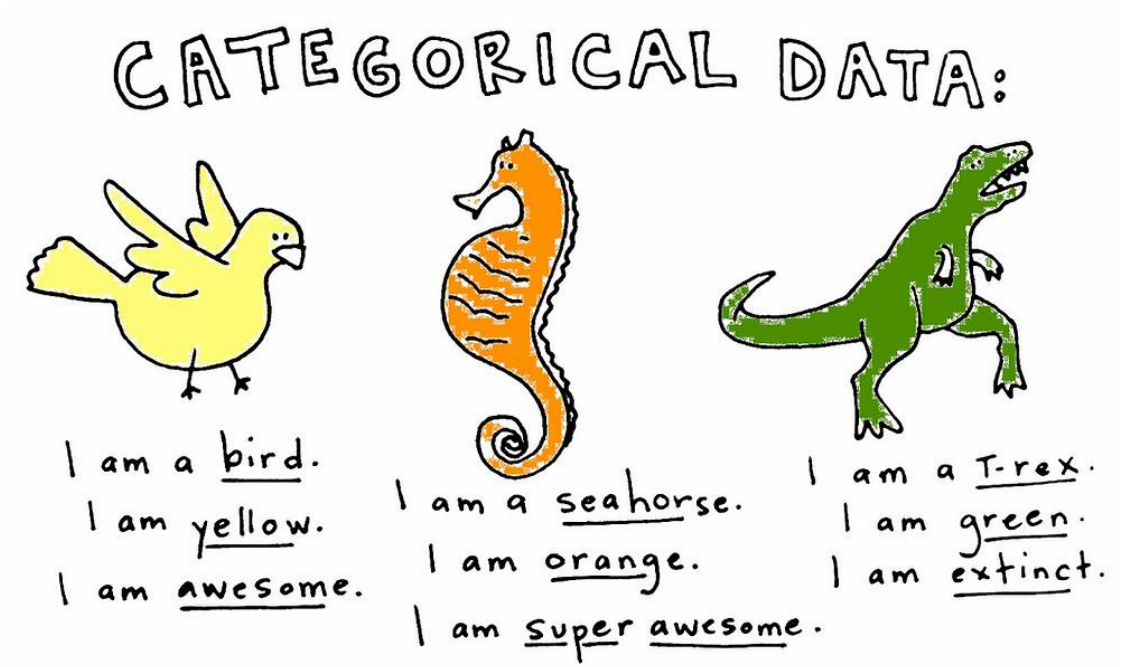
Class 2: Introduction to R and categorical data



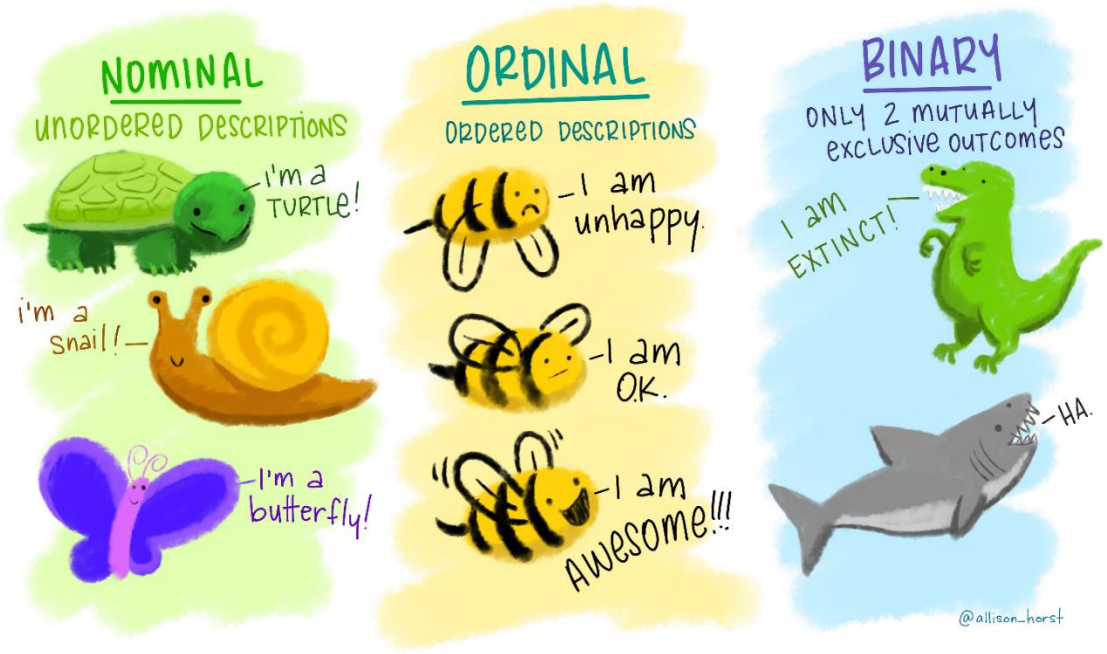
Overview

Quarto documents and a quick review

Introduction to R

If there is time: categorical data

- Proportions
- Frequency tables
- Bar charts and pie plots
- Categorical data in R



Announcement: Practice sessions

The practice sessions will be held in Kline tower room 1105

- 11th Floor of Kline Tower

Addison's sessions

- Tuesday: 3:00–5:00 PM
- Wednesday: 2:00–4:00 PM

Lynda's sessions

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

Announcement: homework 1

If you haven't done so yet, please remember to fill out the background survey under the quizzes on Canvas

Homework 1 has been posted

- Due on Gradescope on Sunday, September 7th at 11pm
- We will discuss how to get the homework soon

Questions?

Getting class code and the homework

To load the class functions use:
> library(SDS1000)

To get the class 2 material, on the console type:
> goto_class(2)

To get homework 1, on the console type:
> goto_homework(1)

The homework file is called: homework_01.qmd

Quarto

Quarto

Quarto (.qmd files) allow you to embed written descriptions, R code and the output of that code into a nice looking document

Creates a way to do reproducible research!

