in the world

	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

Questions:

1. What are the cases?
2. What are the variables?
3. Are the variable categorical or quantitative?
4. What is the population?

To learn more about the data see this video

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

We can access individual vectors of data using the $ symbol

```
continents <- gapminder_2007$continent   # same as using  c("Asia", "Europe", etc.
```

Since this is categorical data we could create frequency tables, bar plots, etc.

```
continent_table <- table(continents)

barplot(continent_table)
```

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

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

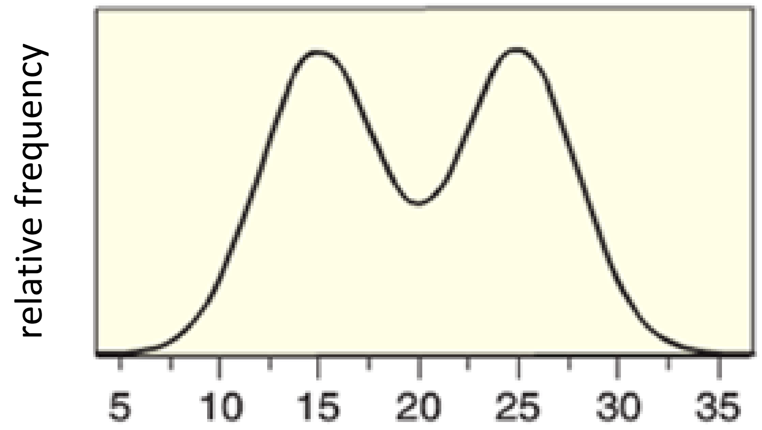
```
# pull a vector of life expectancies from the data frame
    life_expectancy <-  gapminder_2007$lifeExp
```

Histograms

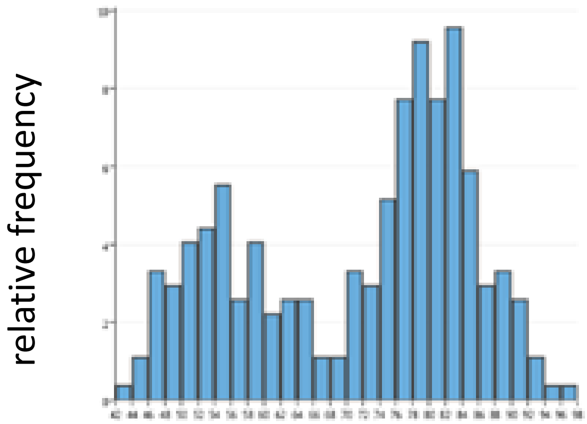
Histograms are a way of visualizing a sample of quantitative data

