Introduction to R and categorical data

Overview

Review

Intro to R continued

Categorical data

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

Announcement

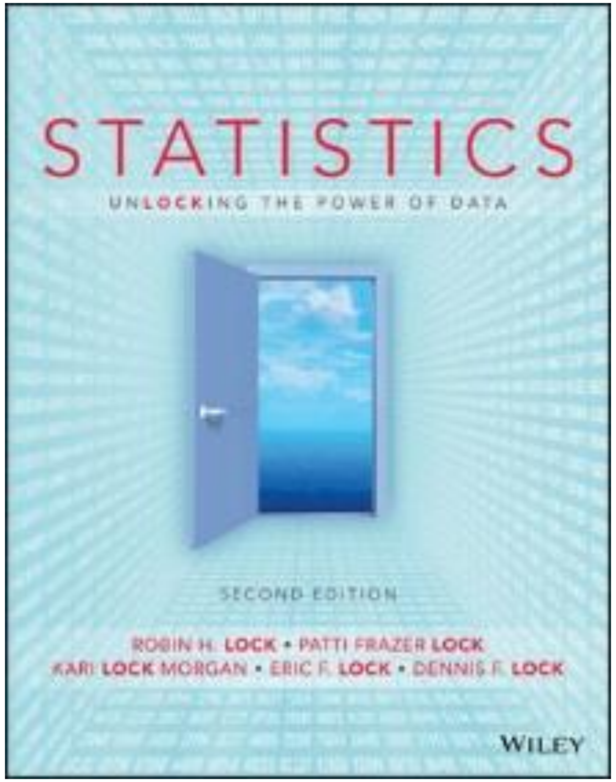
If you haven’t done so yet, please remember to fill out the background survey under the quizzes on Canvas

Any questions about the Lock5 practice problems?

Practice problems from Lock 5, first edition:

1.1, 1.3, 1.5, 1.11, 1.25, 1.26

Has everyone ordered the book?


STATISTICS
UNLOCKING THE POWER OF DATA
SECOND EDITION
ROBIN H. LOCK • PATTI FRAZER LOCK
KARI LOCK MORGAN • ERIC F. LOCK • DENNIS F. LOCK
WILEY

Quiz time! (not to be turned in)

1. What is a population? All individuals/objects of interest (Truth)

2. What is a sample? A subset of the population (shadows)

3. What is statistical inference? Making judgments about the population using data from the sample

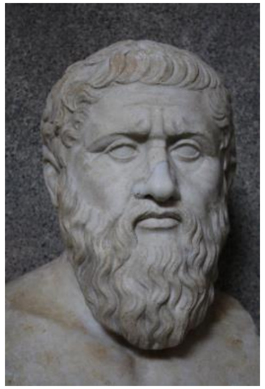
4. What are the rows of a data table called? Cases/observational units

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

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

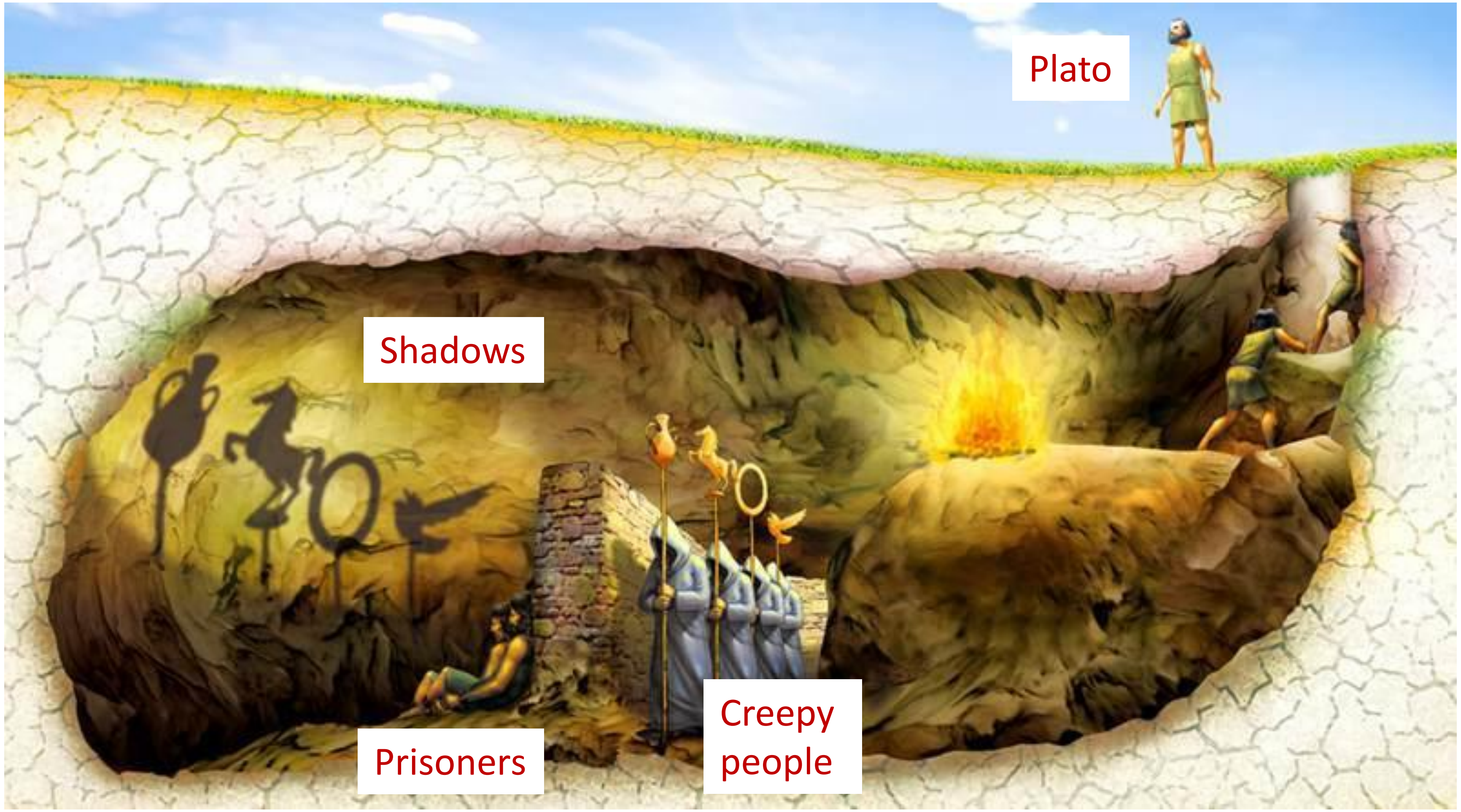
- Categorical variables fall into discrete categories
- Quantitative variables are numbers

