Class 2

Introduction to R and categorical data

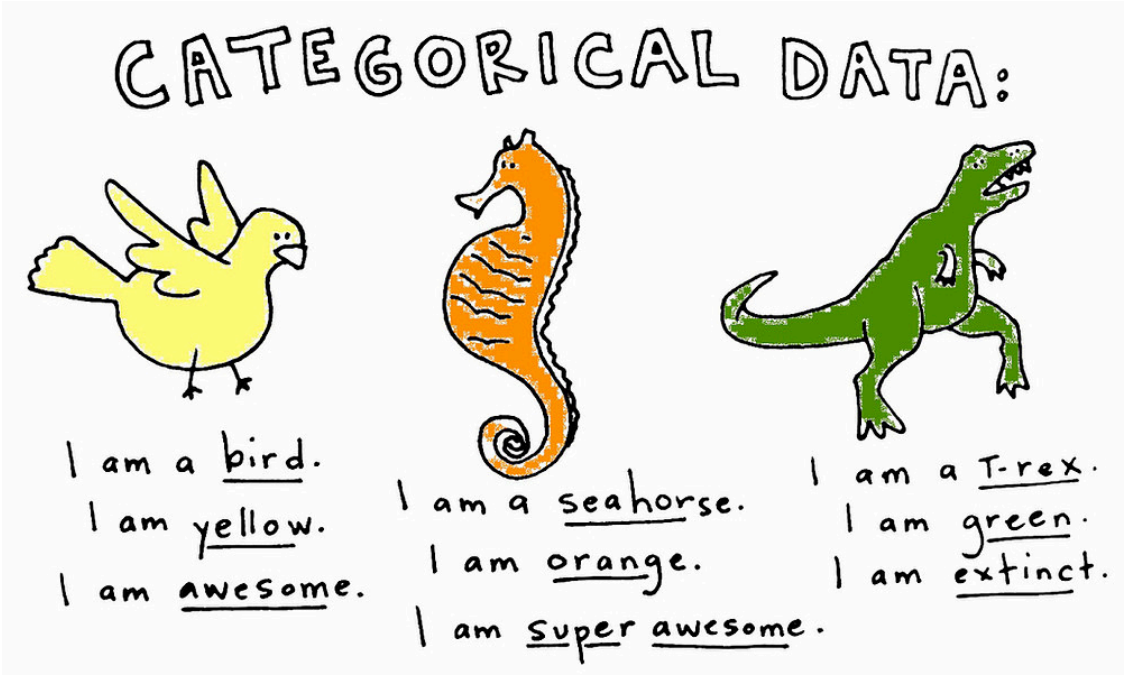
Class 2 topics

Introduction to R

Categorical data

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

If there is time: R Markdown



How is watching the pre-recorded videos going?

Class tip: Rather than scrolling up to watch the videos and answer the quiz questions, open up two tabs on your web browser!

Questions:

- Have you installed R and R Studio?
- Have you gotten the SDS100 package installed and tried homework 0?
- Have you watched the pre-recorded videos for today's class?


ANY
QUESTIONS?


REVIEW

Quiz time! - Do we have these concepts down?

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?**

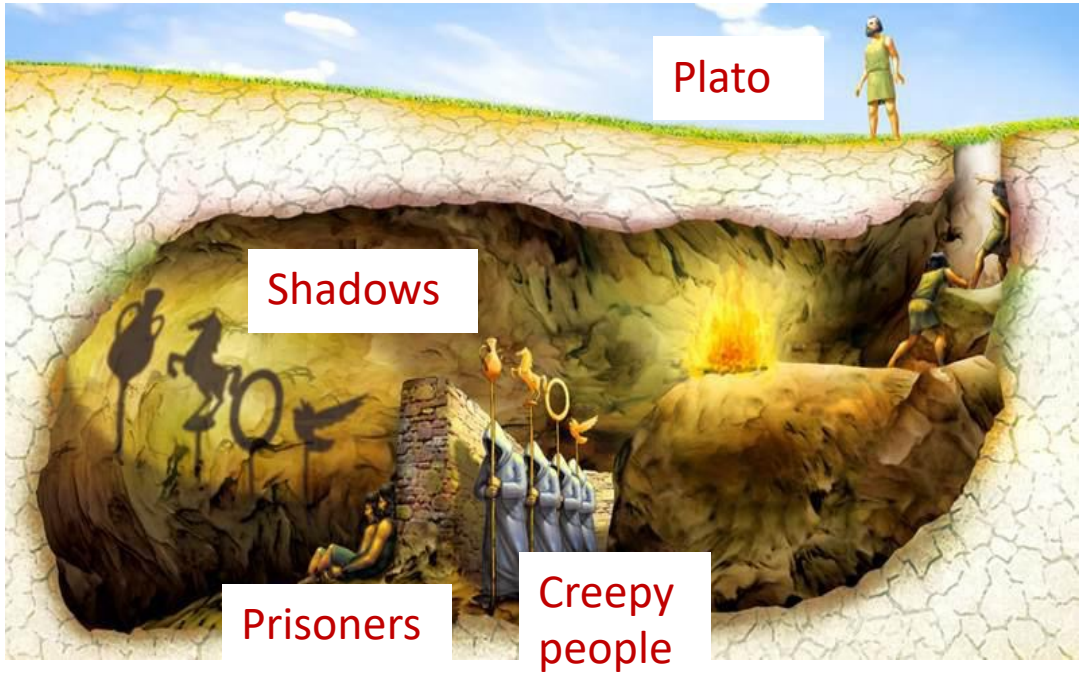
6. **What is the difference between categorical and quantitative variables?**

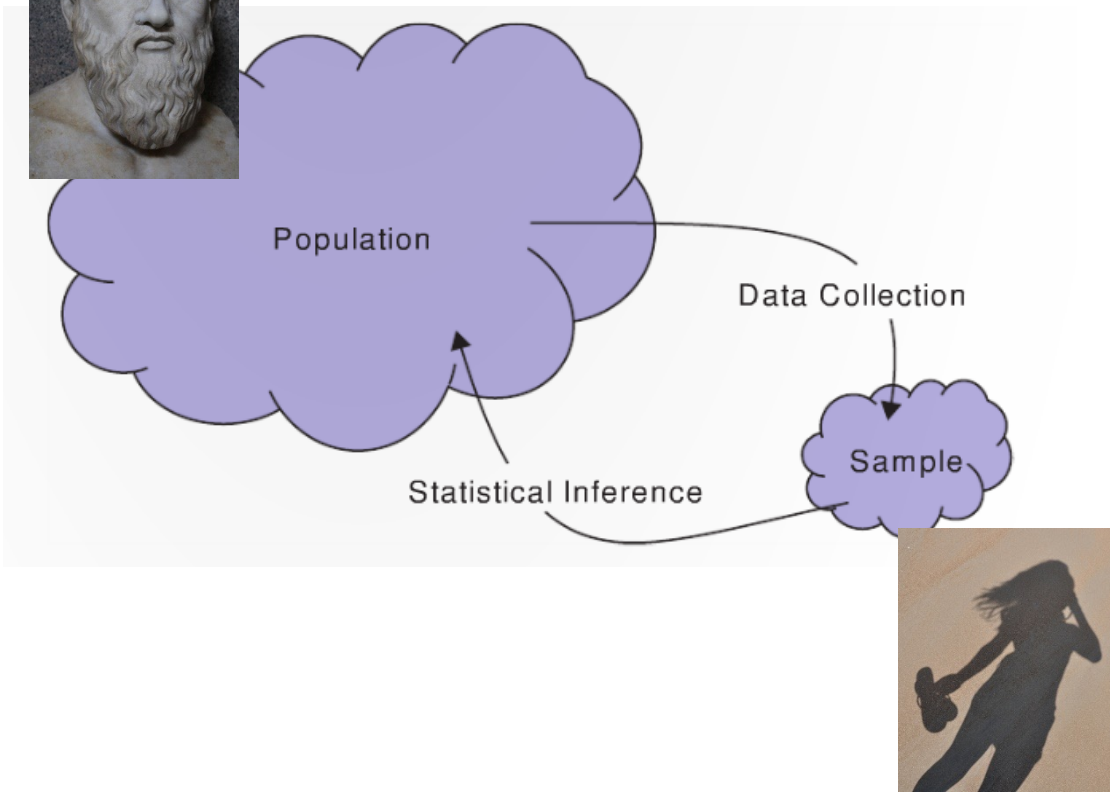
7. **Who is this?**

viktor-crumb:

This is possibly the greatest thing I have seen on the internet.

