STAT 2593

Lecture 001 - Populations, Samples, and Processes

Dylan Spicker

Populations, Samples, and Processes

Learning Objectives

1. What is statistics, and why do we care?
2. What are the components of statistical inference?
3. How are data categorized?
4. What are the different roles for statistics?
5. What are the different types of statistical studies?

The Big Question

What is **statistics** and why *should* you care?

ChatGPT

6σ

DOW JONES

THE LEGEND OF ZELDA TEARS OF THE KINGDOM

A Concrete Example

How might we be able to answer the question:

Does pay discrimination exist?

A Concrete Example

- Can we survey every single person and ask them their wages, then look at discrepancies?

A Concrete Example

- Can we survey every single person and ask them their wages, then look at discrepancies?

- Can we look at all the wages at a single company, and see if there appears to be discrimination?

A Concrete Example

- Can we survey every single person and ask them their wages, then look at discrepancies?
- Can we look at all the wages at a single company, and see if there appears to be discrimination?
- Can we look at available internet data, on salary websites?

The Problem

None of these techniques will give us accurate information about the problem as a whole.

What is Statistics?

Statistics is the process through which **data** are collected and analyzed in order to derive (useful) **insight** regarding a **population** (or process) of interest.

Important Concepts: Data

- **Data:** A collection of facts.

Important Concepts: Data

- **Data:** A collection of facts.
- **Variables:** Characteristics which vary between individuals, processes, objects, etc.

Important Concepts: Data

- **Data:** A collection of facts.
- **Variables:** Characteristics which vary between individuals, processes, objects, etc.
- **Observation:** An individual piece of data.

Important Concepts: Data

- **Data:** A collection of facts.
- **Variables:** Characteristics which vary between individuals, processes, objects, etc.
- **Observation:** An individual piece of data.
- Data can be **univariate**, if only a single fact is collected for each observations, or **multivariate**, if more than one fact is collected for each observation.

Important Concepts: Populations

- **Population:** A well-defined collection of objects of interest.