- 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



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

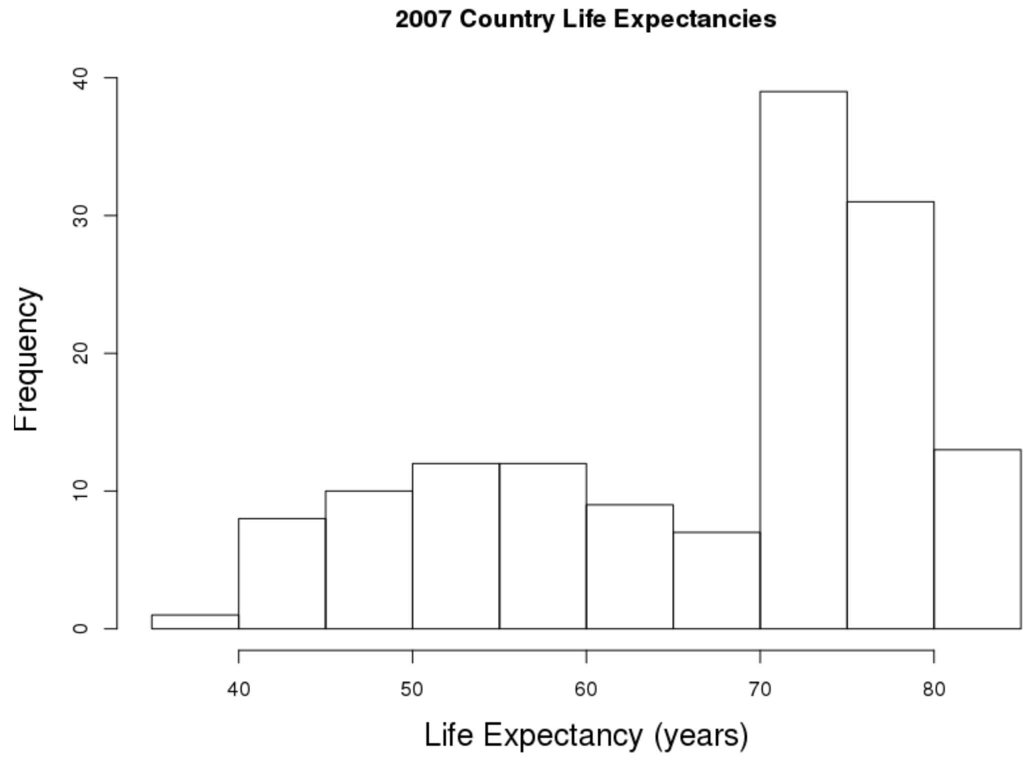
- 35-40, 40-45, 45-50, … 75-80, 80-85

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



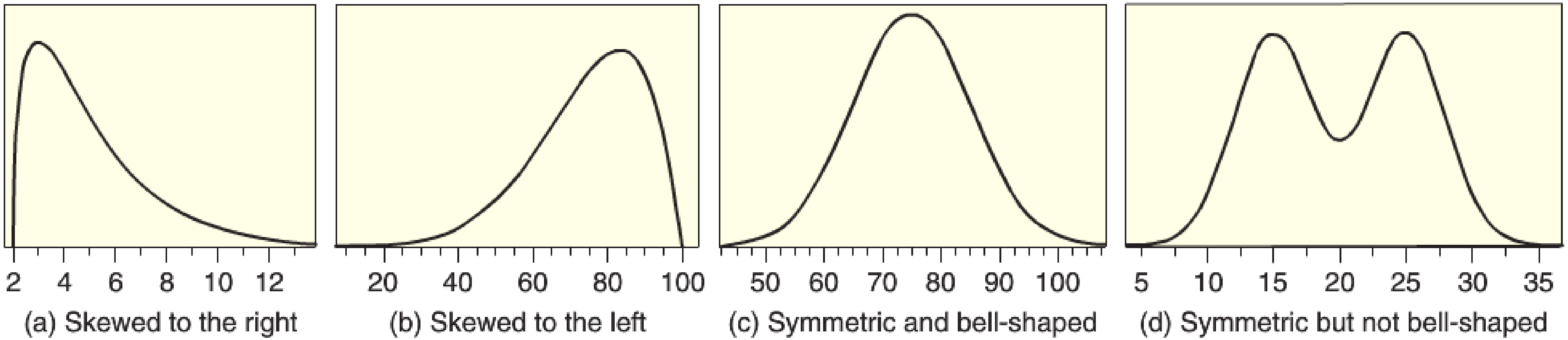
R: `hist(v)`

Gapminder: life expectancy in different countries

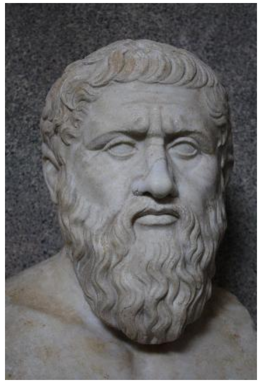
Let's create a histogram of the life expectancy in different countries using the hist() function

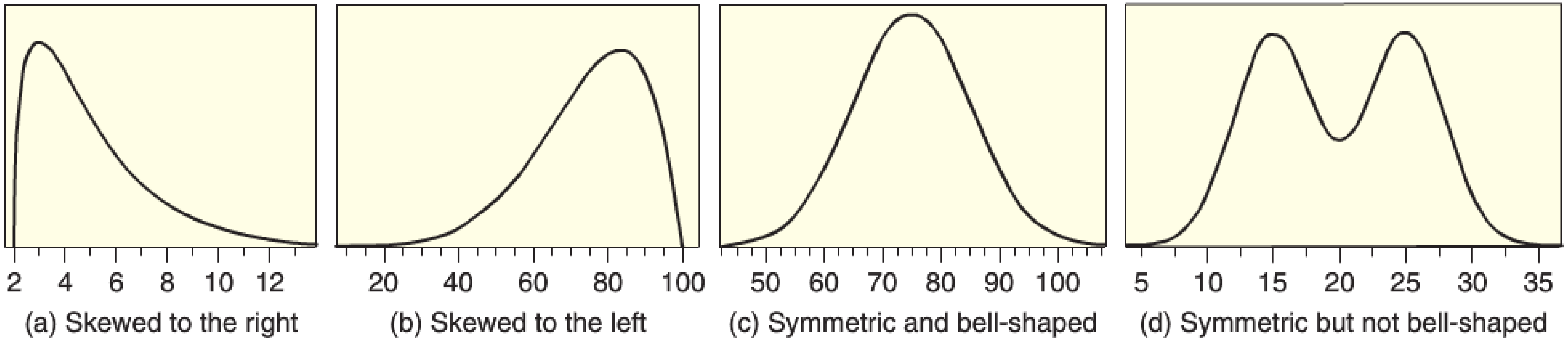
```
# pull a vector of life expectancies from the data frame
    life_expectancy <-  gapminder_2007$lifeExp

# create the histogram
   hist(life_expectancy,
         xlab = "Life expectancy")
```