Quarto

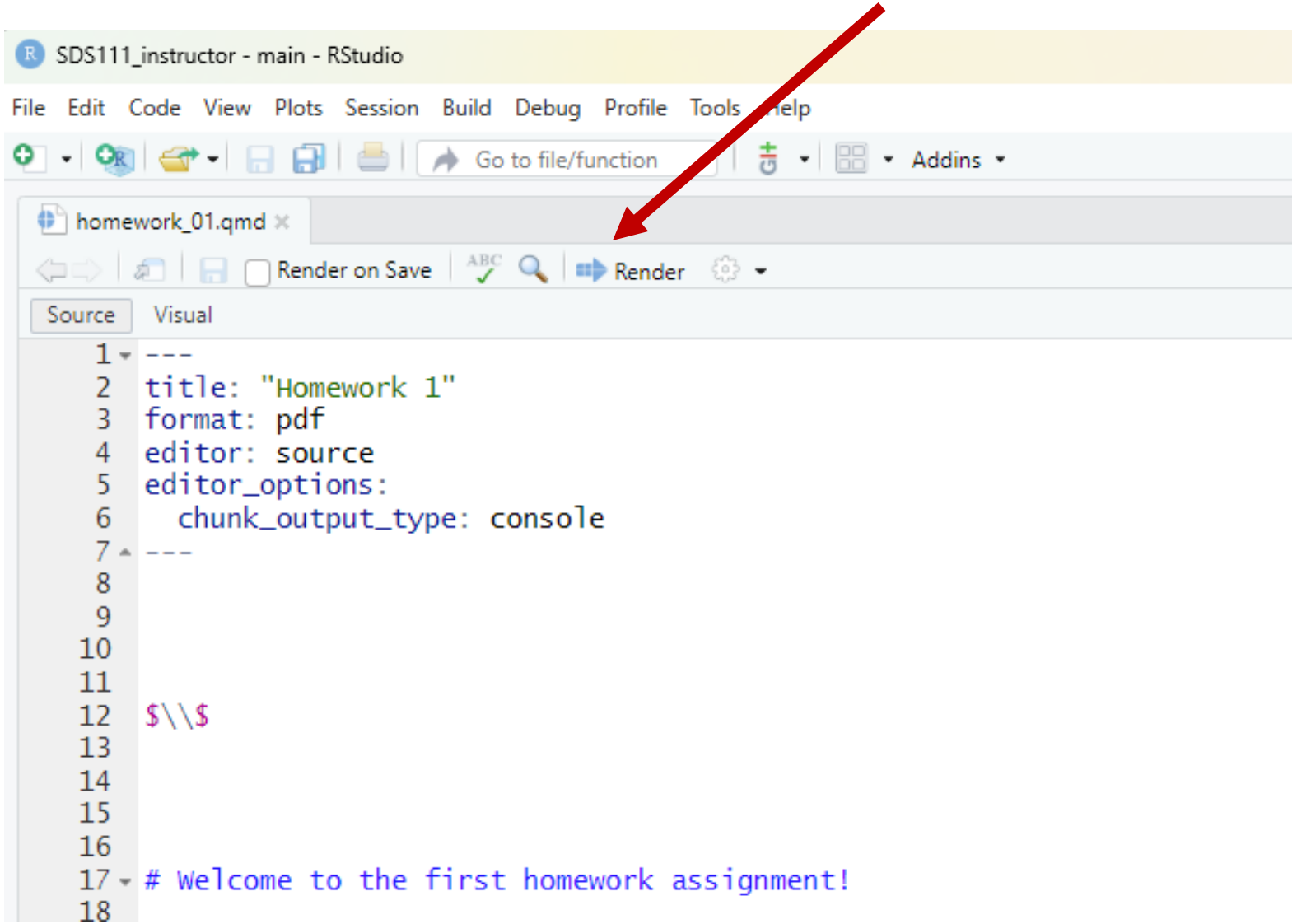
Everything in R chunks is executed as code:

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

Everything outside R chunks appears as text

Render to a pdf

Turns a Quarto document to a pdf



Formatting in Quarto

We can add formatting to text outside the code chunks

Examples:

```
## Level 2 header

**bold**

$\pi$

$x_{outcome}$
```

LaTeX

Avoid hard to debug code!

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

I.e., render your to pdf document often!

Announcement: Homework 1

Due Sunday September 7th at 11pm

- I recommend getting started early on this!

To download the homework please do the following:

> library(SDS1000)

> goto_homework(1)

From the file panel, open the homework_01.qmd and try rendering it to pdf

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!

Questions?


REVIEW

Quiz time! (not to be turned in)

1. What is a population?

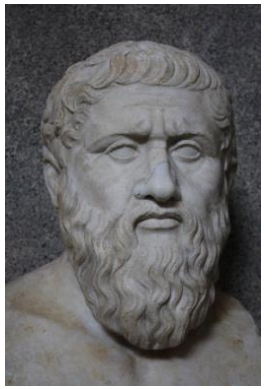
2. What is a sample?