We feel comfortable with this?





Side note: Jakob Bernoulli, Ars Conjectandi, 1713

"If all events are observed for all eternity…all will occur in certain ratios…Plato himself may have predicted this."


JACOBI BERNOULLI,
Profeſſ. Baſil. & utriuſque Societ. Reg. Scientiar.
Gall. & Pruſſ. Sodal.
MATHEMATICI CELEBERRIMI,

ARS CONJECTANDI,

OPUS POSTHUMUM.

Accedit

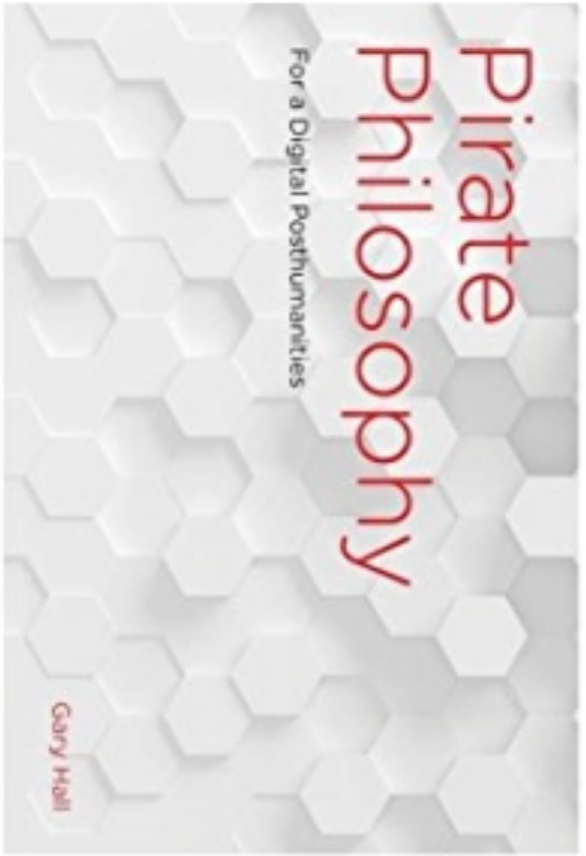
TRACTATUS
DE SERIEBUS INFINITIS,

Et EPISTOLA Gallicè ſcripta
DE LUDO PILÆ
RETICULARIS.

FESTINA LENTE

BASILEÆ,
Impenſis THURNISIORUM, Fratrum.
cIↃ IↃcc xIII.

Question



★★★★☆ 5

Q: How do pirates know that they are pirates?

A: They think, therefore they ARRRR!