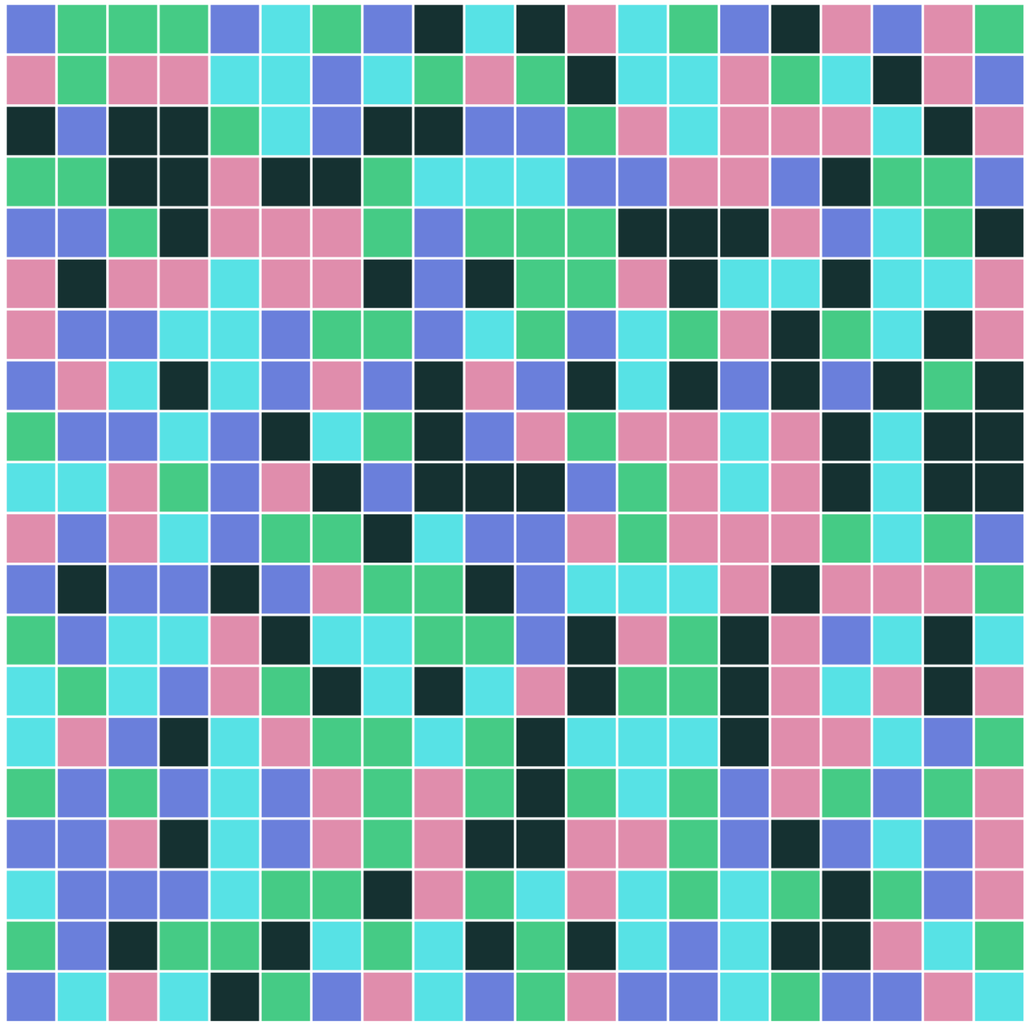
Important Concepts: Populations

- **Population:** A well-defined collection of objects of interest.
- **Census:** Collection of data for *all* members of a population.

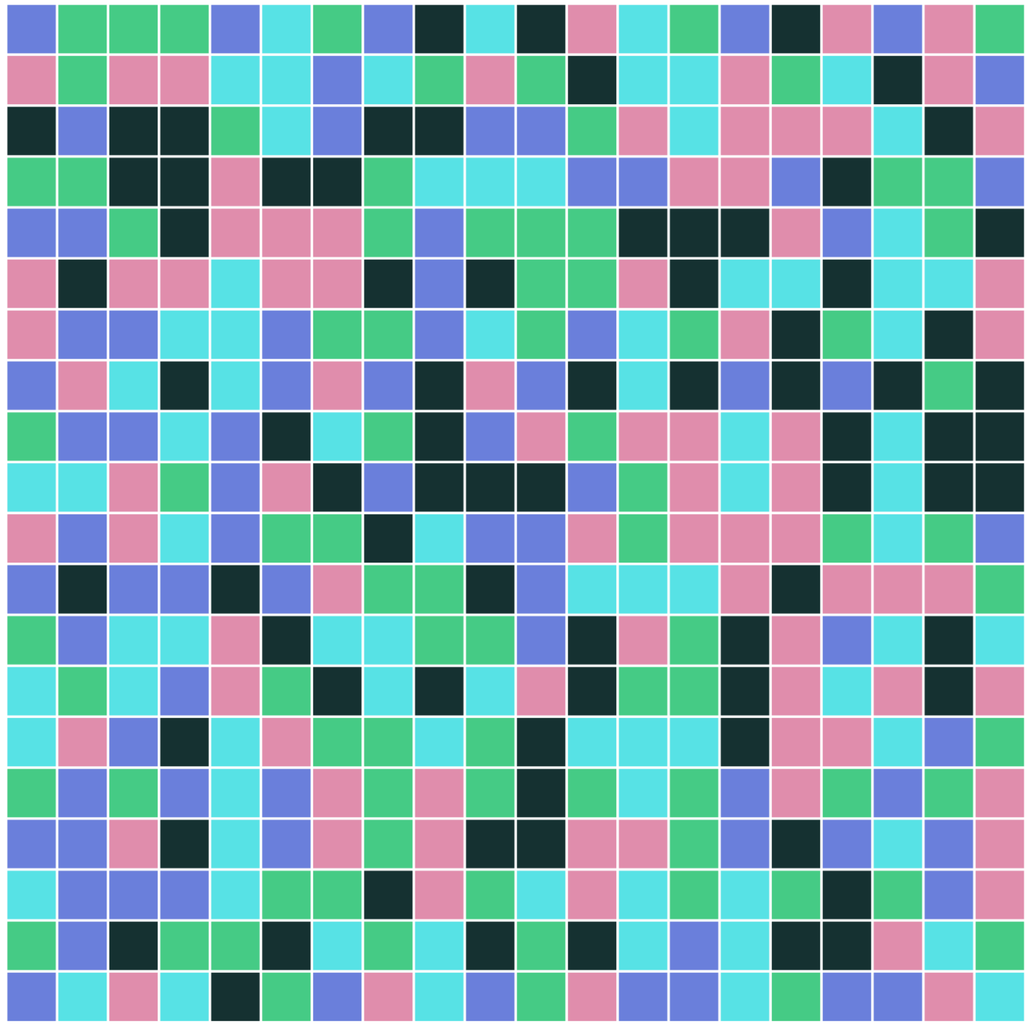
Important Concepts: Populations

- **Population:** A well-defined collection of objects of interest.
- **Census:** Collection of data for *all* members of a population.
- **Sample:** A subset of the population.