3. What is statistical inference?

4. What are the rows of a data table called

5. What are the columns of a data table called?

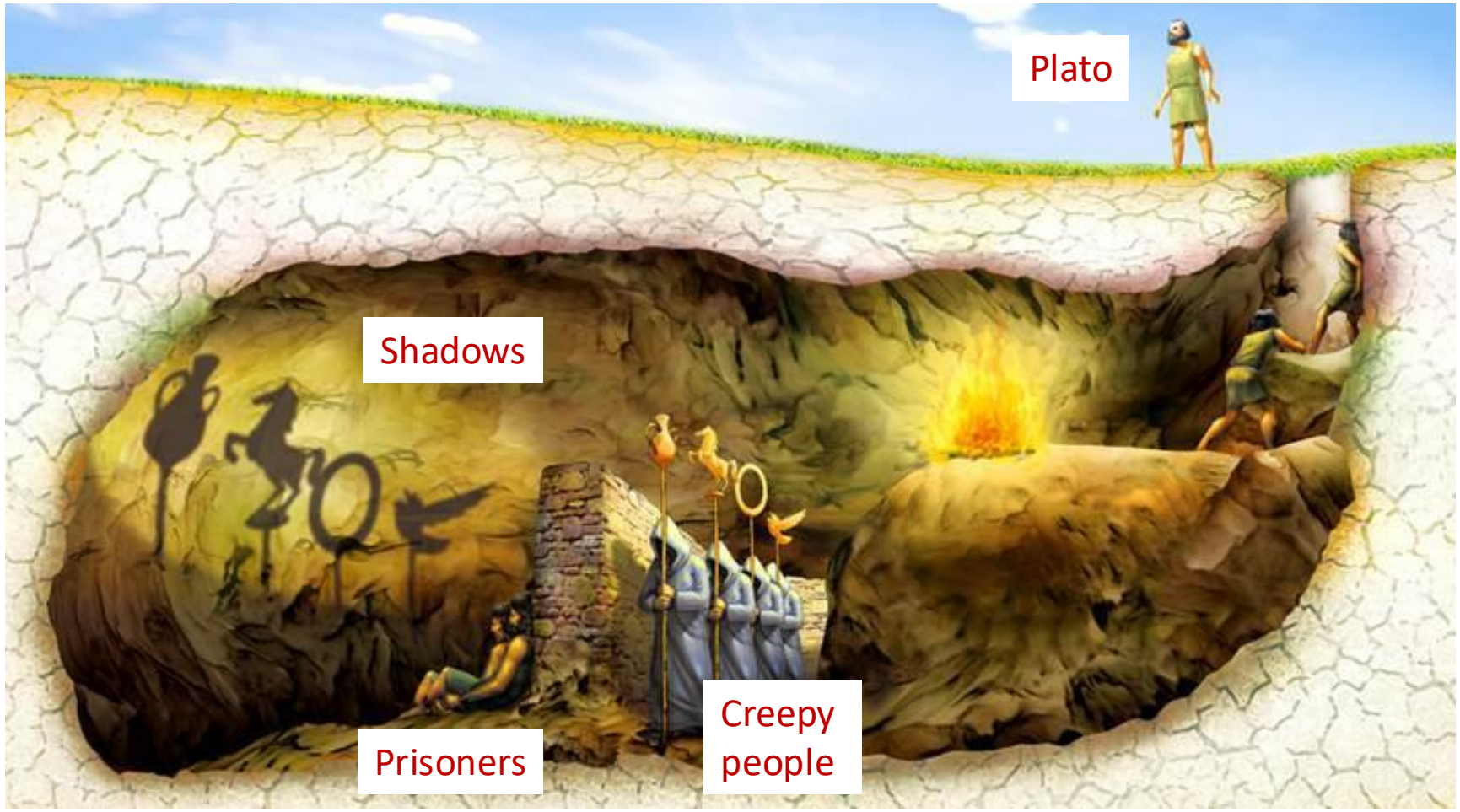
6. What is the difference between categorical and quantitative variables?



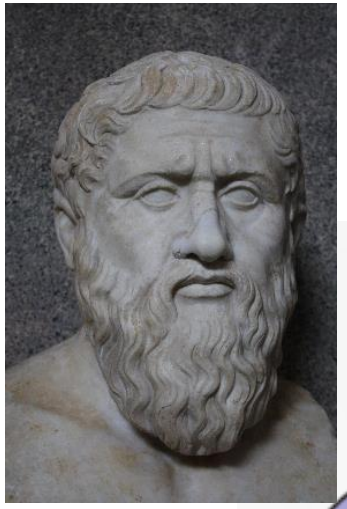
Plato

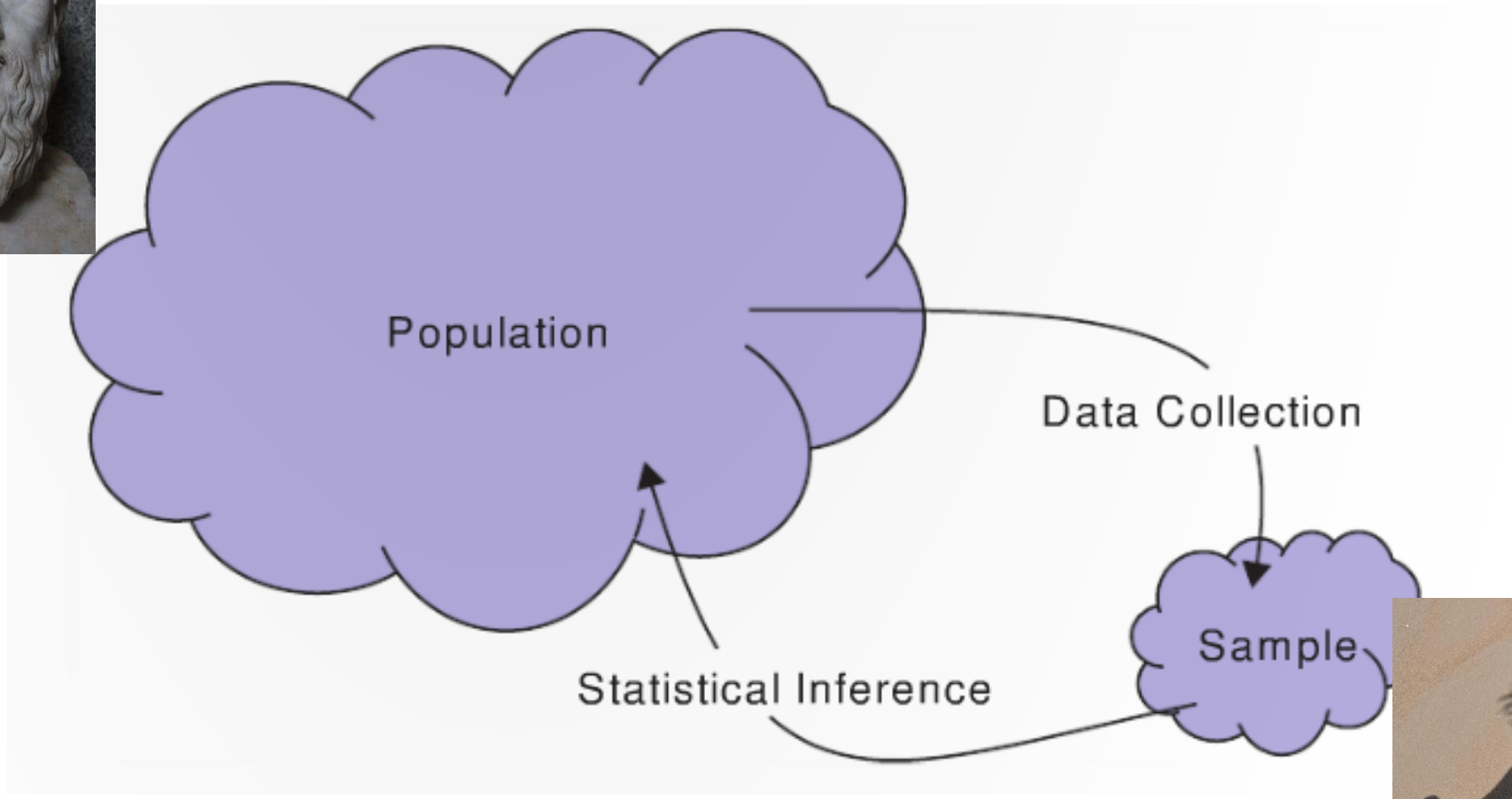
7. Who is this?