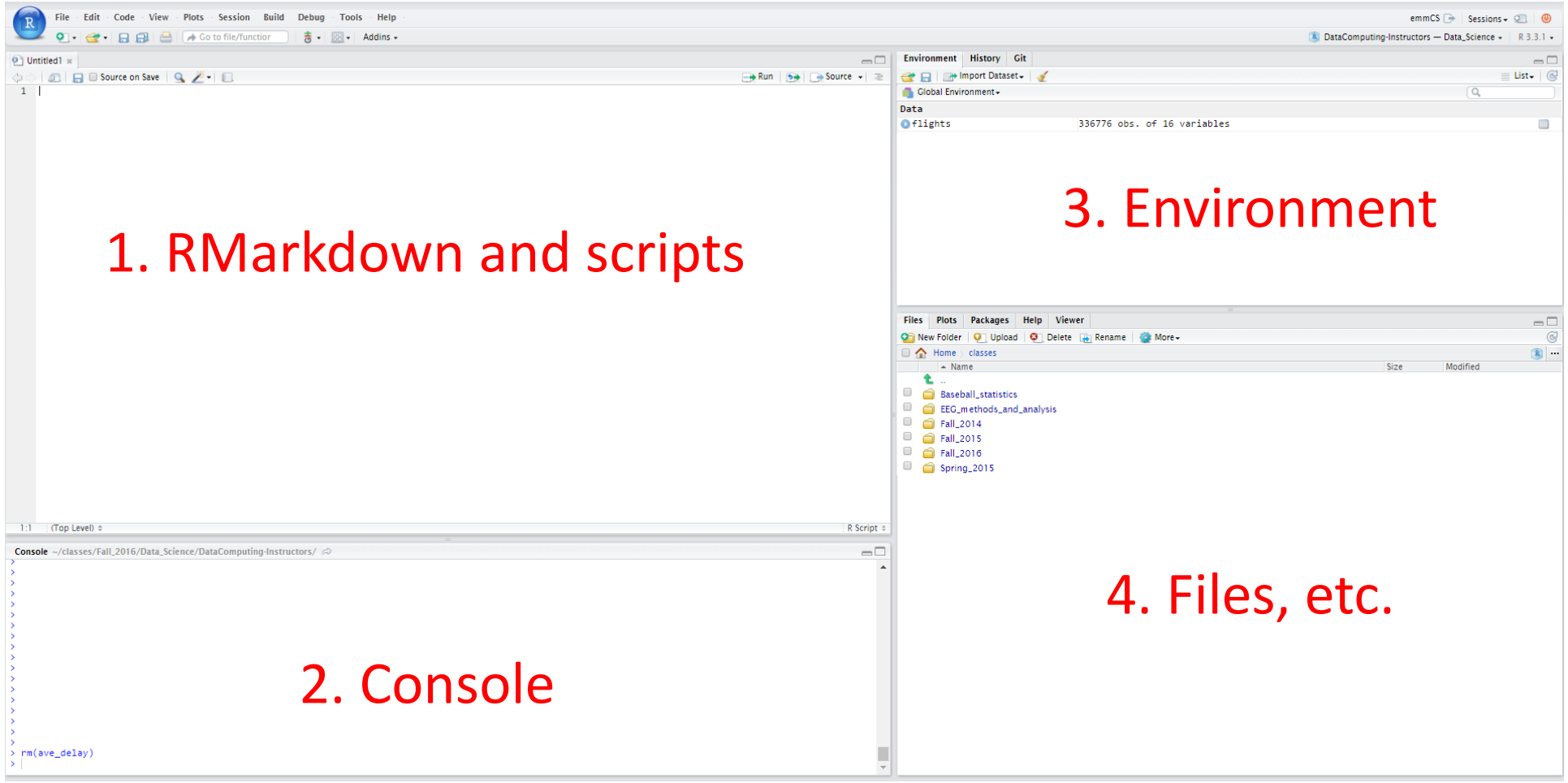
Introduction to

R and R Studio

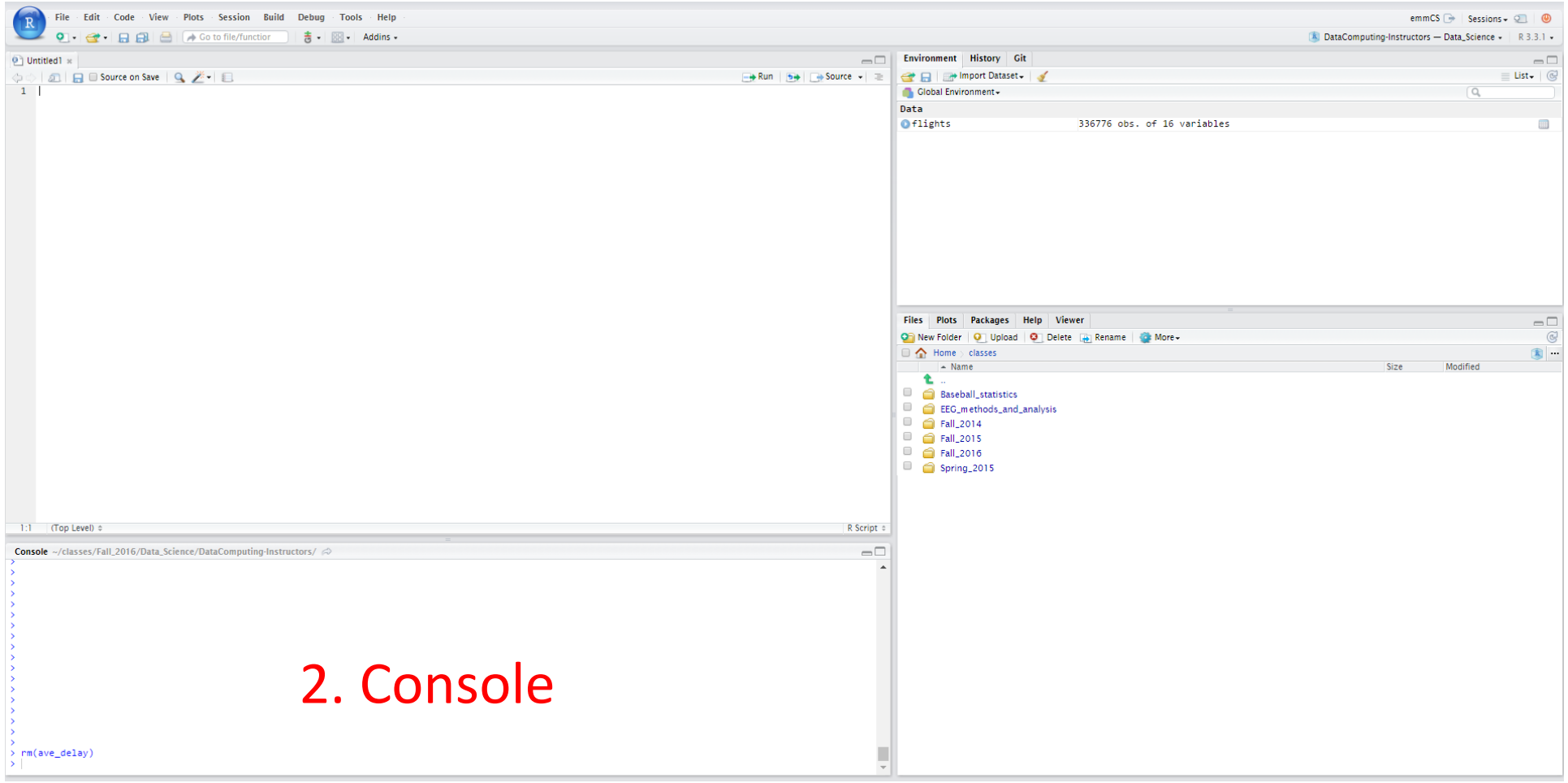
R: Engine

RStudio: Dashboard

RStudio layout



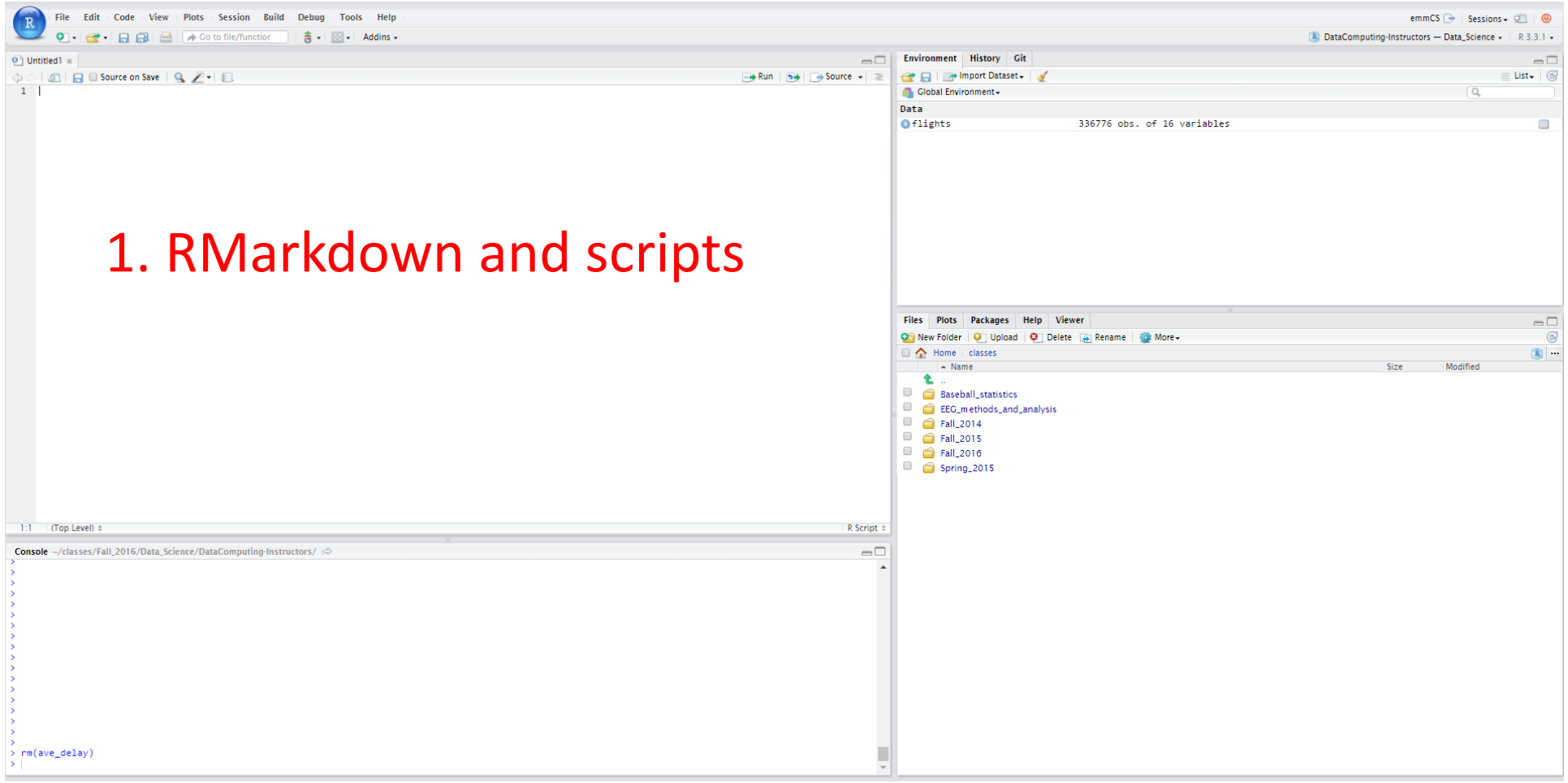
RStudio layout



R as a calculator

```
> 24 * 7
> 360 + 5
```

RStudio layout



R Basics

Please open R Studio and follow along!

If haven't been able to install R yet use

- YCRC R Studio server (link is at the top of the front page of the class Canvas site)
- MyApps: https://rdweb.wvd.microsoft.com/arm/webclient/index.html

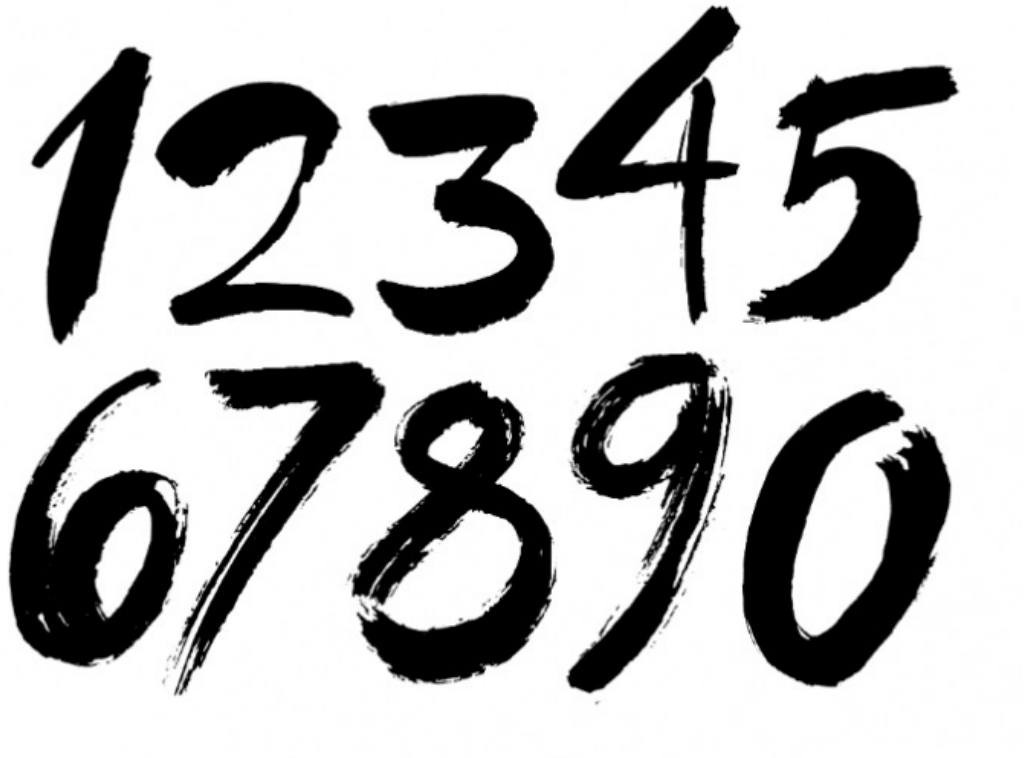
Arithmetic:

```
> 2 + 2
> 21/7
```

Assignment:

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

Number journey…



Review: Character strings and Booleans

```
> a <- 42
> s <- "Plato knows the Truth, which is what we want to know!"
> b <- TRUE

> class(a)
[1] numeric

> class(b)
[1] logical
```

Functions

How can we get help figuring how what a function does?

