Lecture 2
Describing and Visualizing Distributions

Sampling and Data

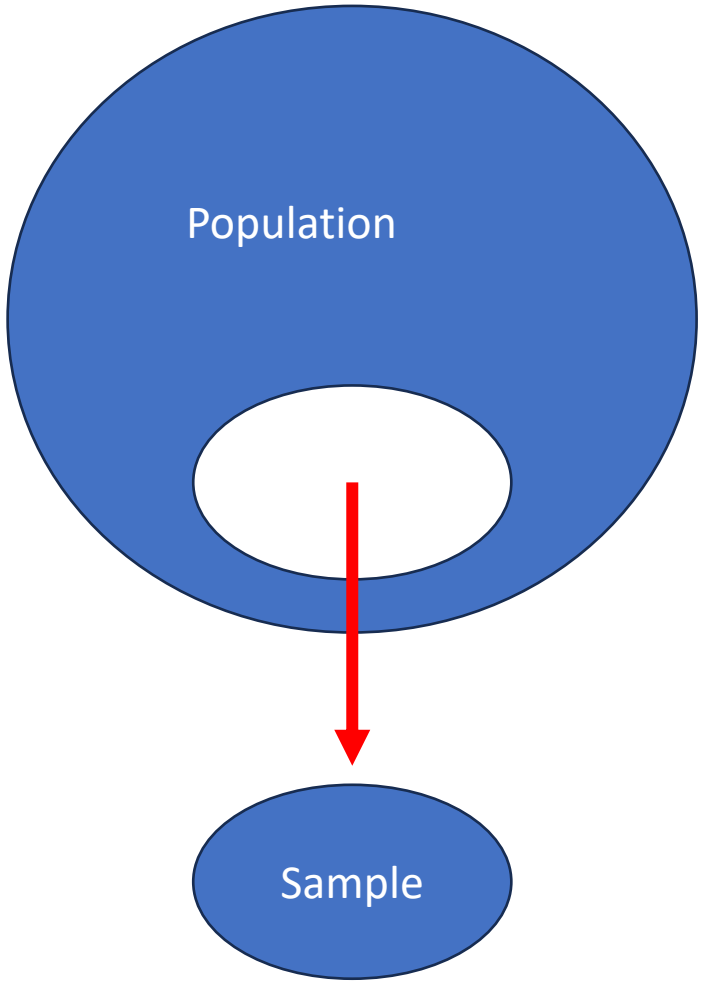
Statistics is generally concerned with studying properties of a **population** – the collection of all possible persons, events, or objects of interest

e.g the set of all possible observations – observed + unobserved

- Populations can be *real* and *finite/countable* (e.g. All employees at a company) or *hypothetical* and potentially *infinite/uncountable* (e.g all possible hands in a game of poker)

A **Sample** is a subset of the population that we actually observe – the observed observations

- The idea of **Sampling** is to select a portion or individuals or objects that are *representative* of the population
- By studying the sample we can gain insights about the population



Example

The population is a set of N observations

$\{x_1, x_2, x_3, \ldots x_N\}$

The sample is a set of n observations

$\{x_1, x_7, x_8, \ldots x_n\}$

Population: $N = 20$

Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	**FN-2414**	**Berchest Station**	**184.9**	**19**	**0.62**	**PV1**
2	FN-2462	Death Star	193.3	20	0.66	PV2
3	FN-2178	Death Star	191.0	20	0.77	CPL
4	FN-2525	Lothal	186.7	23	0.61	PFC
5	FN-2194	Corellia	194.6	21	0.66	PV1
6	FN-2937	Fondor Ship Yard	191.9	22	0.75	PV2
7	**FN-2817**	**Fondor Ship Yard**	**189.5**	**21**	**0.59**	**CPL**
8	**FN-2117**	**Death Star**	**193.5**	**21**	**0.66**	**PFC**
9	FN-2298	Corellia	193.4	24	0.66	PV1
10	FN-2228	Berchest Station	193.2	21	0.71	PV2
11	FN-2243	Death Star	192.8	24	0.69	CPL
12	FN-2013	Corellia	192.3	18	0.62	PFC
13	FN-2373	Lothal	190.3	22	0.60	PV1
14	**FN-2664**	**Berchest Station**	**189.5**	**21**	**0.72**	**PV2**
15	FN-2601	Fondor Ship Yard	189.2	21	0.73	CPL
16	FN-2602	Lothal	188.2	22	0.62	PFC
17	FN-2767	Death Star	189.8	20	0.76	PV1
18	FN-2708	Death Star	186.3	20	0.61	PV2
19	**FN-2090**	**Fondor Ship Yard**	**197.7**	**19**	**0.64**	**CPL**
20	FN-2952	Corellia	194.5	19	0.57	PFC