Common shapes for distributions

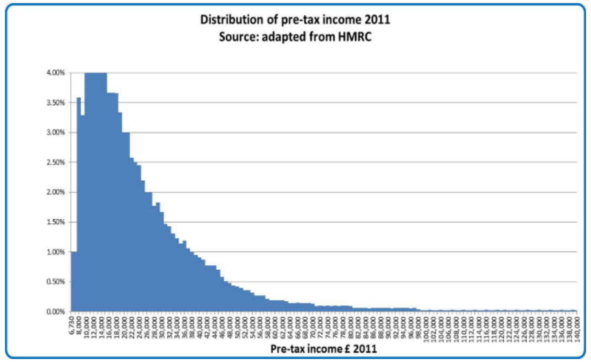


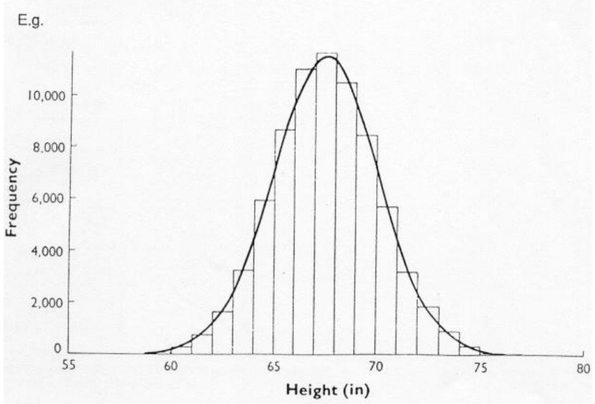
Plato and shadows: distributions and histograms