7. Who is this?

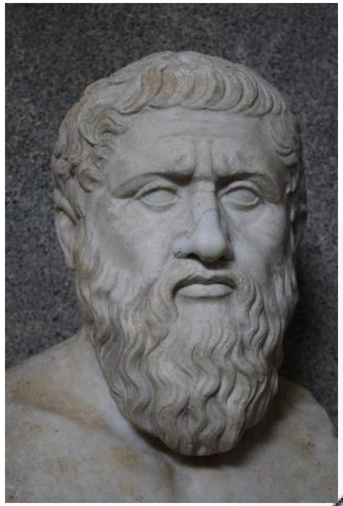


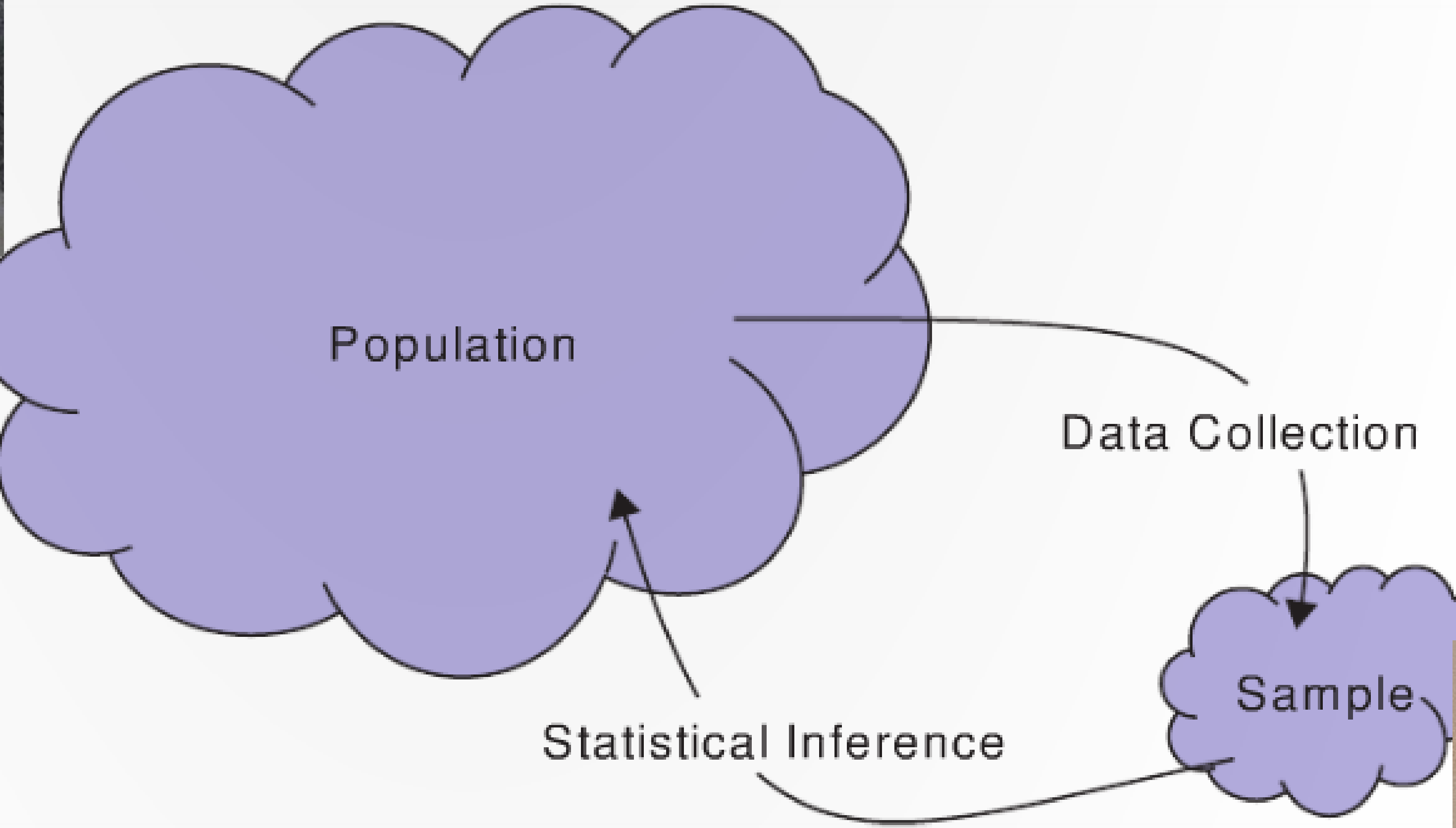
Plato

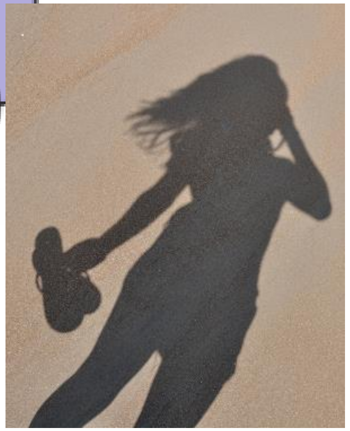
Plato’s cave



From The Republic (~ 380 BCE)







We will come back to this soon, but first back to R!

Review: R Basics

Log into R Studio Cloud: http://bit.ly/SDS100

Arithmetic:

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

Assignment:

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

Review: 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
```

Question

Q: What kind of grades the pirate get in Introduction to Statistics?

A: High Seas

Q: Worst joke of the semester?

A: Not likely

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?
```