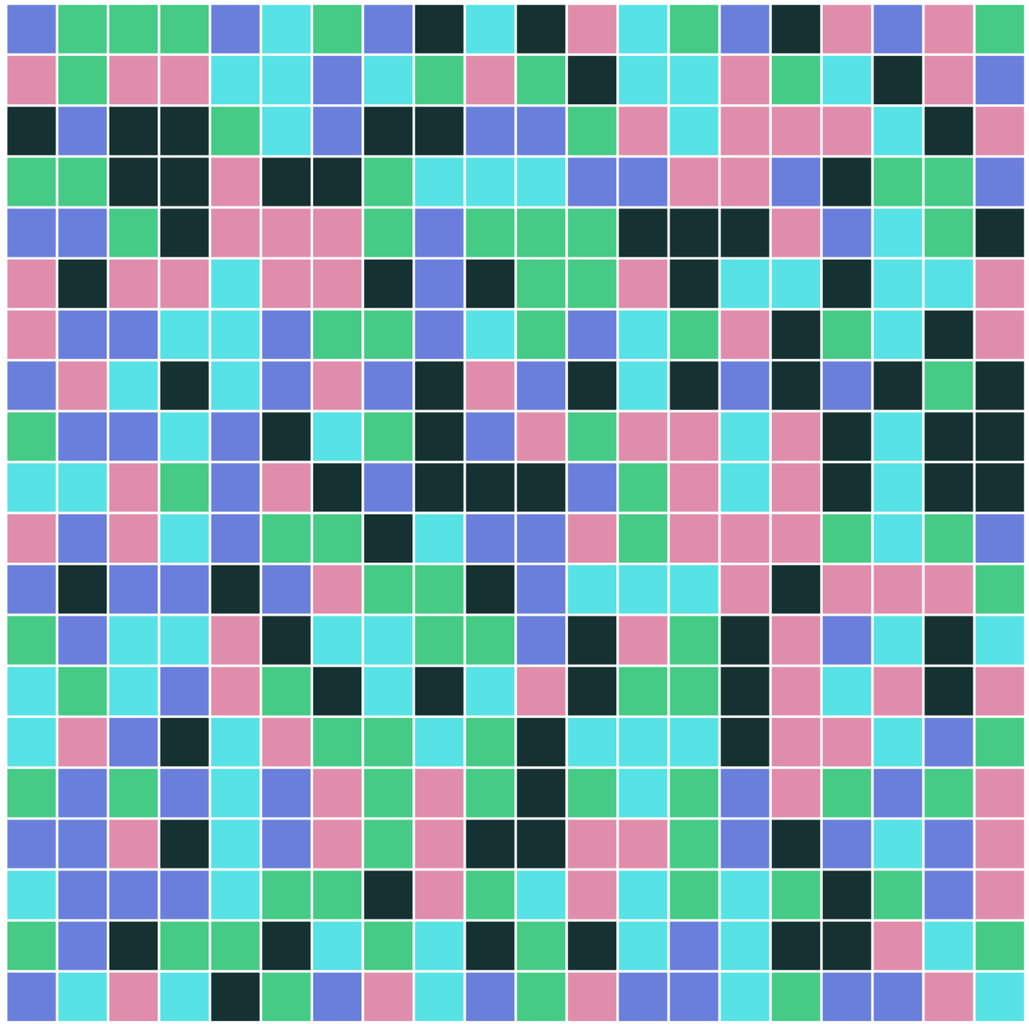
Population versus Sample, Visually

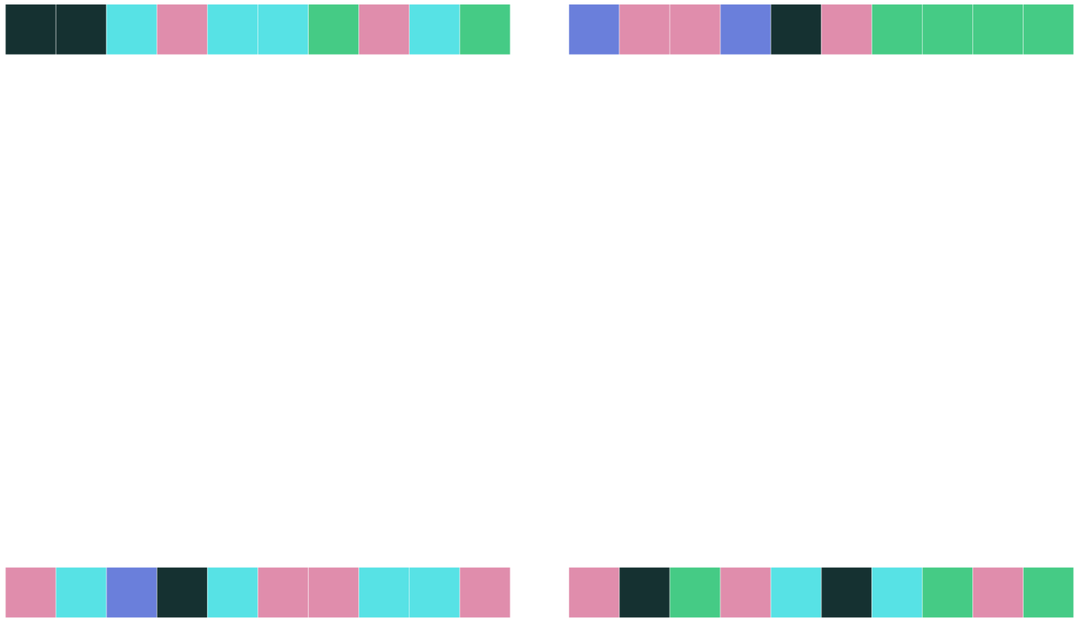


Population versus Sample, Visually