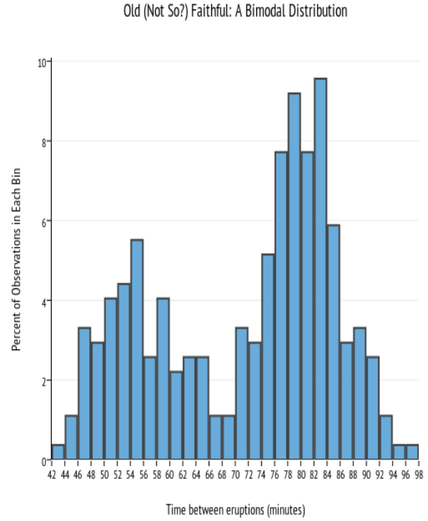




Income distribution

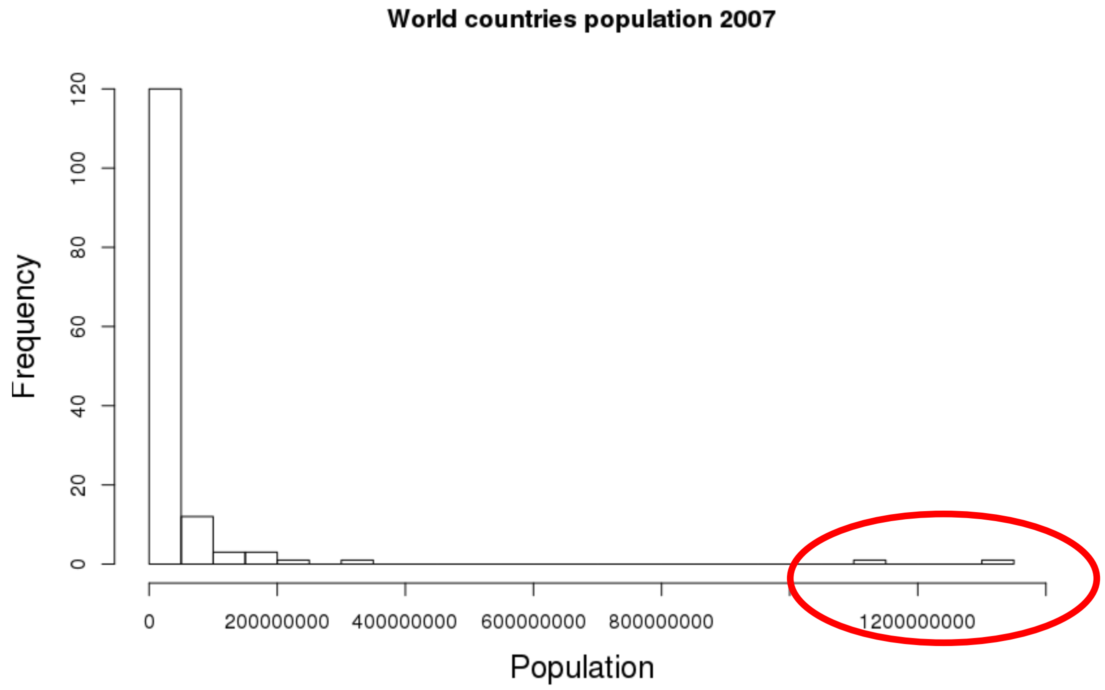






Outliers

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



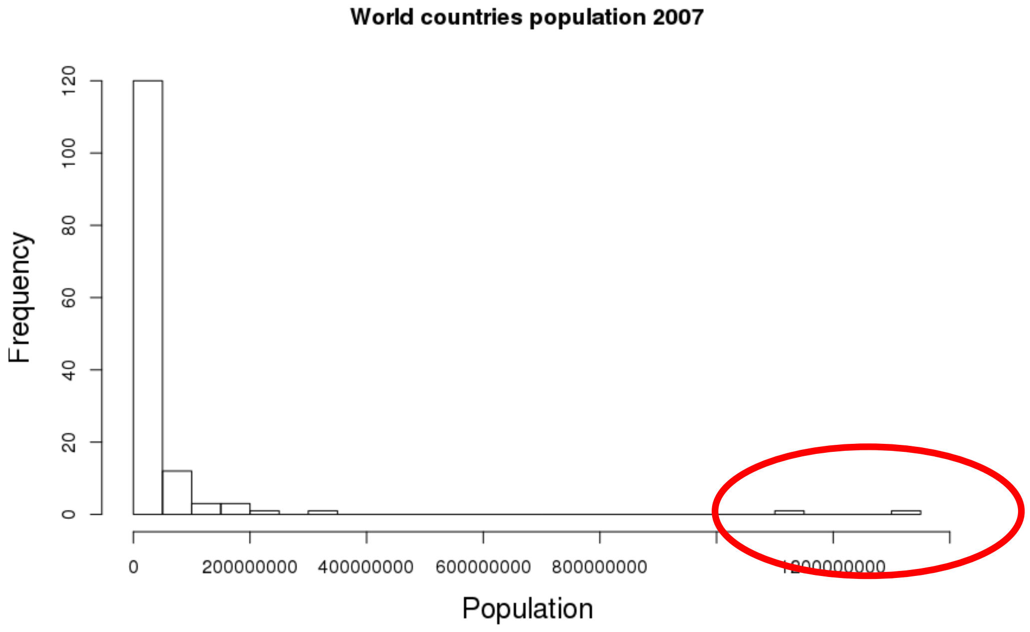
Outliers can potentially have a large influence on the statistics you calculate!

Outliers

Q: What should you do if you have outliers in your data?

A: See if we can tell why the outliers exist by examining the data!

- If the outliers are due to a mistakes, one can remove them
- If they are not due to mistakes, one should explain why they exist, and potentially try the analyses with and without the outliers to see if the analysis is affected



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?