Sample: n = 5

Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	FN-2414	Berchest Station	184.9	19	0.62	PV1
7	FN-2817	Fondor Ship Yard	189.5	21	0.59	CPL
8	FN-2117	Death Star	193.5	21	0.66	PFC
14	FN-2664	Berchest Station	189.5	21	0.72	PV2
19	FN-2090	Fondor Ship Yard	197.7	19	0.64	CPL

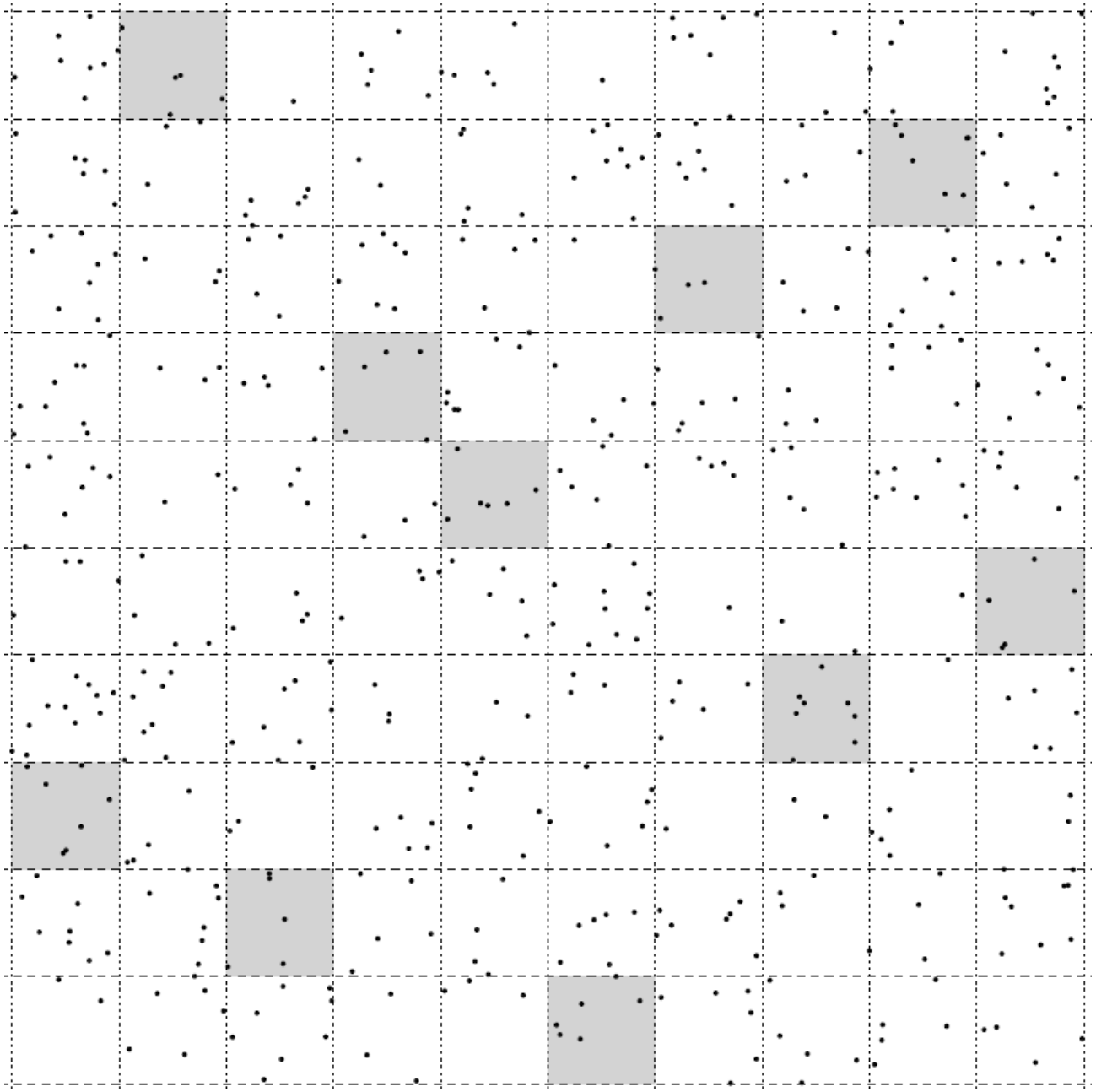
Example 2

Consider a rectangular-shaped piece of land that has been divided into 100 smaller rectangular units, each containing something of interest (e.g., trees, burrows, archaeological artifacts).

A subset of 10 of those smaller units was selected and the number of objects in each of these units was counted.

Population Size: $N = 100$

Sample Size: $n = 10$



Why is statistics so valuable?

- Most of the time, we can't measure everyone or every unit in the population and therefore must limit our measurements to a sample.