> ? sqrt

How else can you figure out how to use a function?

Vectors

Vectors are ordered sequences of numbers or letters

The c() function is used to create vectors

```
> v  <-  c(1, 2, 3, 4, 5)
> s  <-  c("I", "love", "Statistics")
```

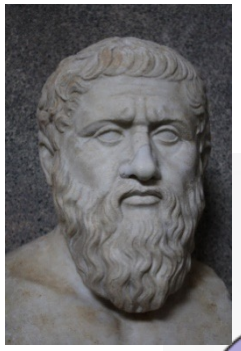
One can access elements of a vector using square brackets []

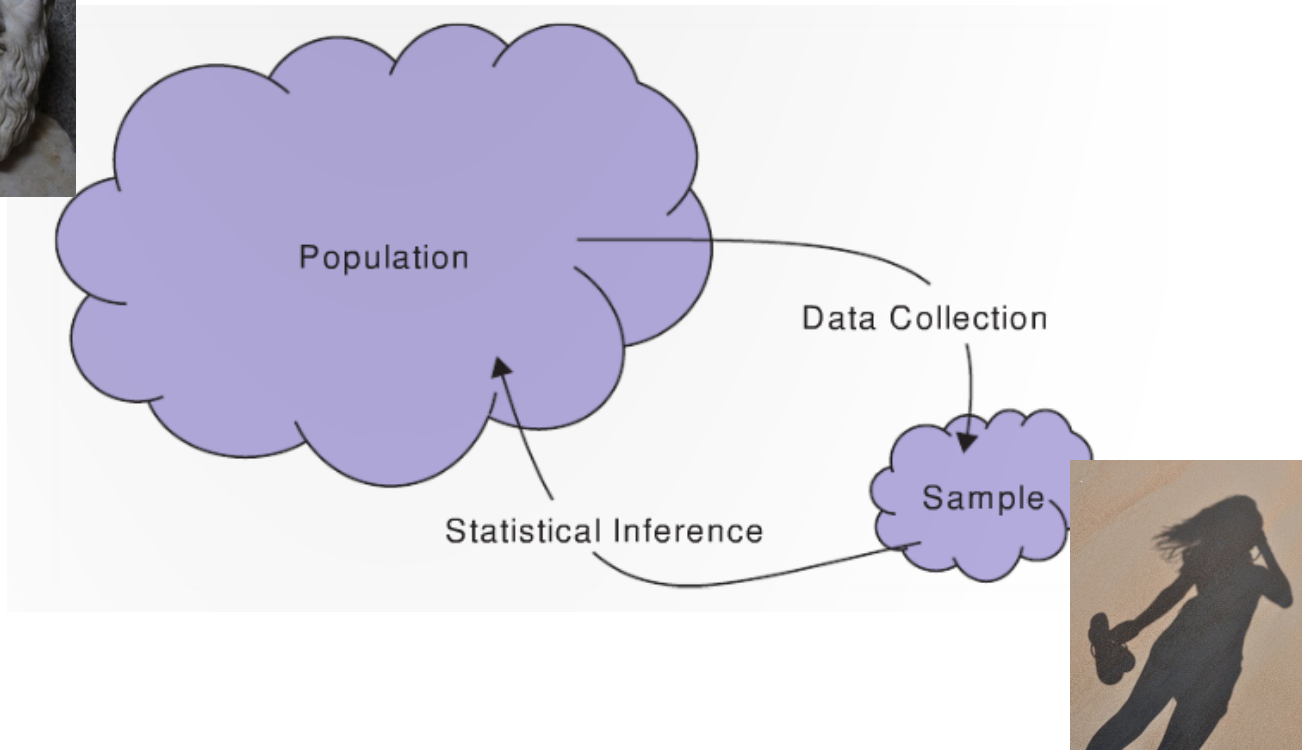
```
> s[2]        # what will the answer be?
```

What will this return?