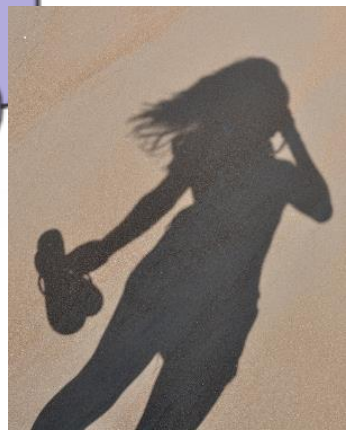
Plato's cave



From The Republic (~ 380 BCE)




Population
Data Collection
Sample
Statistical Inference



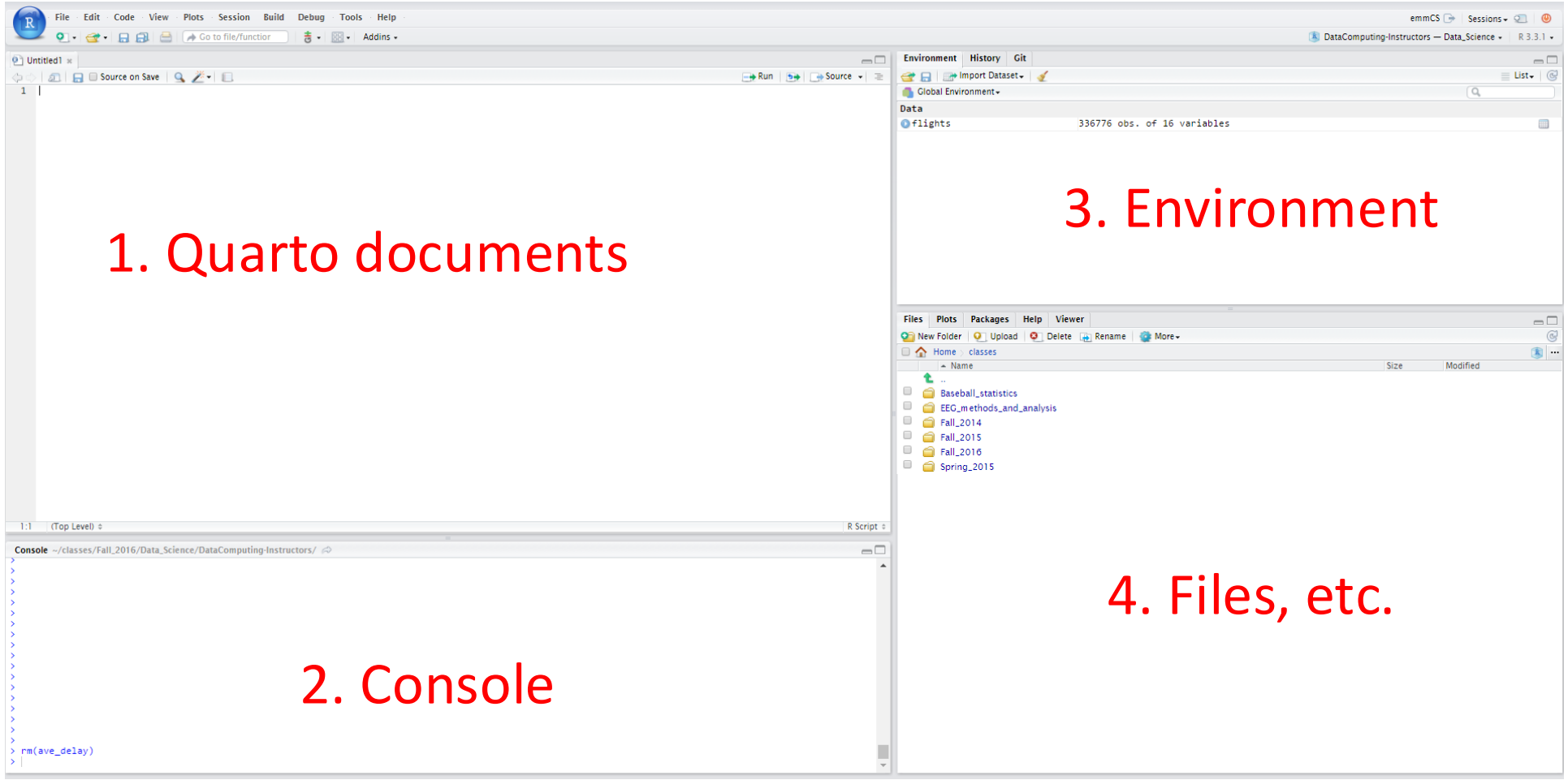
Introduction to

R and R Studio

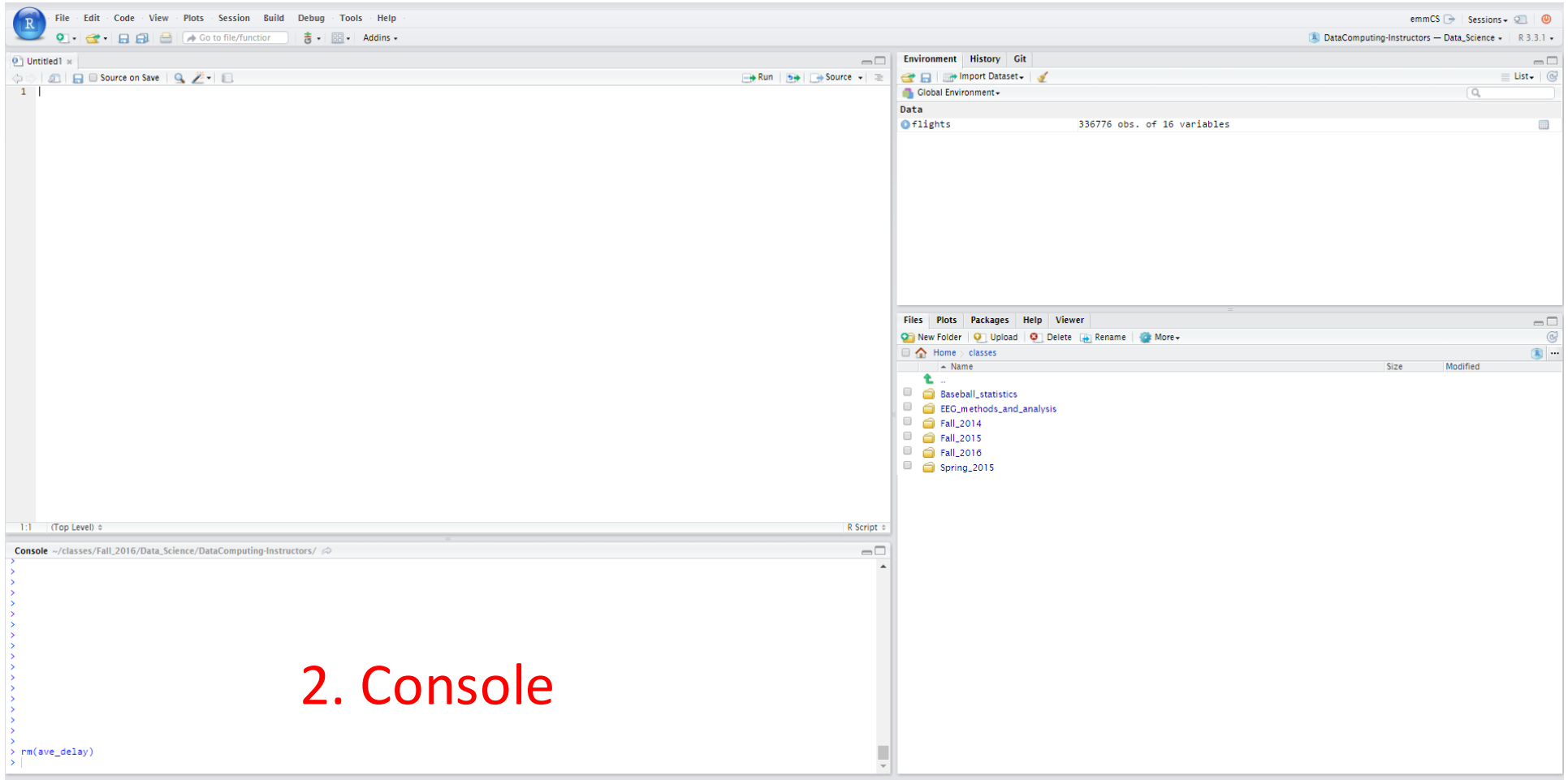