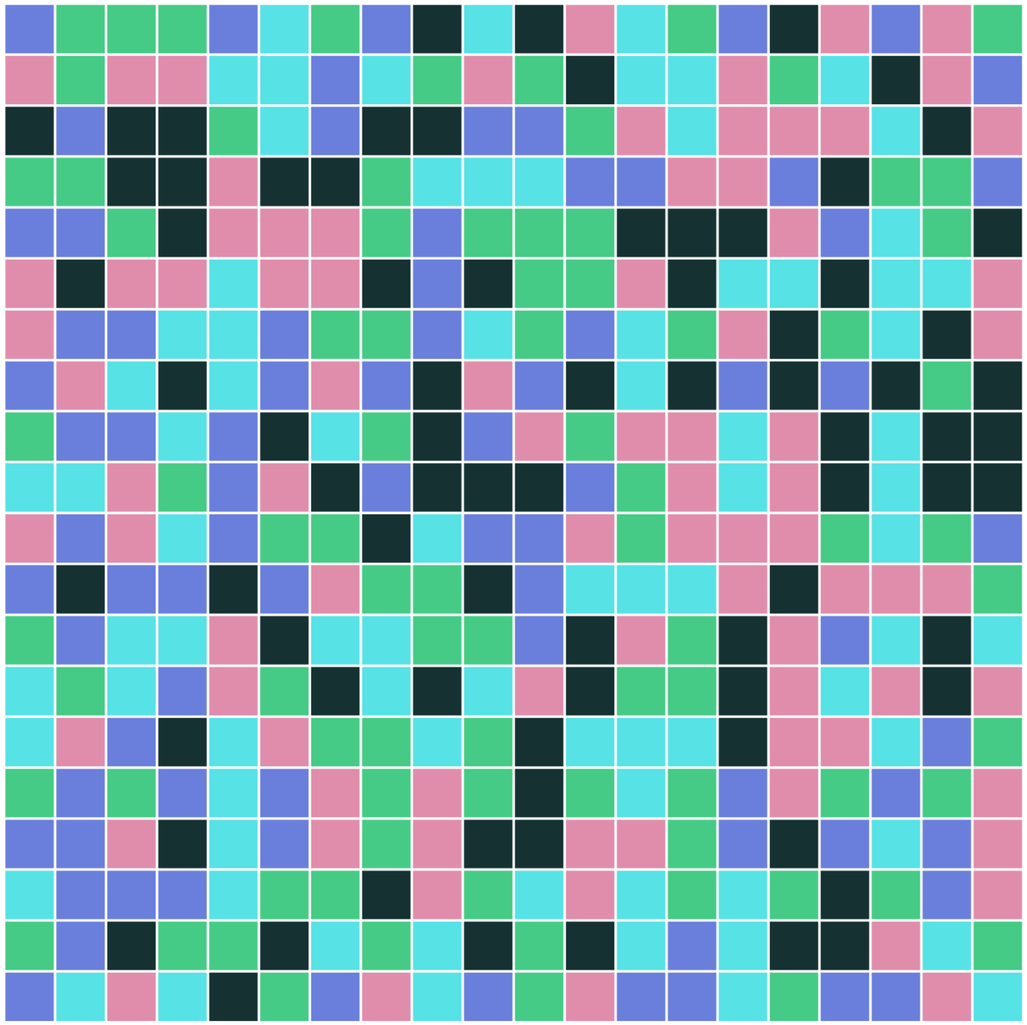


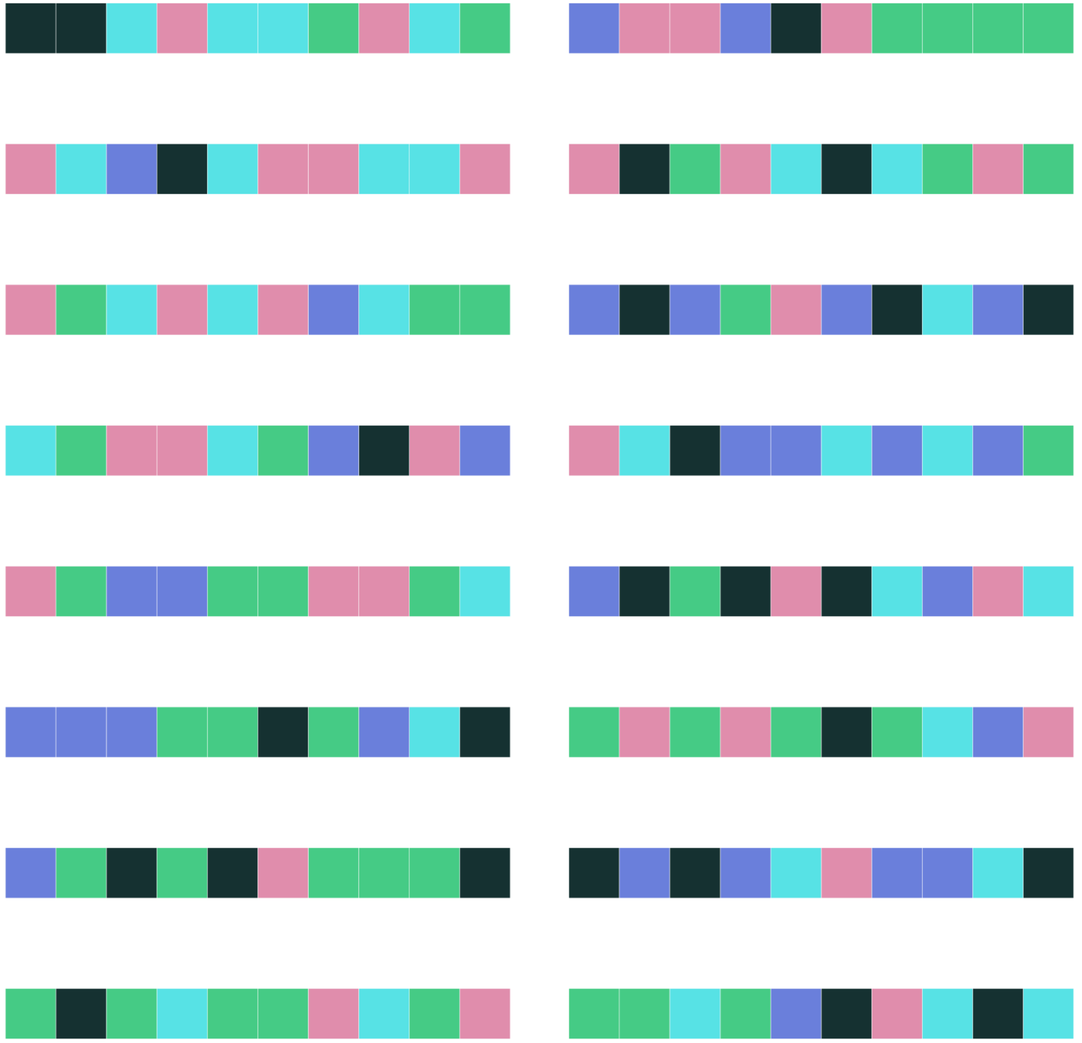
Population versus Sample, Visually





Population versus Sample, Visually





Practice

For each of the following scenarios, identify the population of interest, the sample, and the unit of observation. Is the described sampling effective for studying the population of interest?