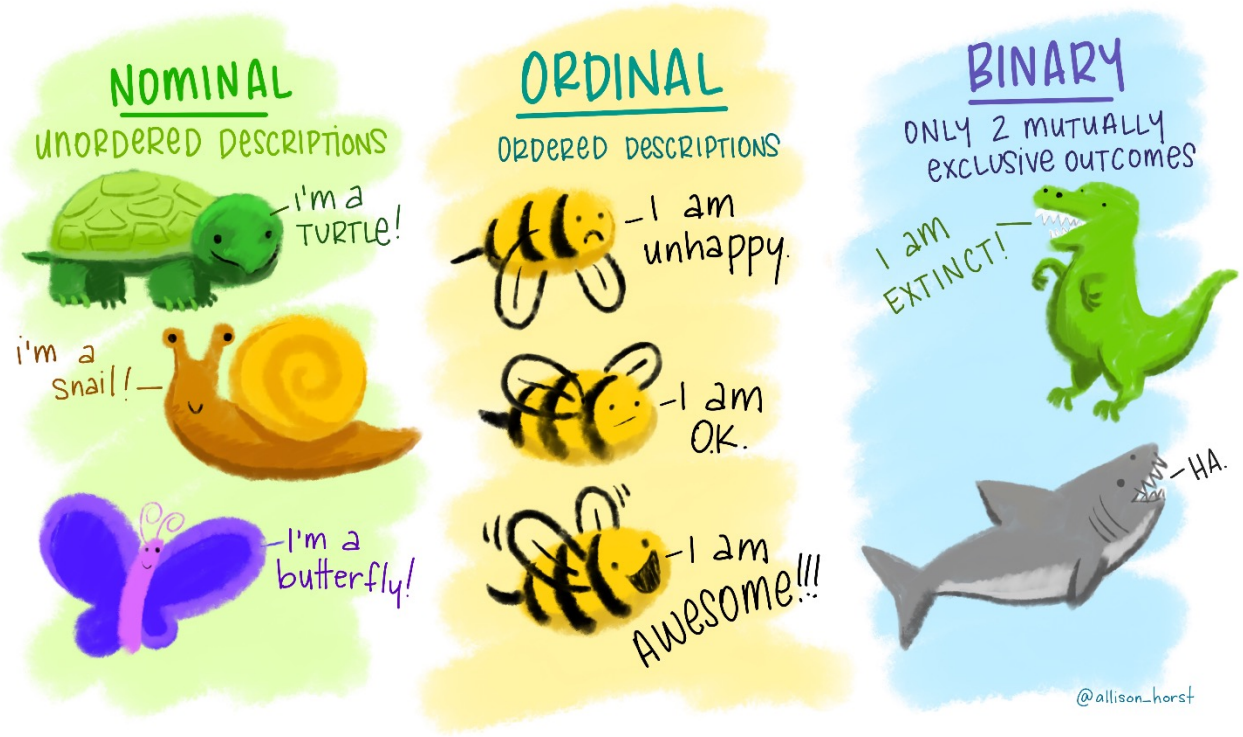
```
> v < 3
```

Now back to fundamental concepts in Statistics…

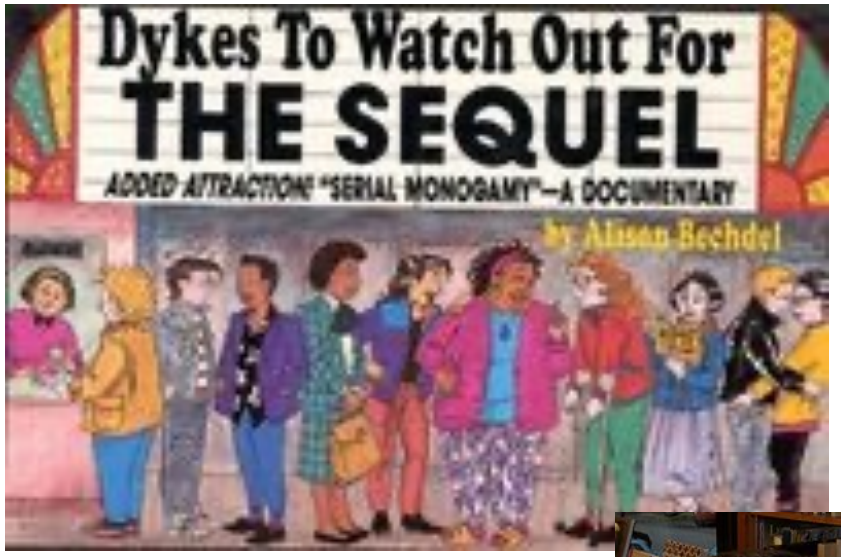




Categorical variables



Motivation: our favorite comic strip from the 80's



The Bechdel test

Where do samples/data come from?

Suppose we had a random sample of 1794 movies

- The ***sample size*** is 1794 (n = 1794)

Sampling example

Variable of interest: Did the movie pass the Bechdel test?

What is the population/process?

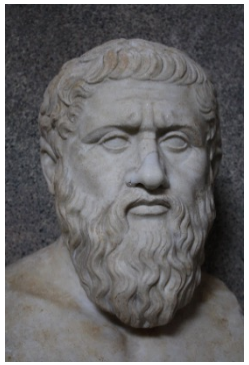
	title	binary
1	21 & Over	FAIL
2	Dredd 3D	PASS
3	12 Years a Slave	FAIL
4	2 Guns	FAIL
5	42	FAIL
6	47 Ronin	FAIL
7	A Good Day to Die Hard	FAIL
8	About Time	PASS
9	Admission	PASS
10	After Earth	FAIL

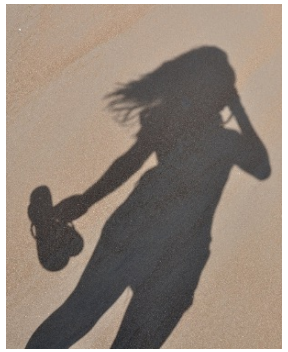
Population parameters vs. sample statistics

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

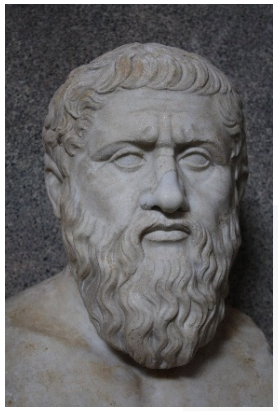
A **parameter** is a number that describes some aspect of a ***population***



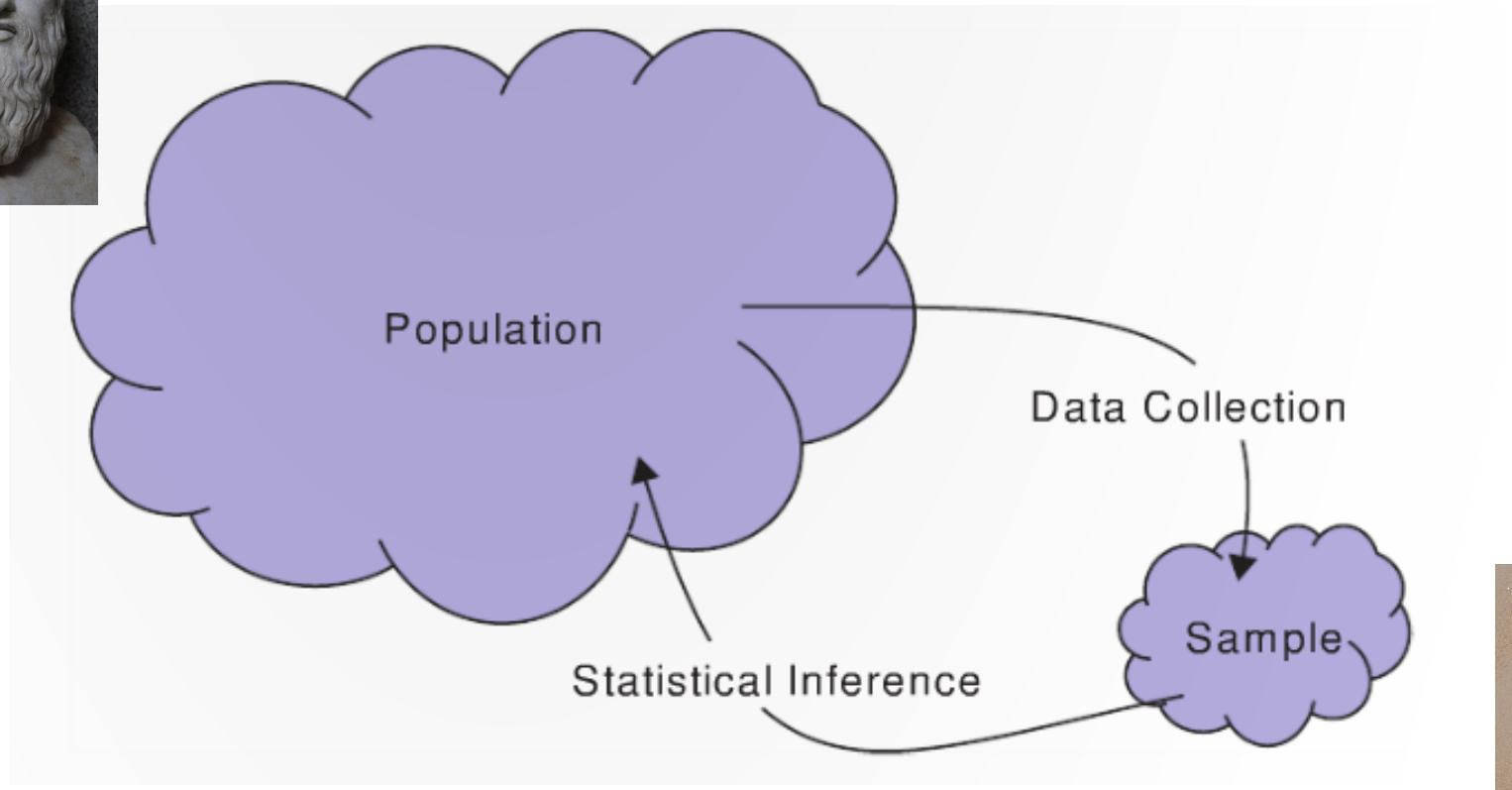


?

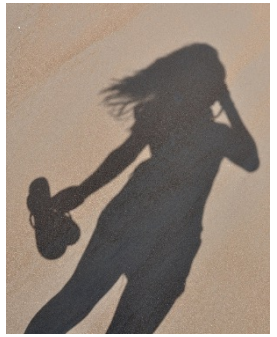
Parameters and statistics



Parameters



statistics



Proportions

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

- E.g., the proportion of movies that pass the Bechdel test

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

Example: proportion of movies passing the Bechdel test

The sample

- PASS, FAIL, PASS, PASS, FAIL, FAIL, …, FAIL

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

- $\hat{p}_{pass} = 803/1794 = 0.448$

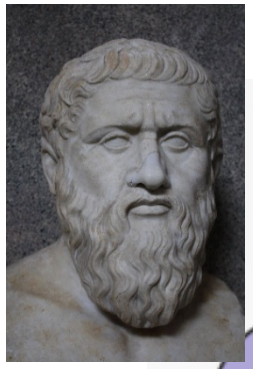
The proportion for a ***population*** is denoted π (the book uses p)