R: Engine

RStudio: Dashboard

RStudio layout



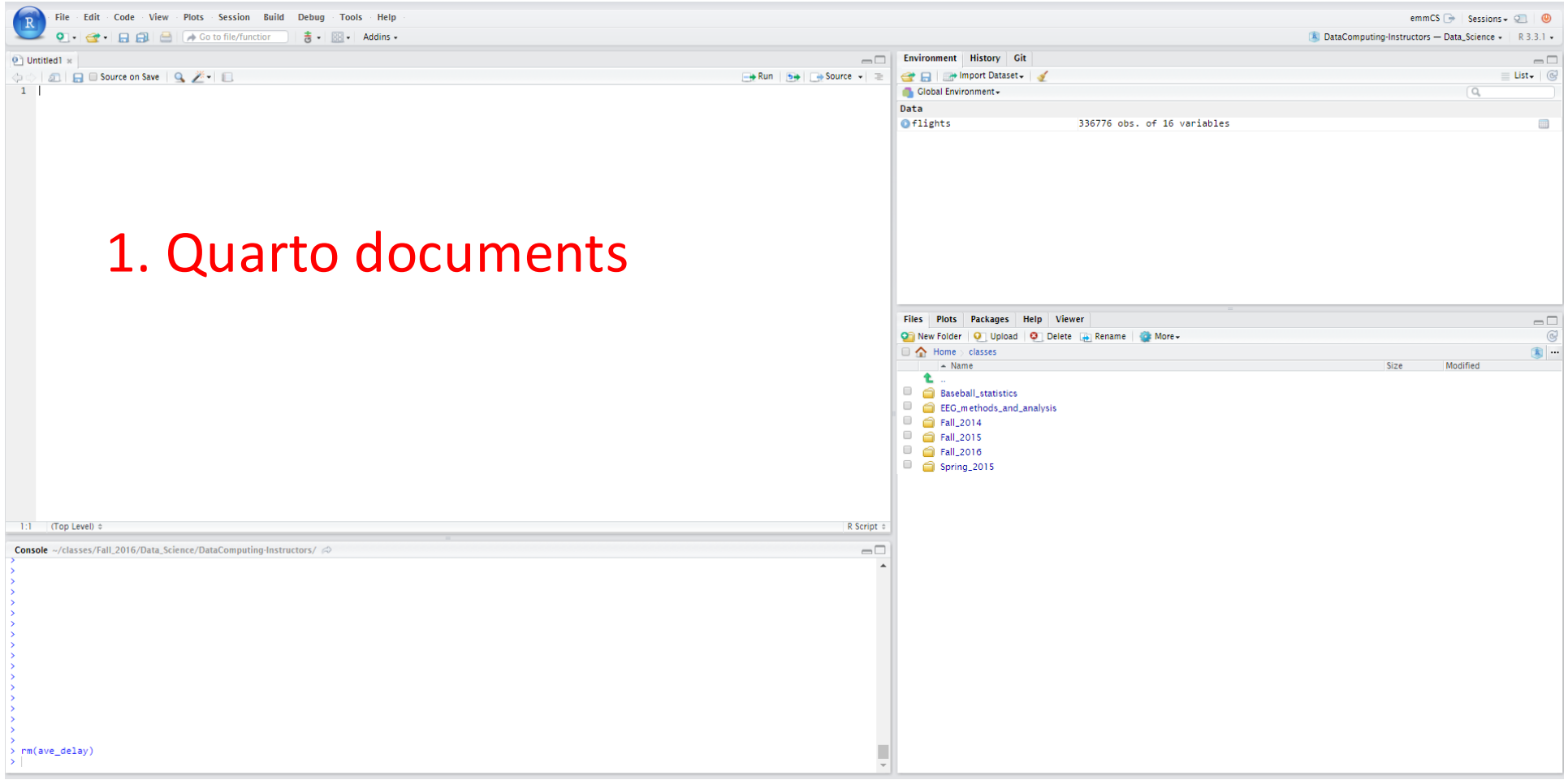
RStudio layout



R as a calculator

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

RStudio layout



1. Quarto documents

R Basics

Please open R Studio and follow along!