We can get multiple elements from a vector too

```
> s[c(1, 2)]
```

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
```

Question

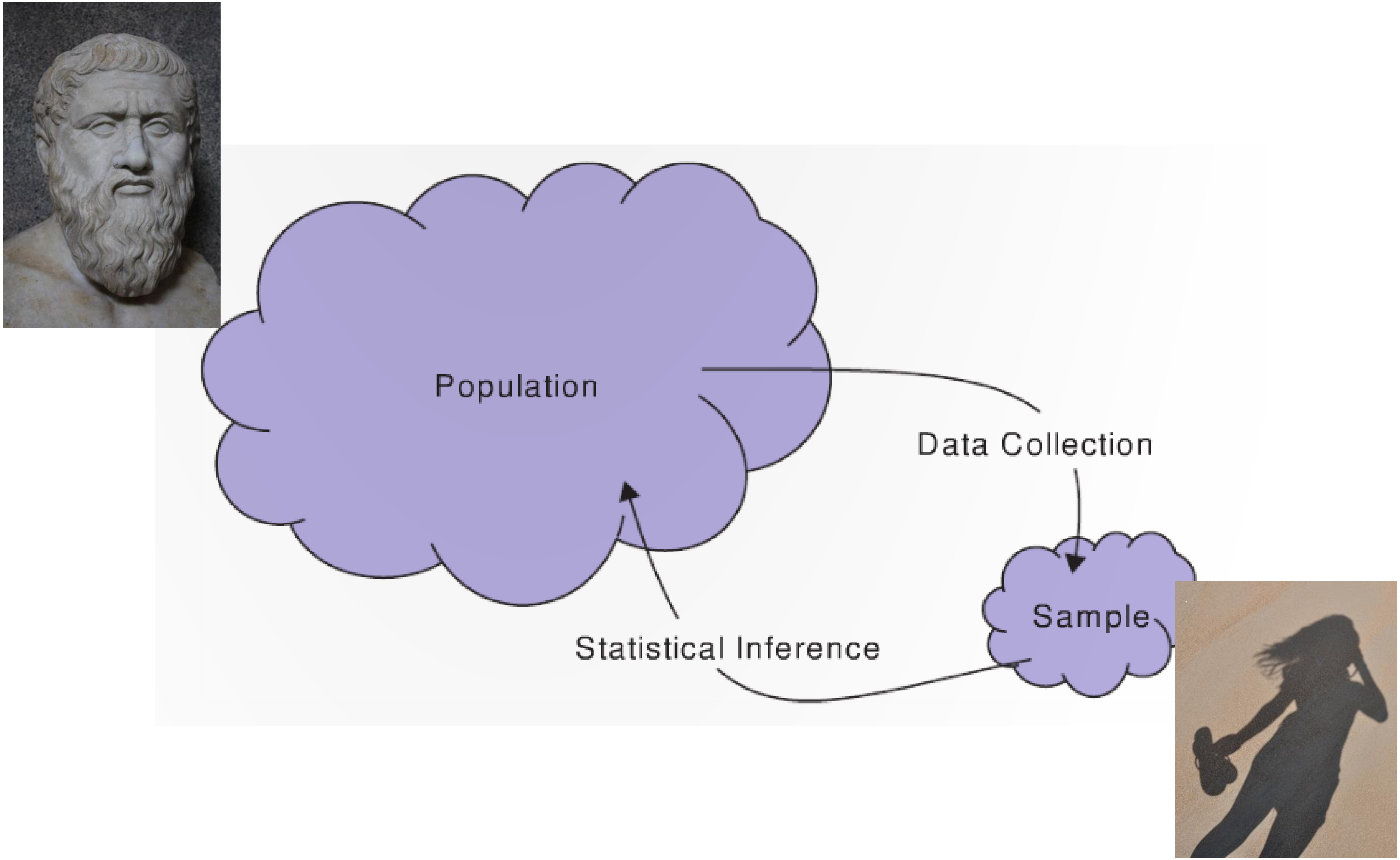
Q: What was the movie, ‘Pirates of the Caribbean’ rated?

A: PG-13

Q: Worst joke of the semester?

A: We are just getting started!

Now back to fundamental concepts in Statistics…


Population
Data Collection
Sample
Statistical Inference

The sprinkle business (fictional)