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

The mean

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

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

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

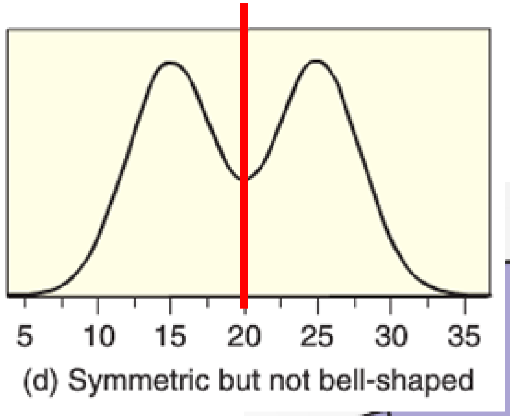
Notation

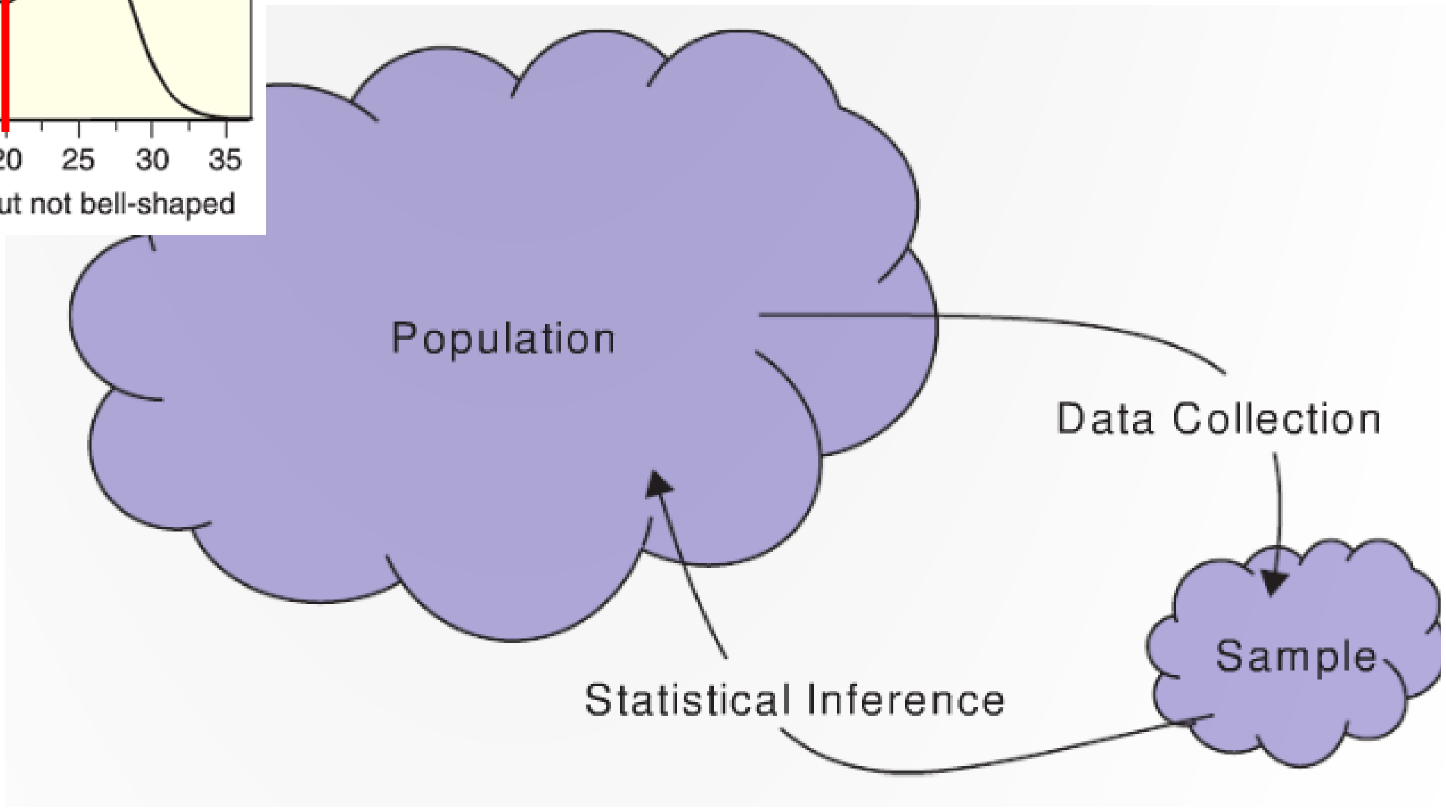
The mean of the ***population*** is denoted μ

The mean of a ***sample*** is denoted $\bar{x}$

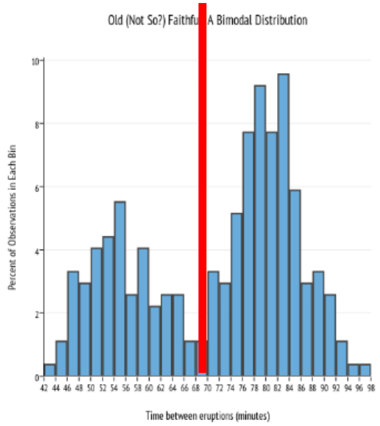
Sample and population mean

μ





$\bar{x}$



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** is a value that splits the data in half