Arithmetic:

```
> 2 + 3
> 7 * 5
```

Assignment:

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

Character strings and Booleans

```
> a <- 7
> s <- "Statistics is great!"
> b <- TRUE

> class(a)
[1] numeric

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

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
```

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("these", "are", "strings")
```

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

```
> s[3]        # what will the answer be?
```

Vectors continued

One can assign a sequence of numbers to a vector

```
> z <- 2:10
> z[3]
```

One can test which elements are greater than a value

```
> z > 3
```

We can see how many elements are in a vector using the length() function

```
> length(z)
```

Categorical variables

The sprinkle business (fictional)

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

Sampling example

Questions:

1. What are the observational units (cases)?

2. What is the variable?

3. Is the variable categorical or quantitative?

4. What is the population?

5. Do you think the samples we are getting are representative of the population?

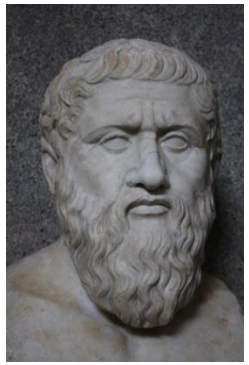
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

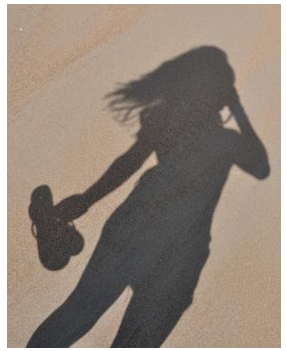
Population parameters vs. sample statistics

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

- Not to be confused with Statistics, which is a field of study

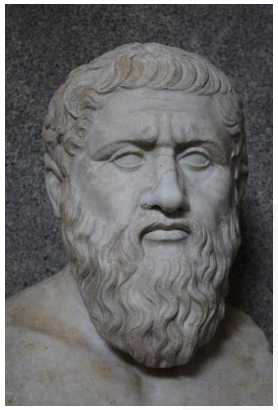
A **parameter** is a number that describes some aspect of a ***population***



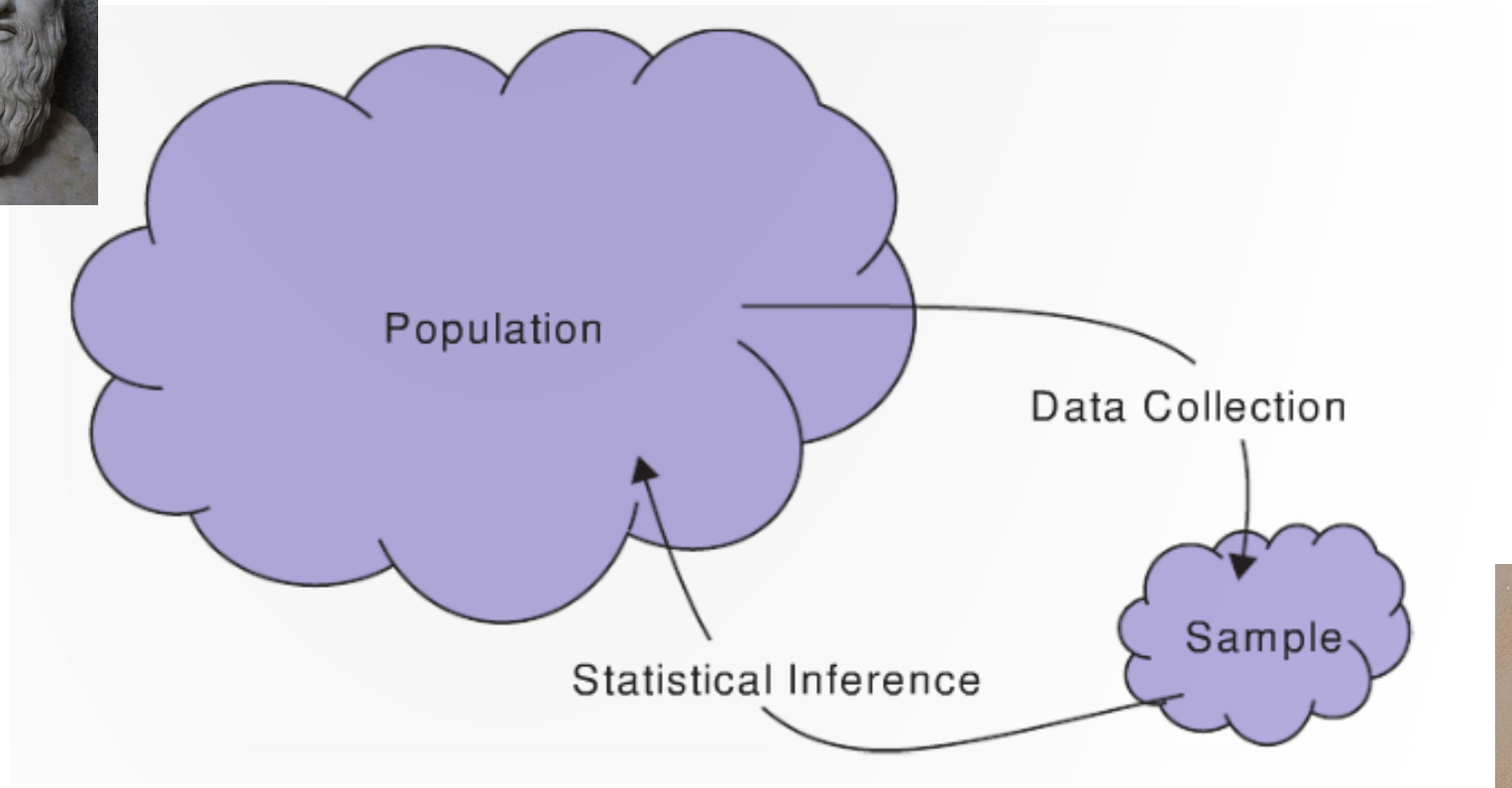


?

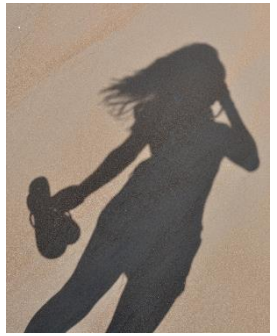
Parameters and statistics



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 $\hat{p}$ (pronounced "p-hat")

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

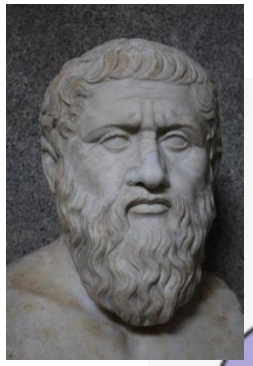
The proportion for a ***population*** is denoted π (the book uses p)

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