ACME CORPORATION

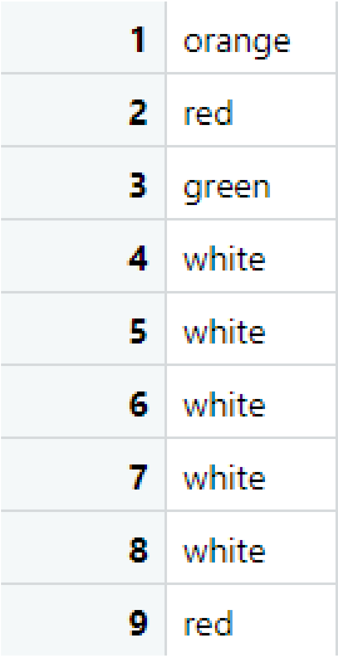
PERFECT Corporation

ACME corporation believes that if they use the same 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

<u>Questions:</u>

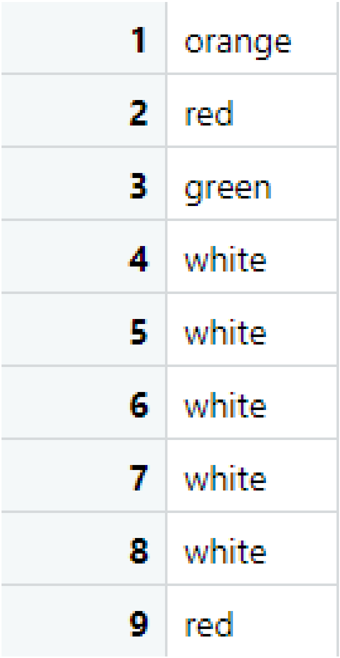
1) What are the observational units (cases)?

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?



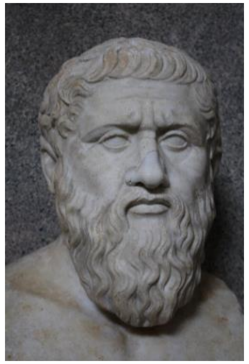
1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