- Statistics primarily deals with **estimation** – the process of inferring an unknown quantity about a population using set of sample data
- The tools for estimation allow us to approximate almost everything about populations **using only samples**.

Where we can go with estimates

- With **estimates** we can
  - Assess differences among groups and relationships between variables.
  - Describe populations. Examples of estimates include averages, proportions, measures of variation, and measures of relationship.
  - Then we can ask and answer questions or formally, test and evaluate hypotheses.

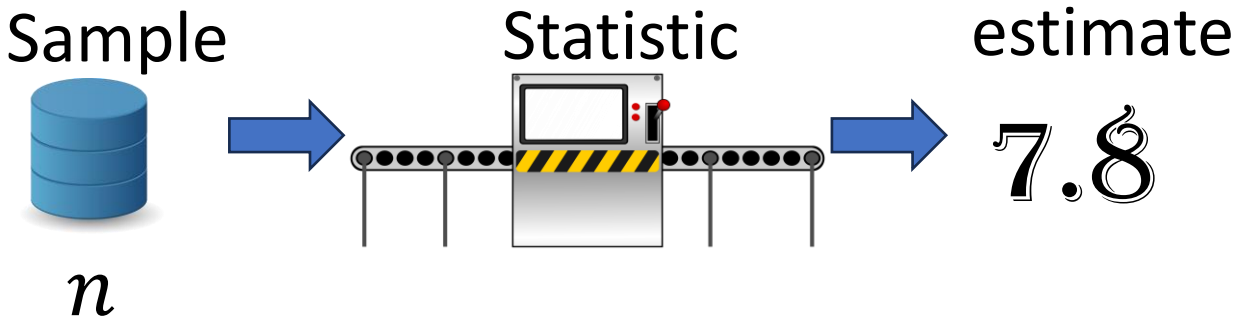
Statistics Vs Parameters

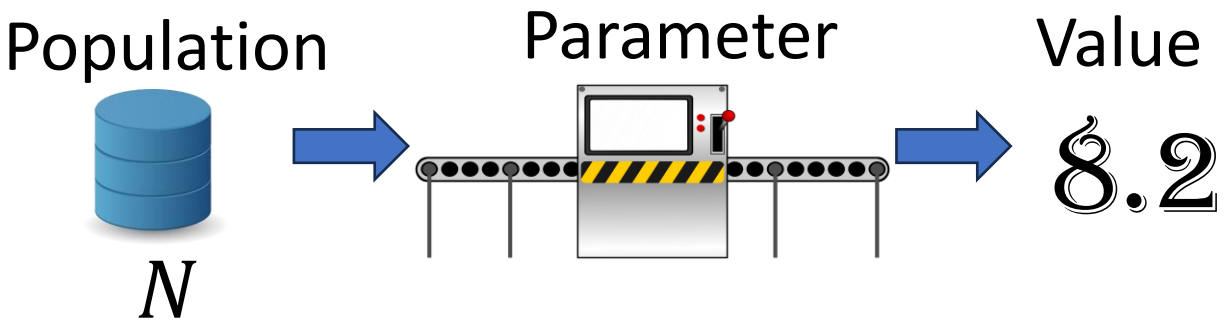
A **statistic** is a numerical characteristic of a **sample** that estimates a population parameter.