- π_{pass} proportion if we had measured all movies in the population/process

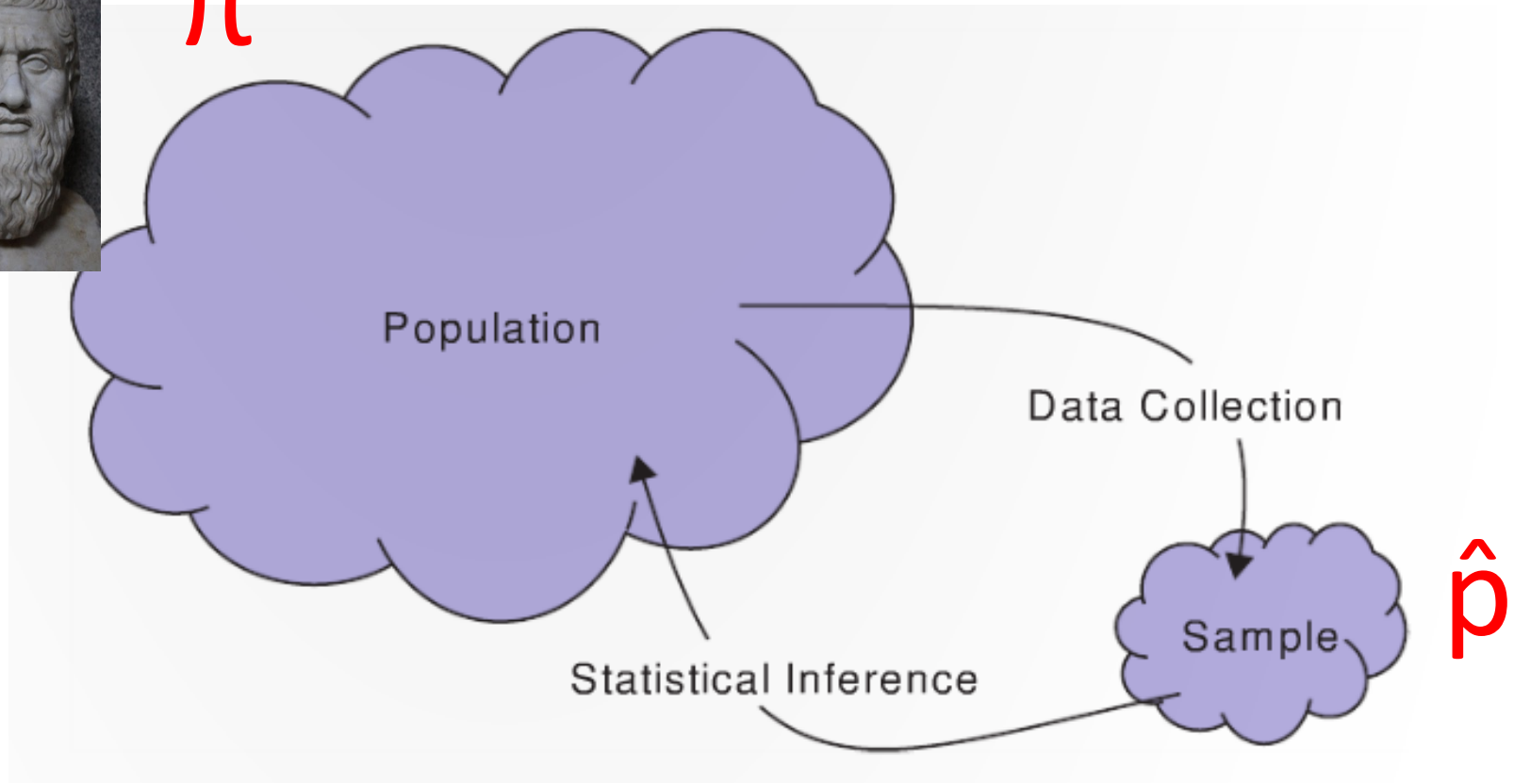
$\hat{p}$ is a ***point estimate*** of π

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

Sample vs. Population proportion



π



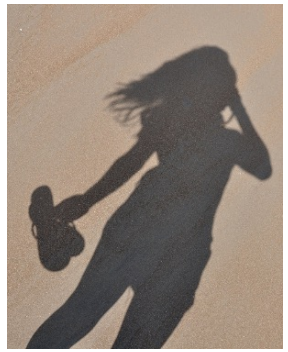
$\hat{p}$

Different samples yield different values for the statistic

$\hat{p}_{s1\_pass} = 0.462$

$\hat{p}_{s2\_pass} = 0.401$

$\hat{p}_{s3\_pass} = 0.498$



Calculating counts on a categorical variable

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

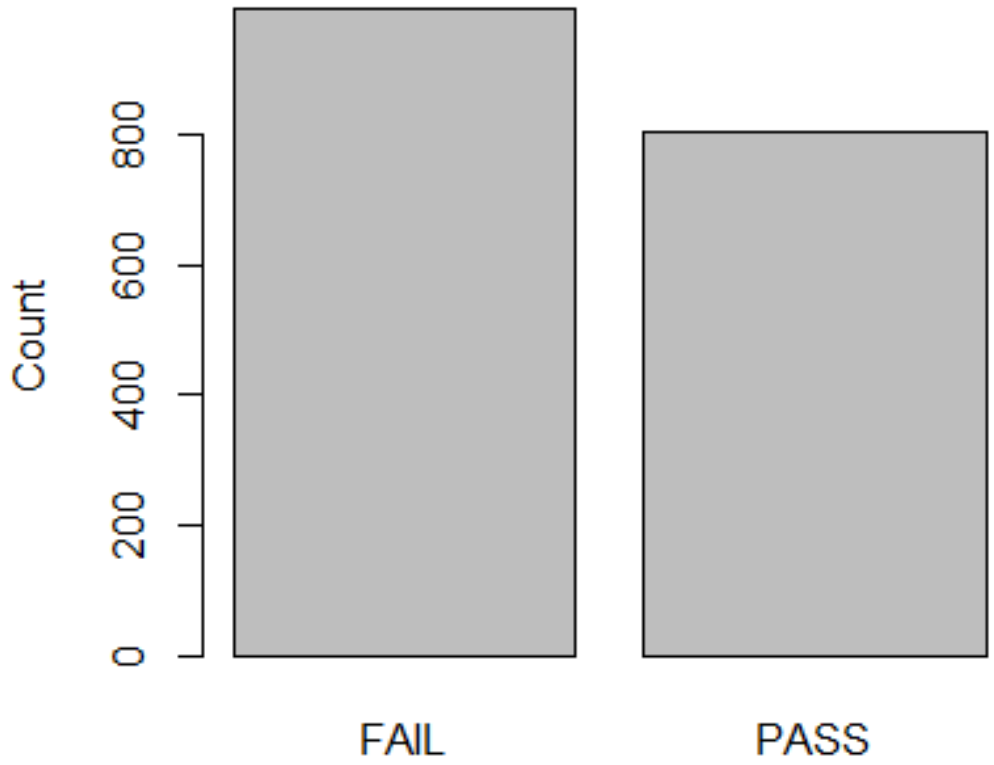
PASS	FAIL		Total
803	991		1794

Calculating proportions (relative frequencies)

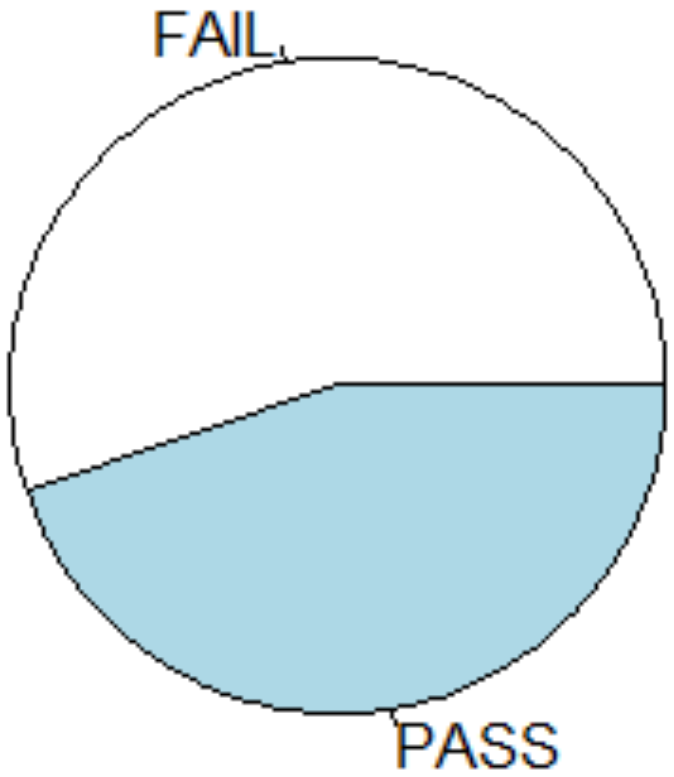
We can convert a frequency table into a ***relative frequency table*** by dividing each cell by the total number of items