Population parameters vs. sample statistics

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

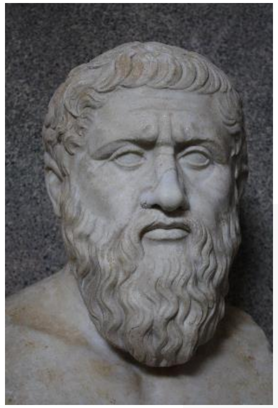
- Not to be confused with Statistics, which is a field of study

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

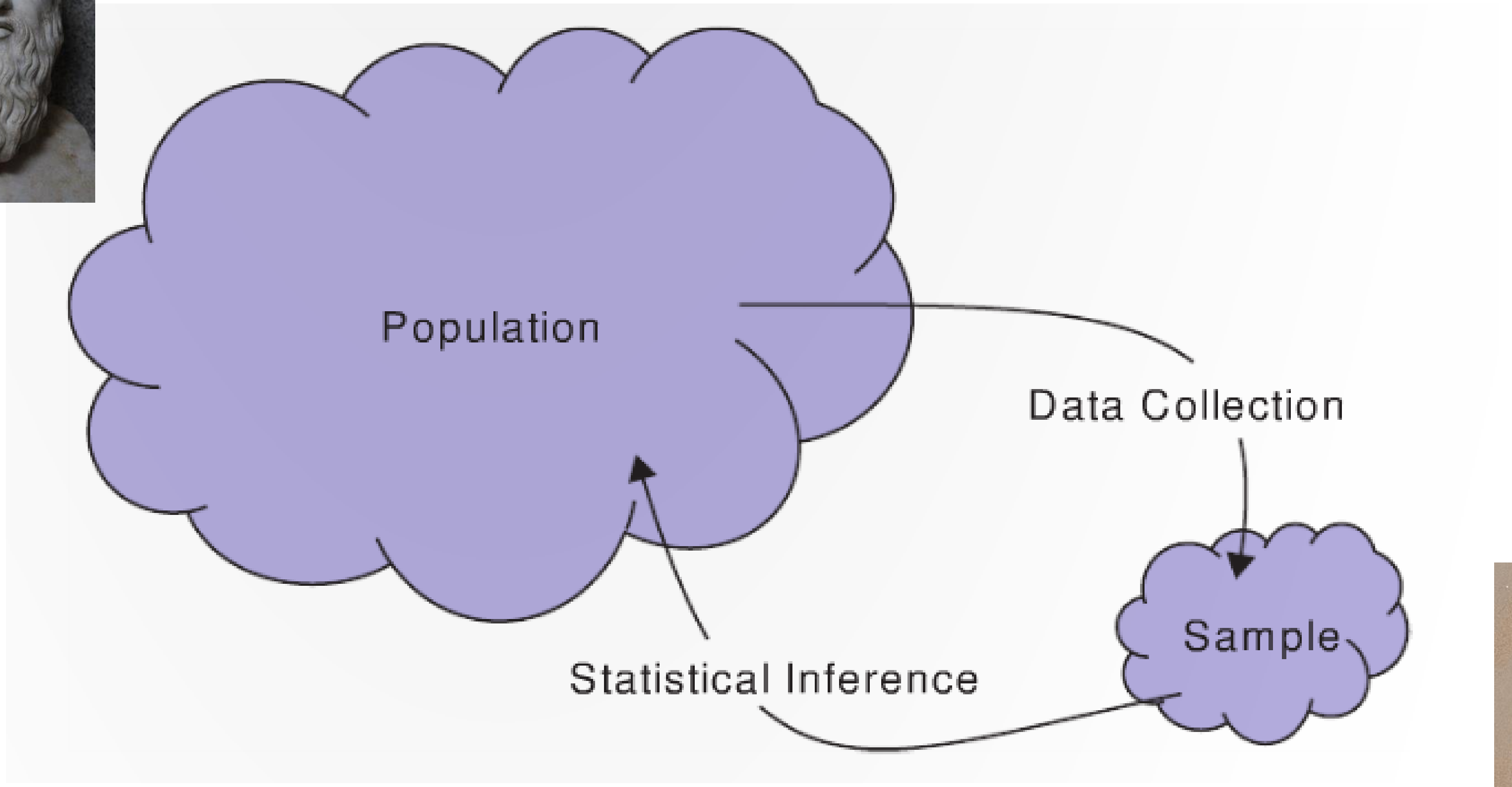


?

Parameters and statistics



Parameters



statistics

Categorical variables

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 **p̂** (pronounced “p-hat”)