A **parameter** is a numerical characteristic of a **population** that can be estimated by a statistic.

Put another way…

A **statistic** is a function of the observations of in a sample while a **parameter** is a function of all observations in the population.





Mean and Proportion

A **proportion** describes the fraction of a whole that represent some property or category. Usually, it is expressed a percentage.

Notation:

$\hat{p}$ - denotes the sample proportion

p - denotes the population proportion

Parameter

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$p = \frac{\text{Number of objects in category}}{N}$$

The arithmetic **mean** is the center of a set of data (we often use the words mean and average interchangeably)

Notation:

$\bar{x}$ - denotes the mean of a sample

μ – denotes the mean of a population (i.e the population parameter)

Statistic

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

$$\hat{p} = \frac{\text{Number of objects in category}}{n}$$

Example: Gallup Pool

- On April 20, 2010, one of the worst environmental disasters took place in the Gulf of Mexico when the Deepwater Horizon offshore oil rig exploded.
- In response to the spill, many activists called for an end to offshore drilling for oil.
- Almost nine months later, turbulence in the middle east caused the price of oil to surge to an all-time high.
- In March 2011, Gallup conducted a survey and found that 60% of Americans favored offshore drilling as means to reduce U.S dependence on foreign oil.
- The poll was based on interviews with 1,021 adults aged 18 and older, living in the continental U.S, and selected using random digit dialing.
- **What is the population under study, and what is the population parameter being estimated? What is the sample statistic ?**