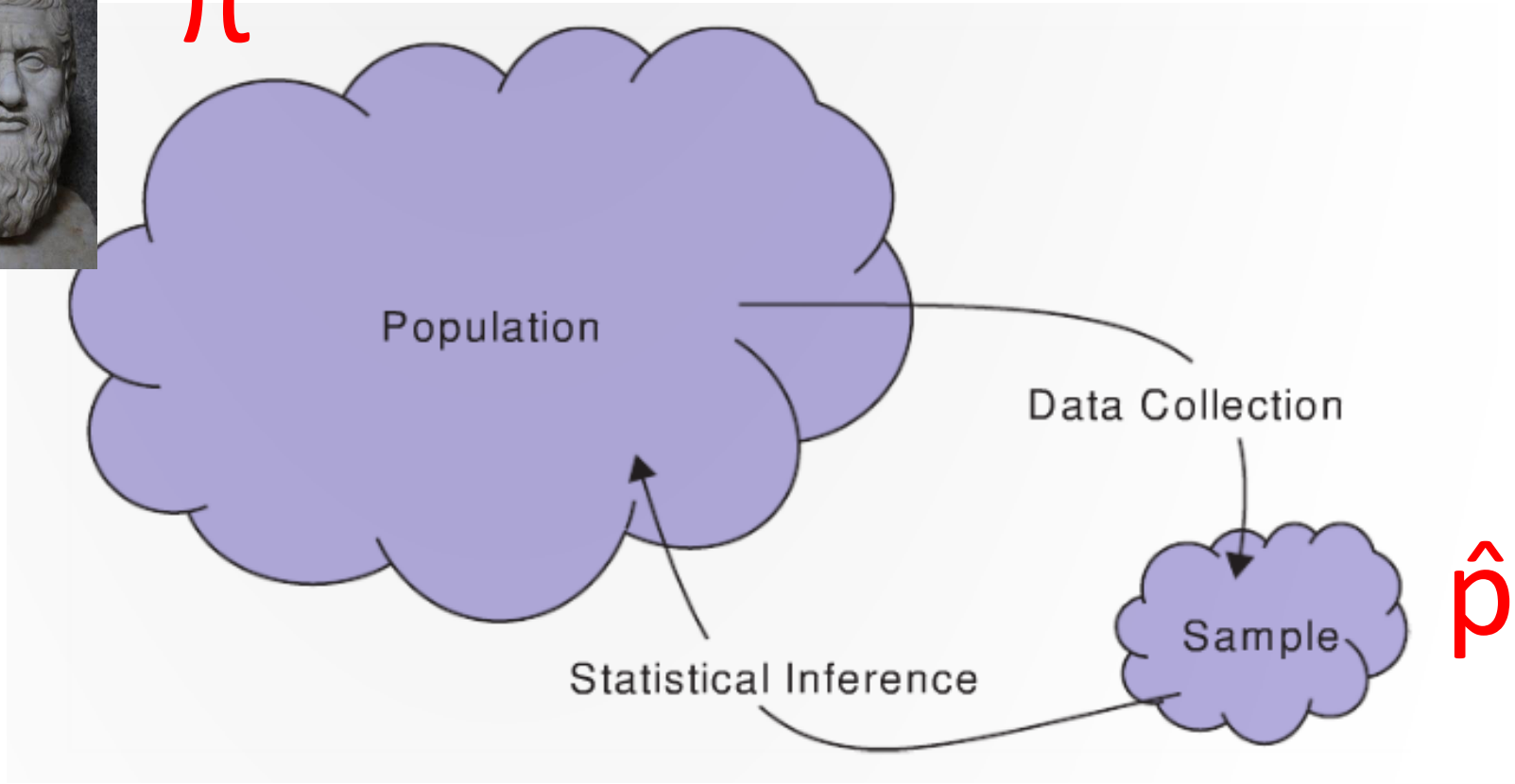
$\hat{p}$ is a ***point estimate*** of π

- 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(v)

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

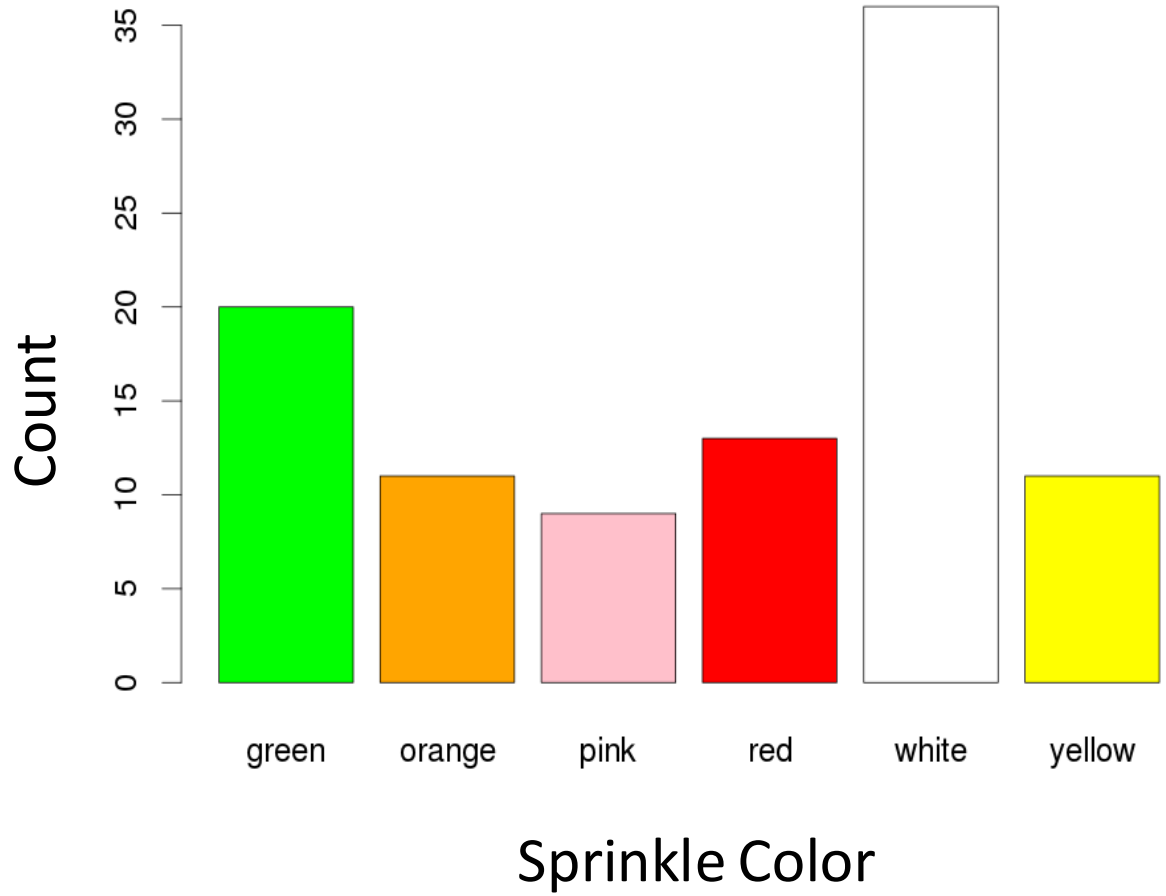
Color	green	orange	pink	red	white	yellow		Total
Count	.20	.11	.09	.13	.36	.11		1

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

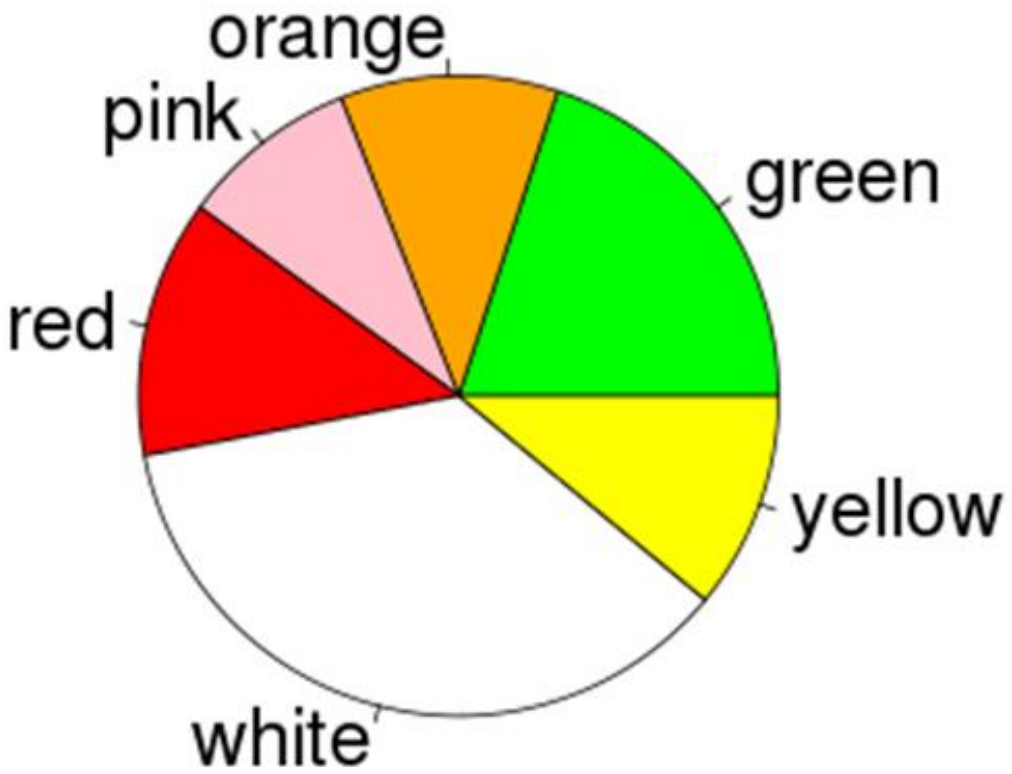


In R: barplot(my_table)

Visualizing categorical data: The pie chart

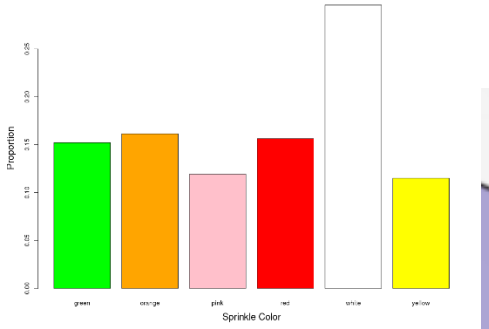
A pie chart plots the proportion of items in each category

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



In R: pie(my_table)

Summary: Sample and Population proportion

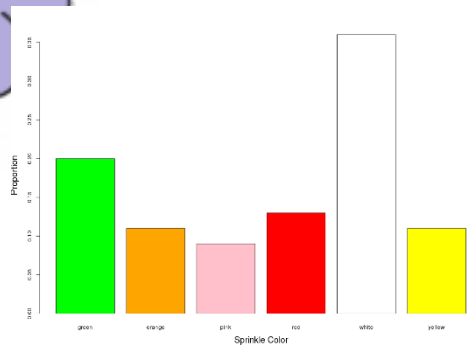


π_{red}

Categorical distribution

Bar chart

$\hat{p}_{red}$



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)
```

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**

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)
```