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

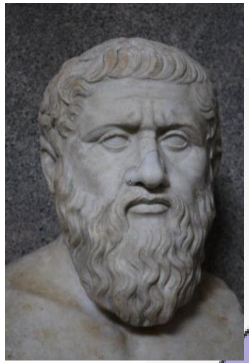
The proportion for a ***population*** is denoted **π** (the book uses p)

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

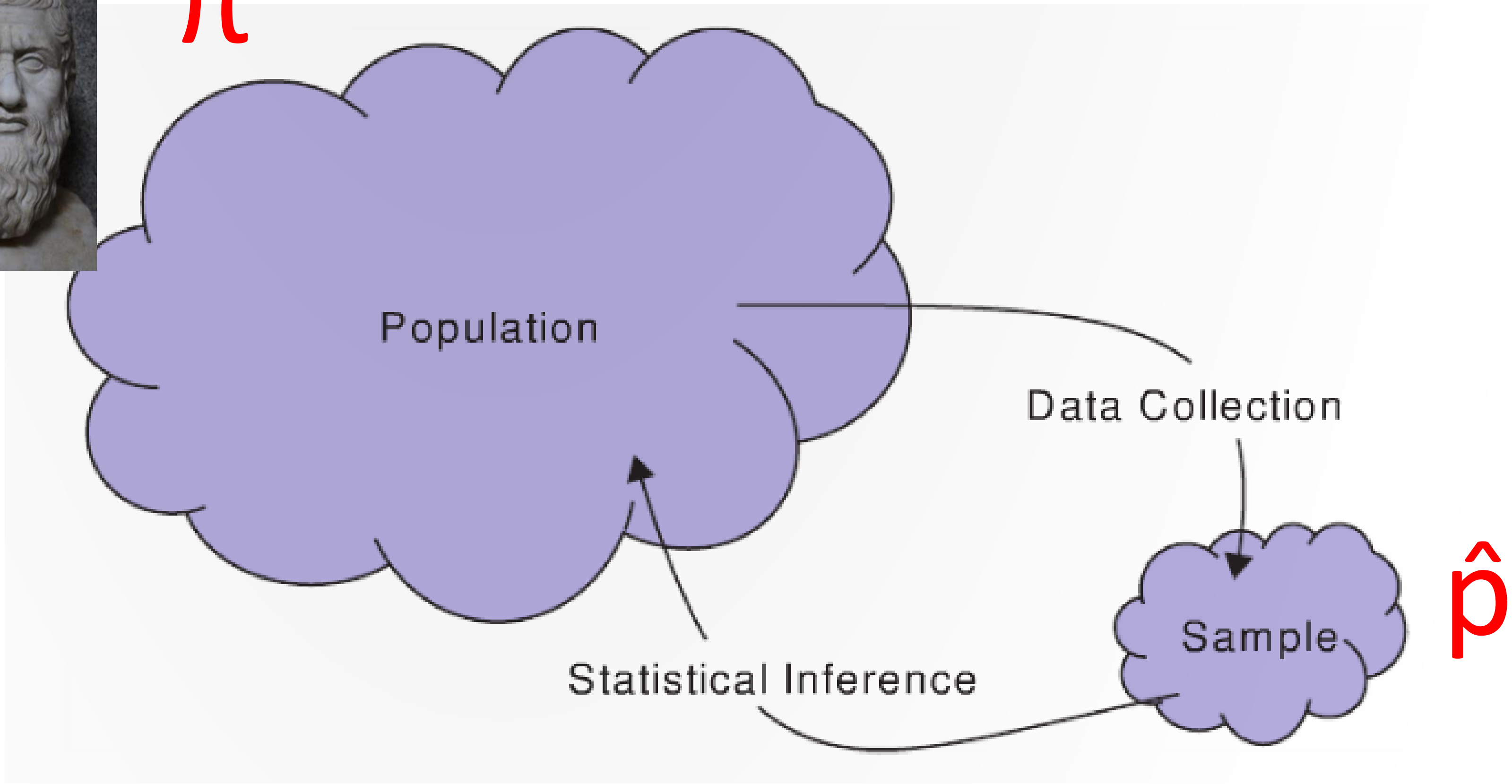
p̂ is a ***point estimate*** of π

- i.e., p̂ 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***

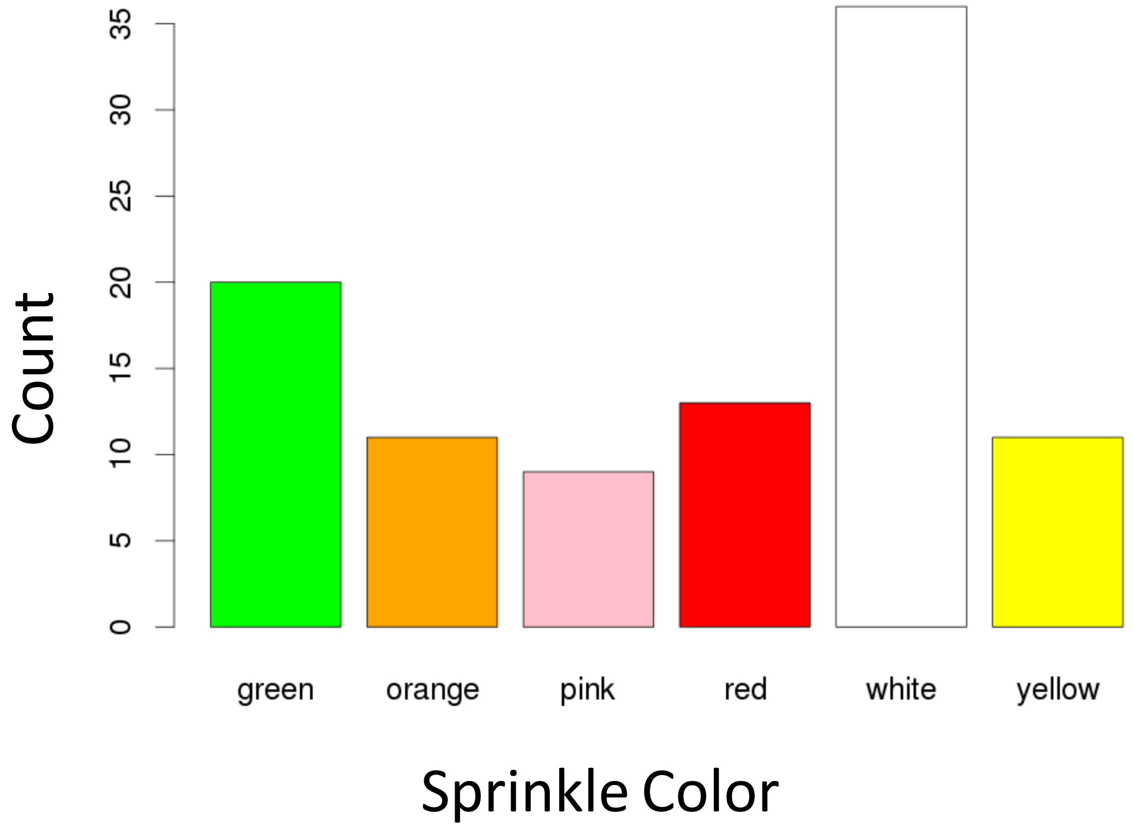
Color	**green**	**orange**	**pink**	**red**	**white**	**yellow**		**Total**