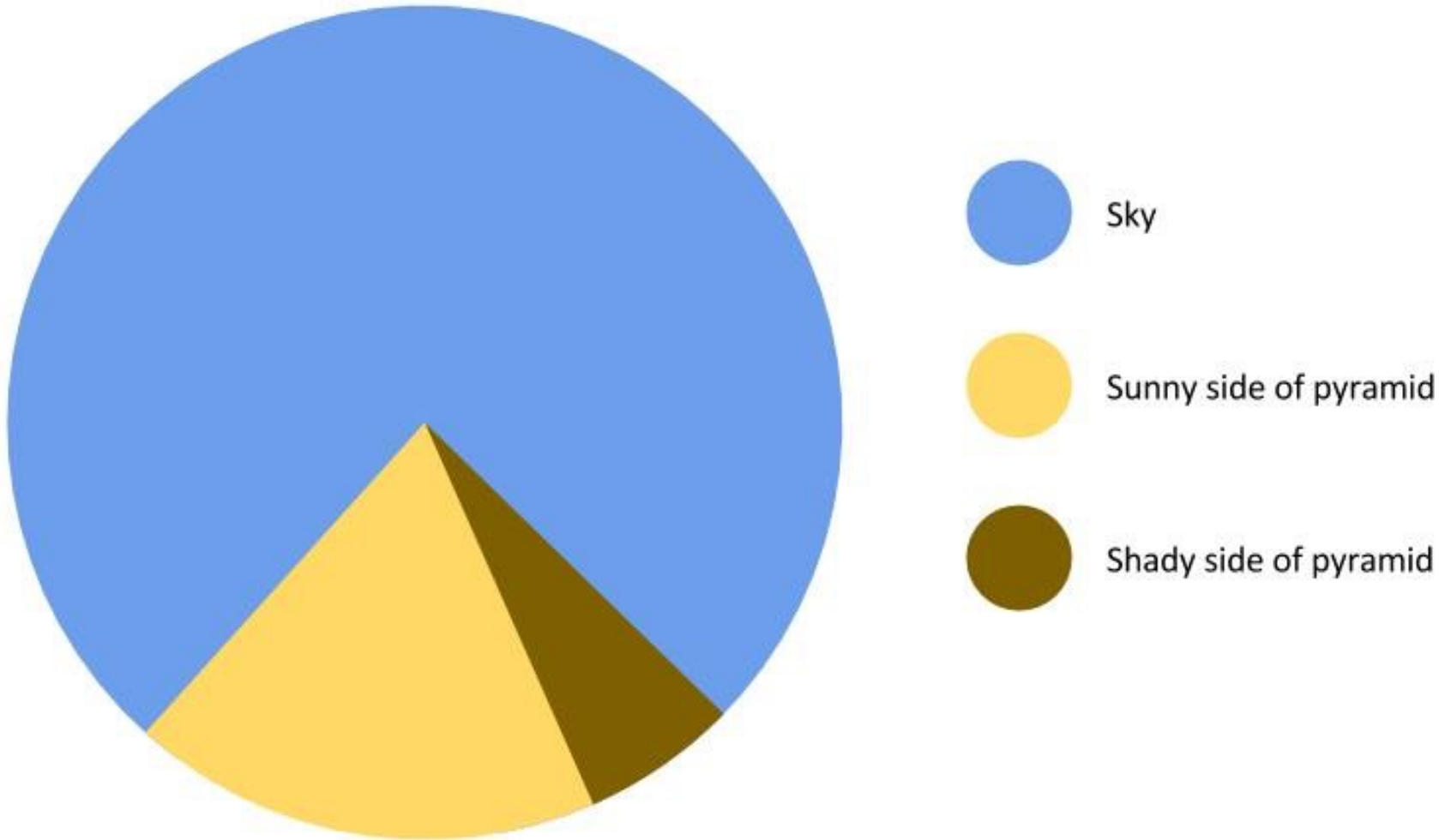
PASS	FAIL		Total
0.552	0.448		1

Visualizing categorical data: The Bar Chart



Visualizing categorical data: The Pie Chart

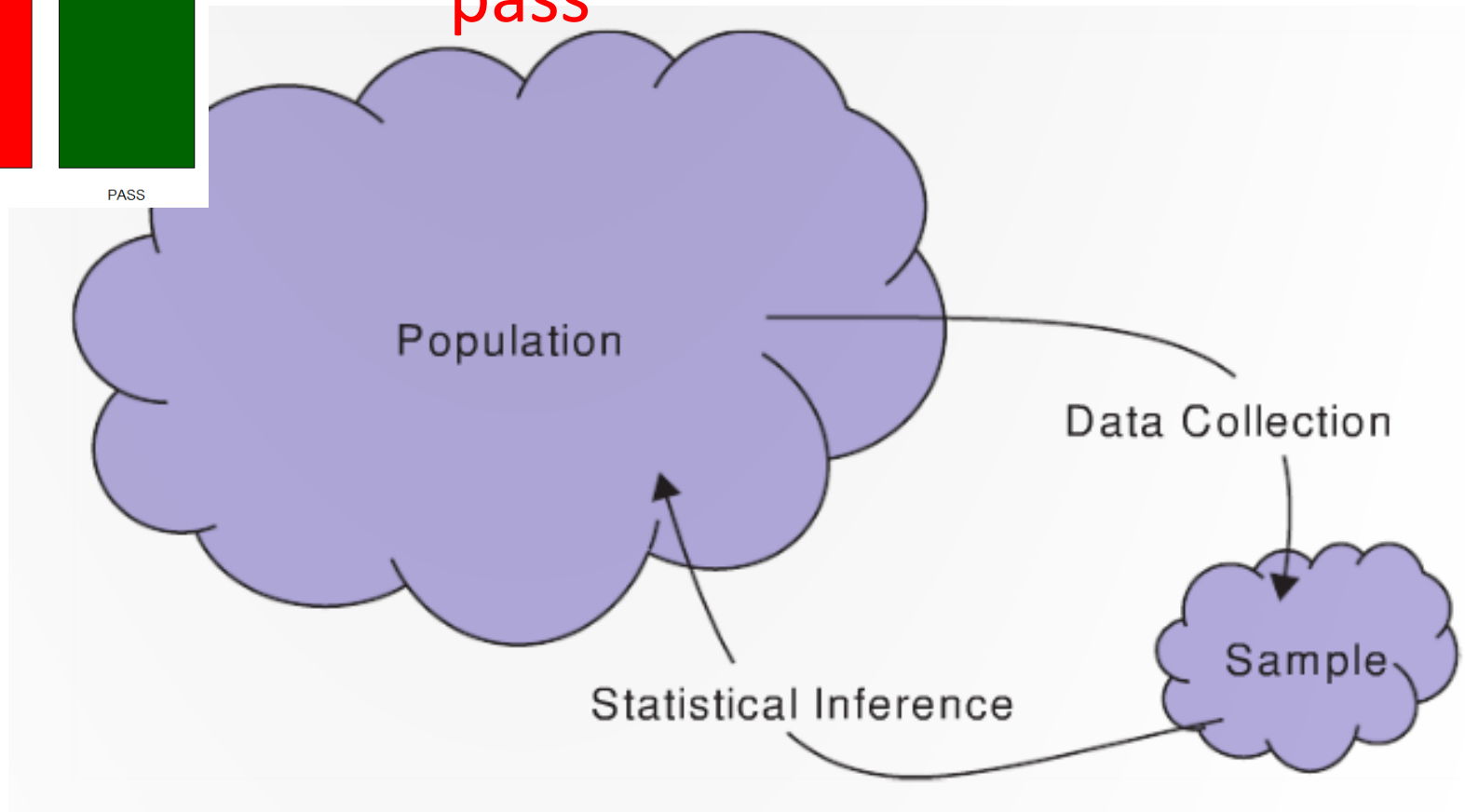



Sky
Sunny side of pyramid
Shady side of pyramid

Summary: Sample and Population proportion

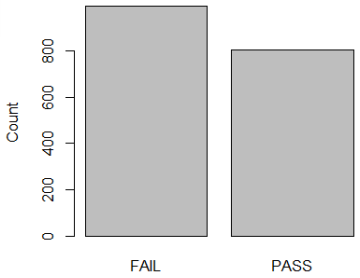
π_{pass}

Categorical distribution



Bar chart

$\hat{p}_{pass}$



Let’s example this in R…

Sampling virtual sprinkles

```
install.packages("fivethirtyeight")
library(fivethirtyeight)

outcome <- bechdel$binary

outcome_count_table <- table(outcomes )
outcome_prop_table <- prop.table(sprinkle_count_table)

barplot(outcome)
pie(outcome)
```