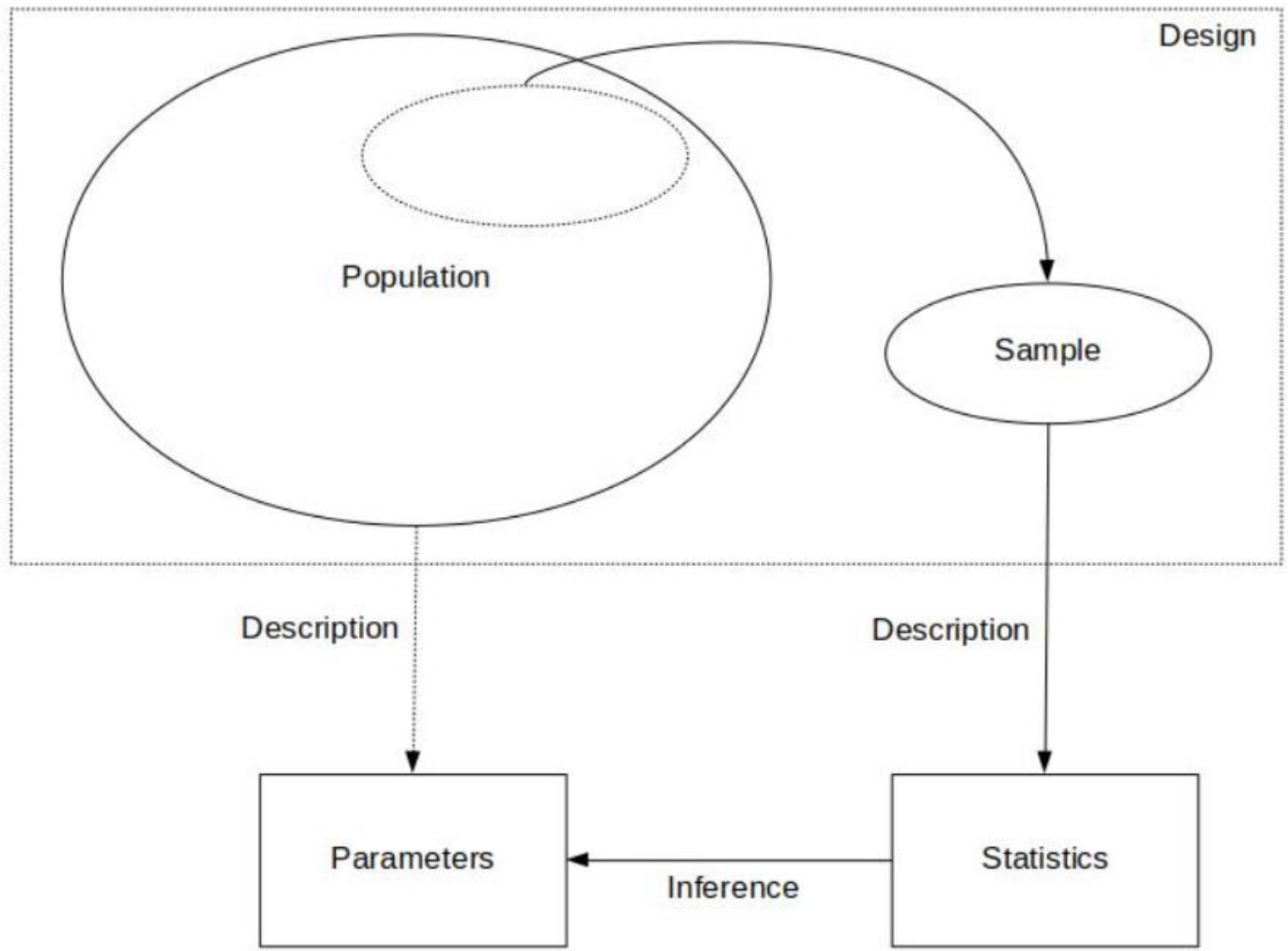
Descriptive Vs. Inferential Statistics

1. **Design** – The process/method in which we plan to collect data to answer our statistical question

2. **Descriptive Statistics –** refers to describing the observations in a sample using statistics or a population using parameters
   1. - collection, organization, summarization and visualization of data

3. **Inferential Statistics** (or statistical inference) – refers to using a sample (usually a statistic) to answer a question about a population (such as estimating the value of a parameter)

   - estimation, hypothesis testing, determining relationships among variables, prediction


Design
Population
Sample
Description
Description
Parameters
Inference
Statistics

Warm Up

In elections, television networks use exit polling (interviewing voters after they leave the voting booth) to declare the winner well before all votes are counted. In the 2010 California gubernatorial election between candidates Jerry Brown (D) and Meg Whitman (R), an exit poll projected Brown to be the winner early into election night. Specifically, the network responsible for the poll interviewed 3,889 voters at the booth and determined that 53.1% favored the democratic candidate.

What is the statistical question?

Who will win the California Race for Governor

What is the population? What is the population size N?

The population is all eligible voters in the state of California

What is the sample ? What is the sample size n?

All voters interviewed at the election booth; 3,889 people

What is the statistic being calculated?