Count	20	11	9	13	36	11		100

Calculating proportions (relative frequencies)

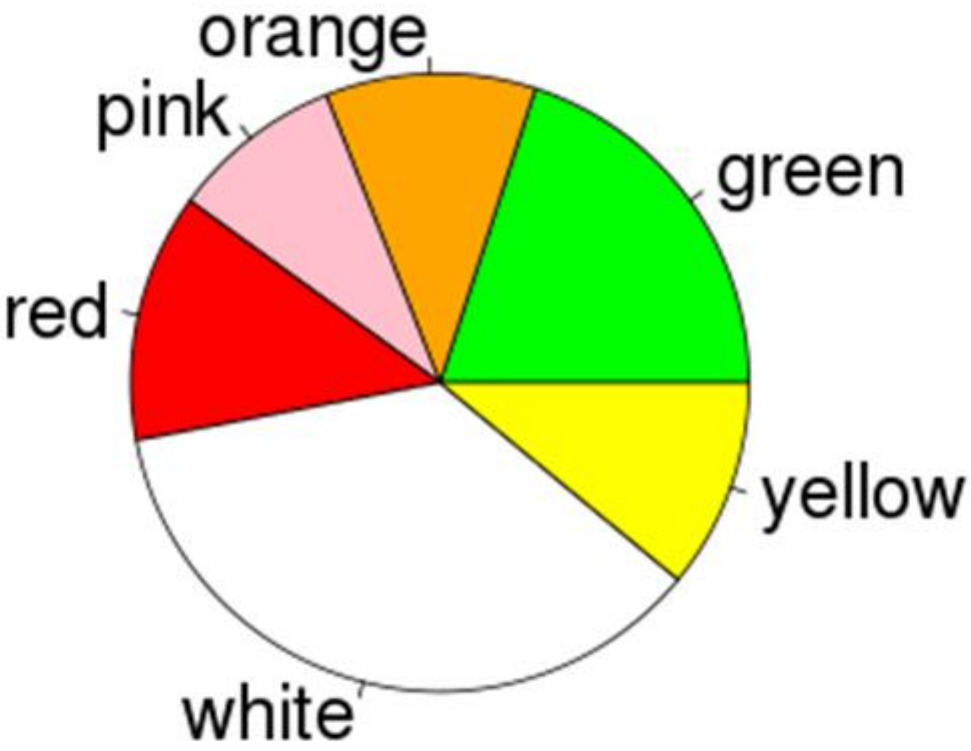
We can convert a frequency table into a ***relative frequency table*** by dividing each cell by the total number of items

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

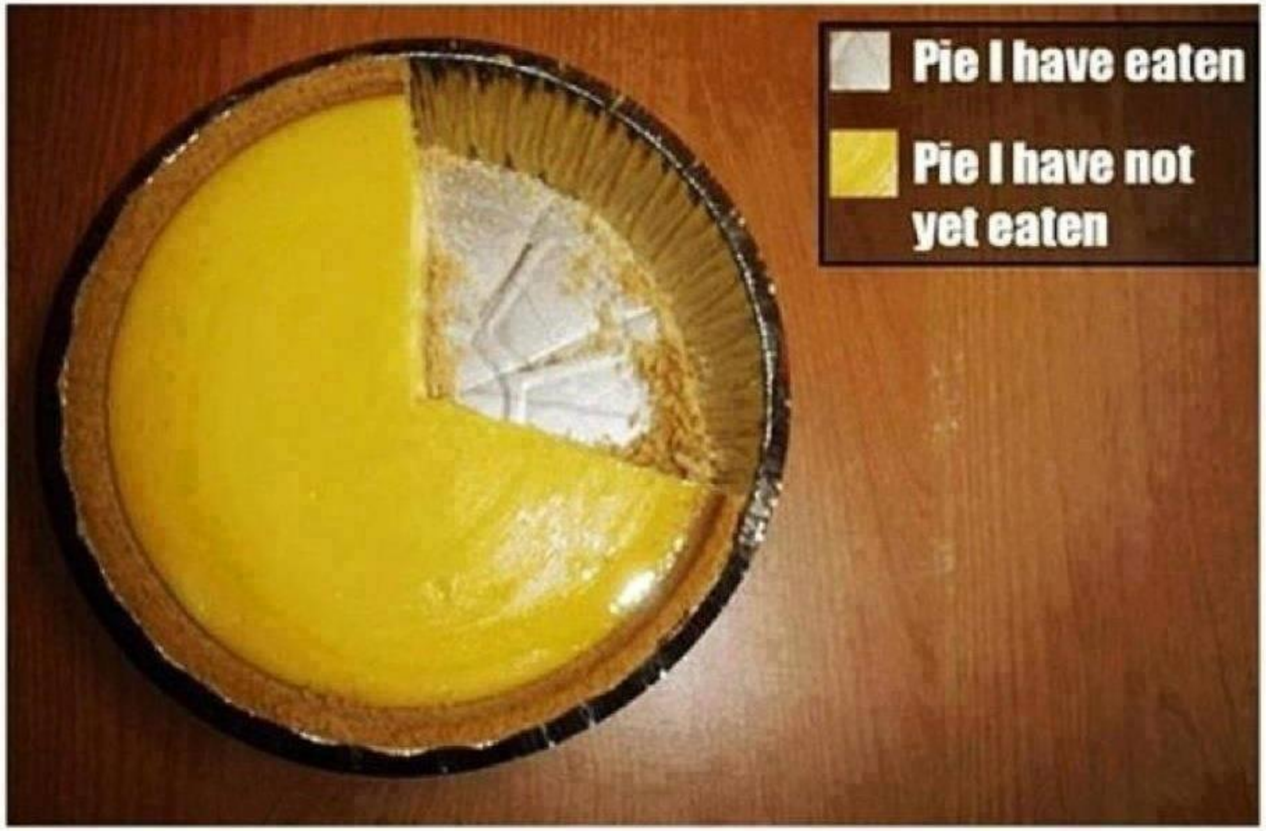
Visualizing categorical data: The Bar Chart



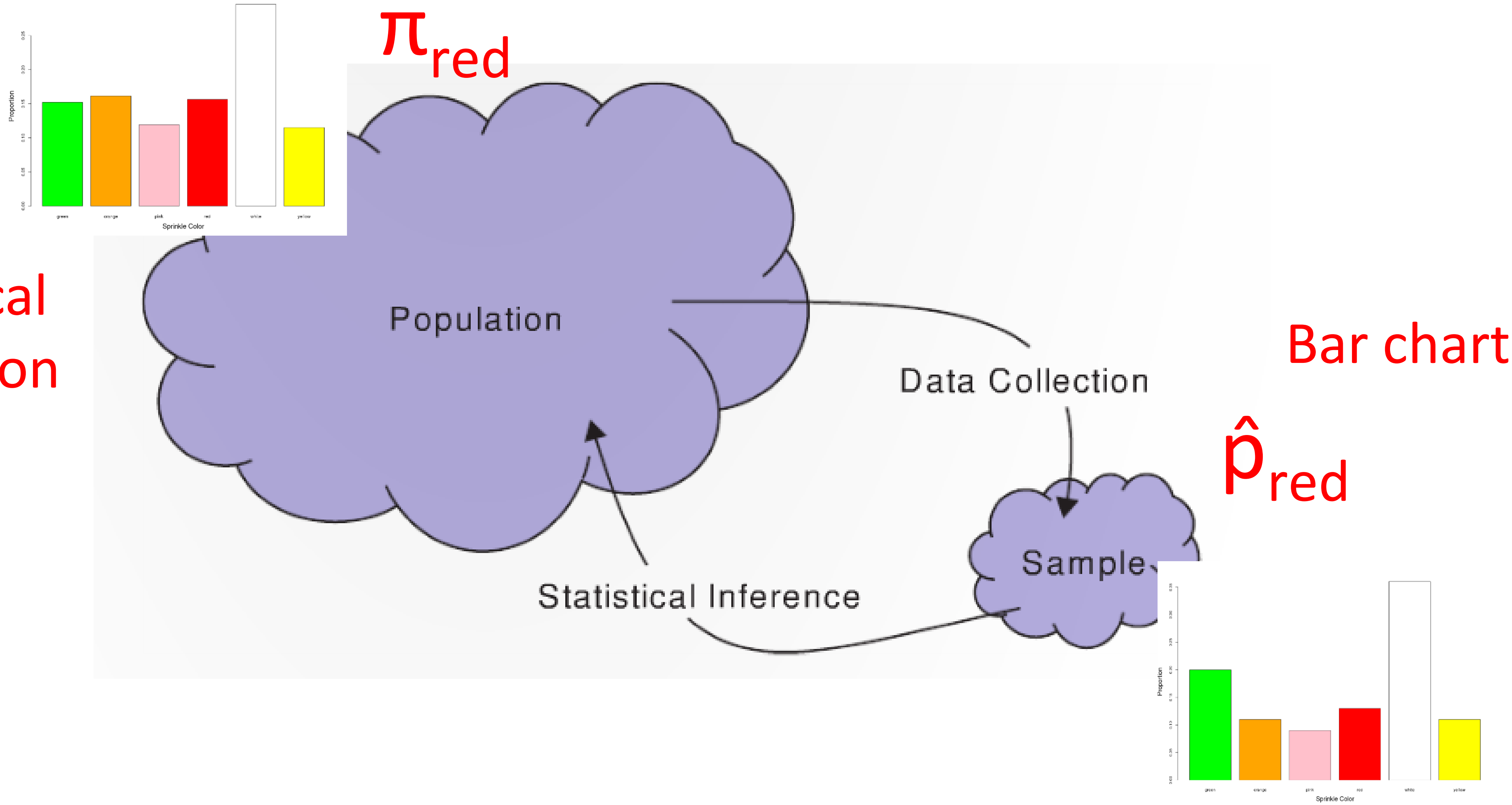
Visualizing categorical data: The Pie Chart



World's Most Accurate Pie Chart



Summary: Sample and Population proportion



π_{red}

Categorical distribution

Bar chart

$\hat{p}_{red}$

Let’s sample virtual sprinkles…

Back to R Studio Cloud! http://bit.ly/SDS100

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 **p̂**
- The population proportion is denoted **π**
- Thus p̂ 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
bar(my_table)
pie(my_table)
```