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("PASS", "FAIL", "FAIL", "PASS", "PASS", ... )

# 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)
```


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
```
```

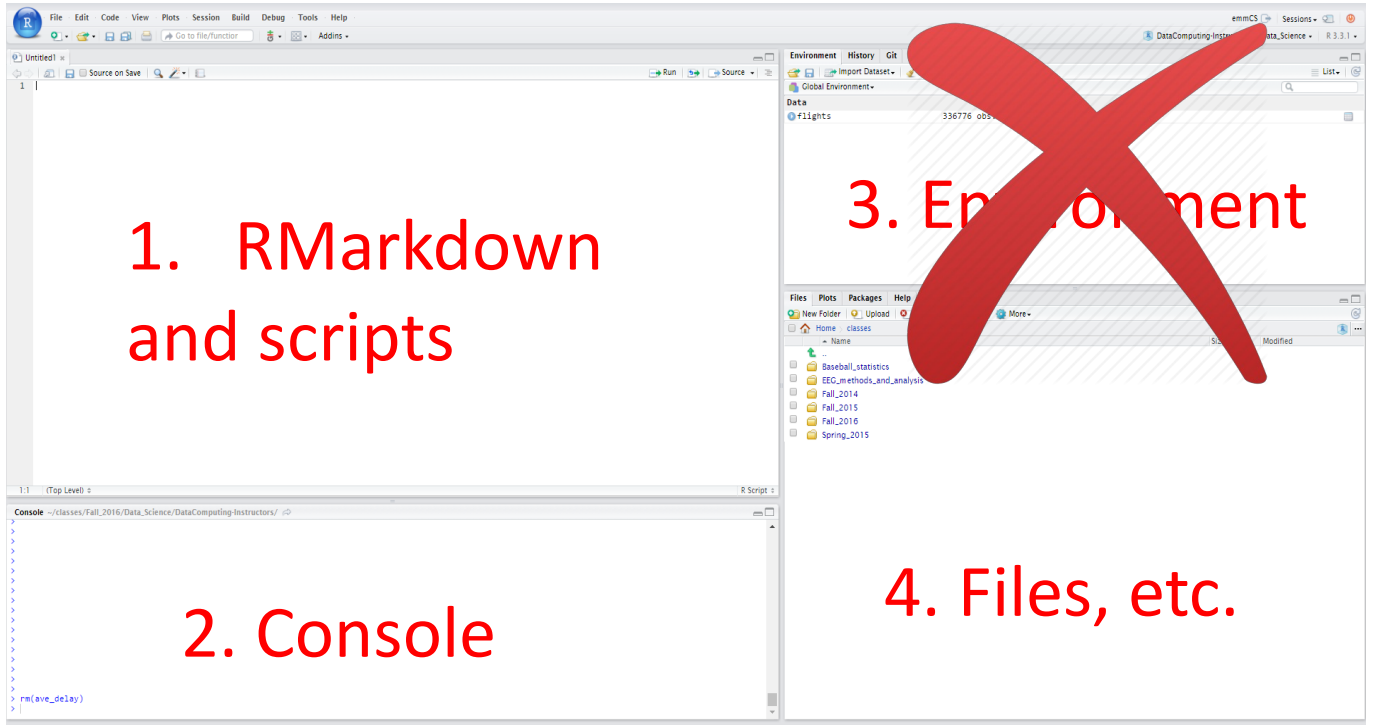
Everything outside R chunks appears as text.

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:

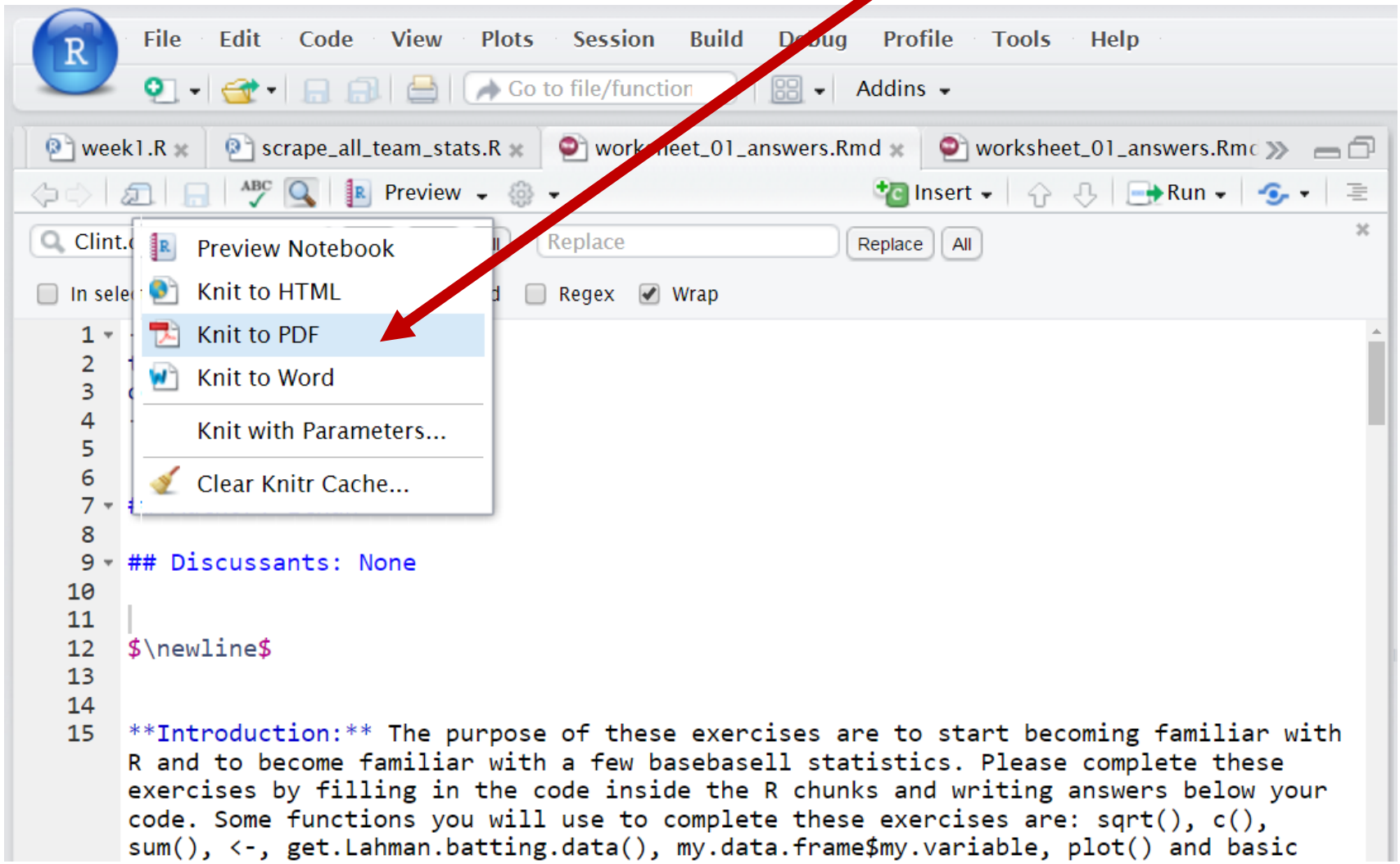
π

$\hat{p}$

$\hat{p}_{red}$

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 0

To practice the material from the first week of class I have created an R Markdown document called ‘homework 0’

- SDS100::download_homework(0)

You will not turn in this homework, it is purely for practice!

Do not worry if you run into any technical difficulties with the homework, the purpose of this homework is to work out any issues so you will be all set for the first real homework.