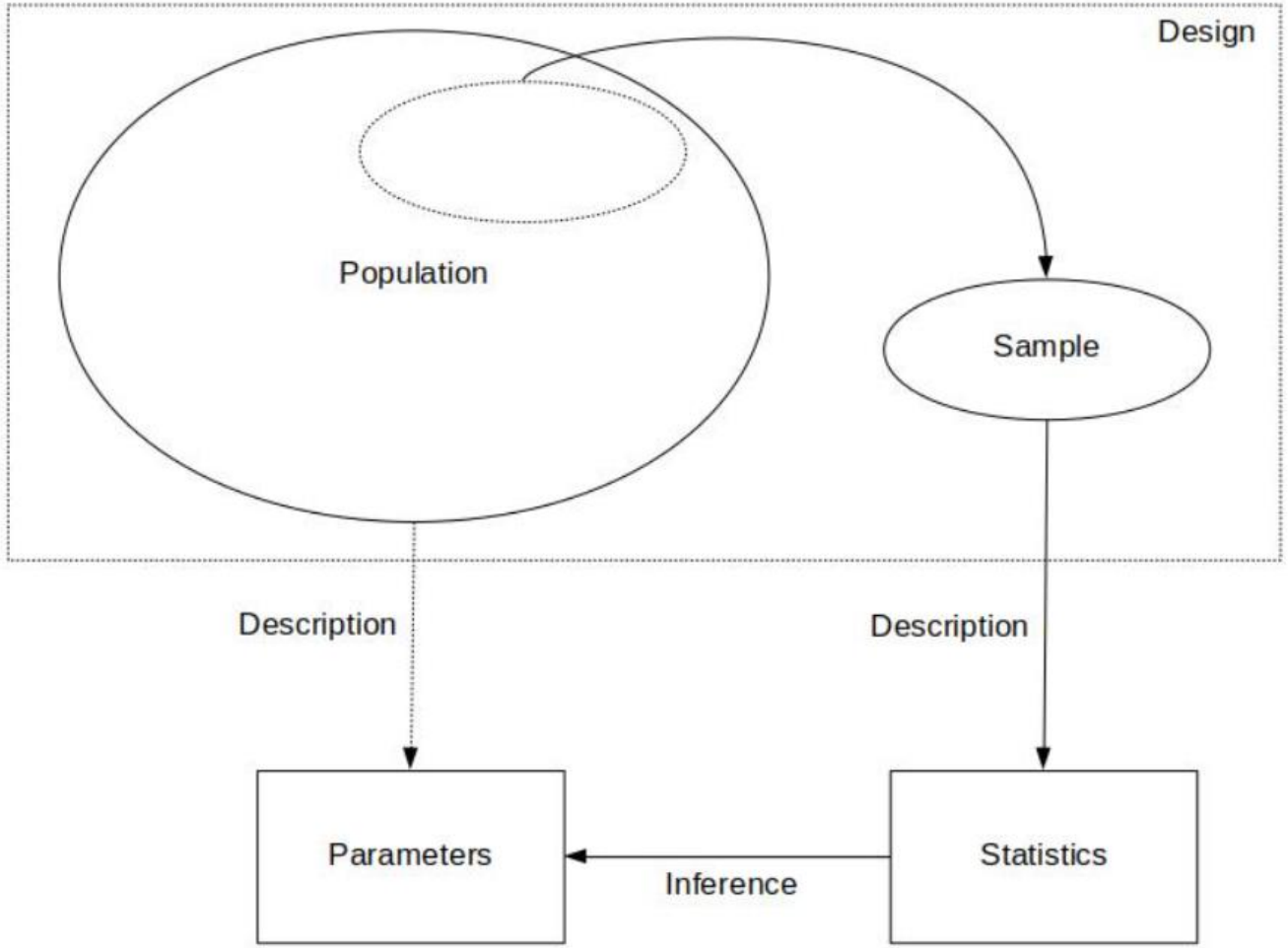
The proportion of 3,889 voters casting their vote for Democrat Jerry Brown

Recap:

1. **Design** – the goal/statistical question we want to answer and how we plan to obtain our data

2. **Description** – a preliminary exploration and summary of the data

3. **Inference** – using statistics to make decisions or predictions about the data