- i.e., half the values in the data are smaller than the median and half are larger

To calculate the median for a data sample of size *n*, sort the data and then:

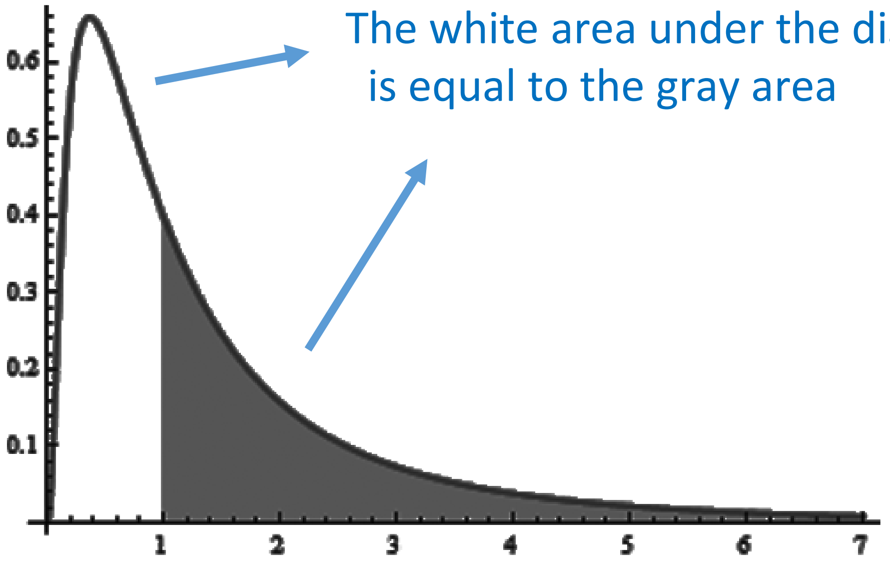
- If n is odd: The middle value of the sorted data

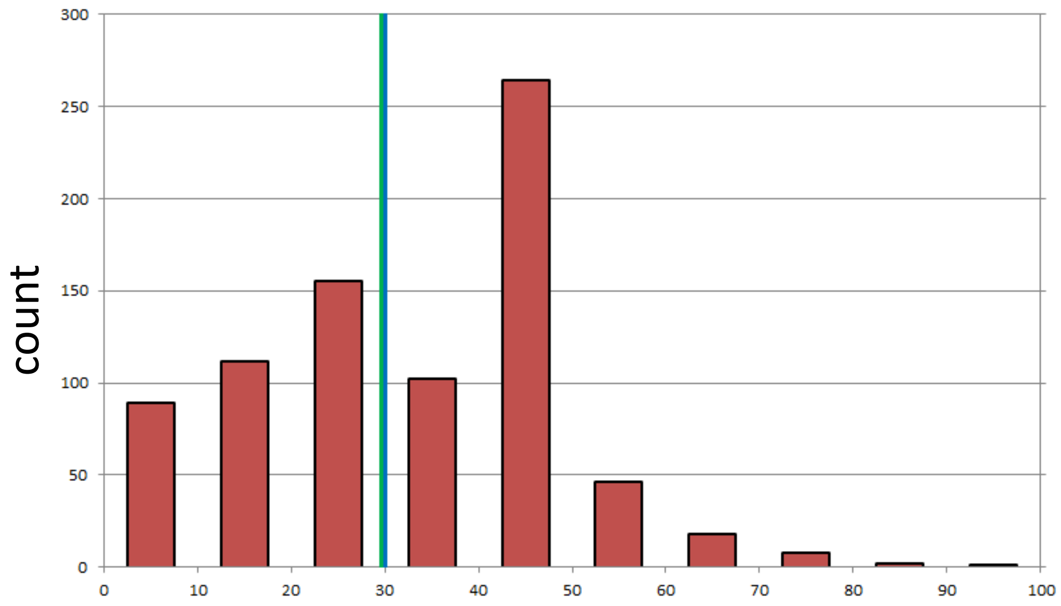
- If n is even: The average of the middle two values of the sorted data

The median

The sum of the heights of the bars on the left is equal to the sum of the heights of the bars on the right

The white area under the distribution is equal to the gray area






```
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

Example of calculating the mean and median

When an individual visits a webpage a ‘ping’ is generated

Below is a random sample of ping counts from 7 people who pinged a website at least once:

```
12, 45, 6, 4, 158, 10, 59
```

Question: What is the mean and median ping count in this sample?

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

Let’s explore calculating the mean and median in R!