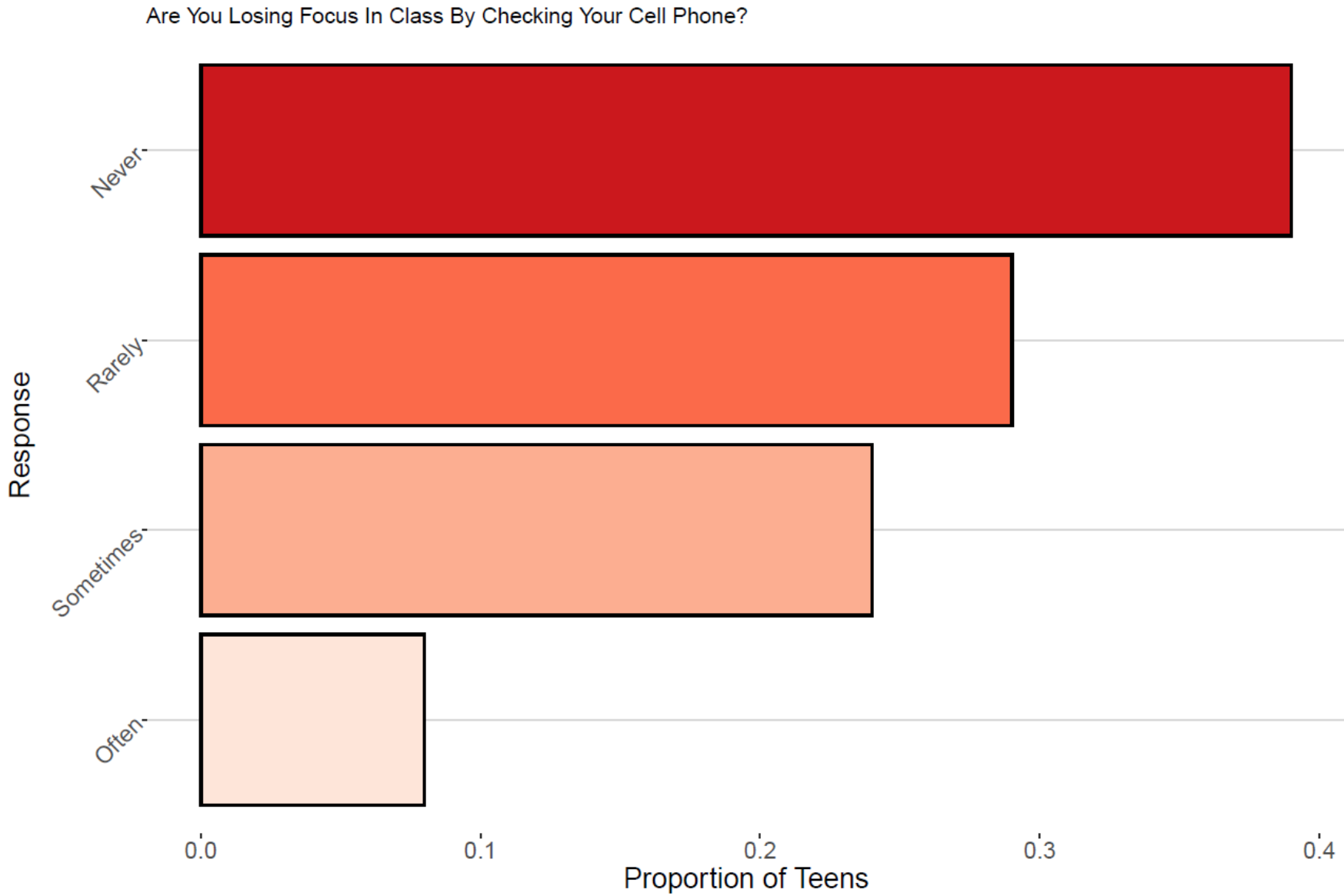
Descriptive statistics

- Methods for summarizing, visualizing and characterizing data
- The data can be from samples OR populations
- Goal: simplify the data without distorting or losing information
- Summaries are easier to understand

Ex. Are teens distracted by their cell phones? A study conducted by the Pew Research center surveyed 743 U.S teenagers ages 13-17 to understand how cell phones in class impacted their ability to concentrate. Each student was asked to rate the impact of cell phone use on their concentration based on a Likert scale

Student	Age	Response
1	13	Never
2	13	Sometimes
3	15	Never
4	17	Often
⋮	⋮	⋮
743	16	Rarely

Possible Responses: Never, Rarely, Sometimes, Often


Are You Losing Focus In Class By Checking Your Cell Phone?
Response
Never
Rarely
Sometimes
Often
0.0
0.1
0.2
0.3
0.4
Proportion of Teens

Features of Distributions

- A **distribution** of a variable gives (a) the values that occur and (b) how often each value occurs

Features of A Distribution

Categorical Variables:

- **Modal category** – the category with the highest frequency

Quantitative Variables:

- **Shape** – do observations cluster into certain areas, are values spread out or more densely packed together?
- **Center** – where is the middle point on the distribution, where does a typical value fall?
- **Variability** – how tightly to observations cluster around the center of the distribution

Displaying Distributions: Frequency Tables

- **Frequency table –** a table listing the distinct values of variable together with the number of observations of each value
- **Frequency** – the number of times a value occurs
- **Relative Frequency** – the proportion of observations that assume a given value

$$\text{RF} = \frac{f}{n},$$

- **Cumulative Relative Frequency** - the proportion of observations equal to or less than a given value (more on this later)

Ex. Are teens distracted by their cell phones?

Response	Frequency	Relative Frequency	Cumulative Relative Frequency
Never	289	0.39	0.39
Rarely	216	0.29	0.68
Sometimes	178	0.24	0.92
Often	60	0.08	1.00

Ex. Mendels Pea Plants - In one of Gregor Mendel's classic studies, he bred 8023 pea plants and observed the color of the pea pods.

Pea Color	Frequency	Relative Frequency
Yellow	6022	0.751
Green	2001	0.249