1. A polling agency conducts a survey to predict the outcome of an upcoming election. They interview 2,000 registered voters across the country.

Practice

For each of the following scenarios, identify the population of interest, the sample, and the unit of observation. Is the described sampling effective for studying the population of interest?

1. A polling agency conducts a survey to predict the outcome of an upcoming election. They interview 2,000 registered voters across the country.

2. Researchers are studying the effects of a new exercise program on heart health. They recruit 50 adults from a local fitness center.

Important Concepts: Inferences

- **Parameters:** A descriptive measure for a population.

Important Concepts: Inferences

- **Parameters:** A descriptive measure for a population.

- **Statistic:** A descriptive measure for a sample.

Practice

For each of the following descriptions, indicate whether the described quantity is a parameter or a statistic. Why?

1. The mean score of your friends on a recent quiz.

Practice

For each of the following descriptions, indicate whether the described quantity is a parameter or a statistic. Why?

1. The mean score of your friends on a recent quiz.

2. The percentage of customers who prefer brand A over brand B in a nationwide survey.

Practice

For each of the following descriptions, indicate whether the described quantity is a parameter or a statistic. Why?

1. The mean score of your friends on a recent quiz.

2. The percentage of customers who prefer brand A over brand B in a nationwide survey.

3. The variation of a species' weight in a national park.

The Process of Statistics

- In statistics we collect **observations** of **variables**, which constitute our **data**.

The Process of Statistics

- In statistics we collect **observations** of **variables**, which constitute our **data**.

- These data come from a **population** of interest, and make-up either a **census** or a **sample**.

The Process of Statistics

- In statistics we collect **observations** of **variables**, which constitute our **data**.
- These data come from a **population** of interest, and make-up either a **census** or a **sample**.
- Our goal is to learn information about a **parameter** of interest, using **statistics** which we can compute.

Questions from Survey

- Difference between parameters and statistics

Questions from Survey

- Difference between parameters and statistics
- “This is a lot of new material.’ ’

Questions from Survey

- Difference between parameters and statistics
- “This is a lot of new material.’’
- Observational studies vs designed experiments

Parameters versus Statistics: Parameter

A **parameter** is a quantity of interest that we typically cannot directly observe since it is the **population quantity**. It is the unknown value that we are interested in, related to the population. Consider:

- Average income for Canadian adults. (Population: Canadian adults)

Parameters versus Statistics: Parameter

A **parameter** is a quantity of interest that we typically cannot directly observe since it is the **population quantity**. It is the unknown value that we are interested in, related to the population. Consider:

- Average income for Canadian adults. (Population: Canadian adults)
- Total number of tree species in Rockwood Park. (Population: Trees in Rockwood Park)

Parameters versus Statistics: Parameter

A **parameter** is a quantity of interest that we typically cannot directly observe since it is the **population quantity**. It is the unknown value that we are interested in, related to the population. Consider:

- Average income for Canadian adults. (Population: Canadian adults)
- Total number of tree species in Rockwood Park. (Population: Trees in Rockwood Park)
- Proportion of defective phones produced by a manufacturing plant. (Population: Phones produced by the plant).

Parameters versus Statistics: Statistic

A **statistic** is a quantity of interest that we can *actually compute* since it is a **sample quantity**. It is the value that is calculated based on the observations we actually make, related to the sample. Consider:

- Average income on a survey of Canadian adults. (Population: Canadian adults; Sample: Those who are surveyed).

Parameters versus Statistics: Statistic

A **statistic** is a quantity of interest that we can *actually compute* since it is a **sample quantity**. It is the value that is calculated based on the observations we actually make, related to the sample. Consider:

- Average income on a survey of Canadian adults. (Population: Canadian adults; Sample: Those who are surveyed).
- Total number of tree species in a $500m^2$ section of Rockwood Park. (Population: Trees in Rockwood Park; Sample: Trees in the $500m^2$ grid).

Parameters versus Statistics: Statistic

A **statistic** is a quantity of interest that we can *actually compute* since it is a **sample quantity**. It is the value that is calculated based on the observations we actually make, related to the sample. Consider:

- Average income on a survey of Canadian adults. (Population: Canadian adults; Sample: Those who are surveyed).
- Total number of tree species in a $500m^2$ section of Rockwood Park. (Population: Trees in Rockwood Park; Sample: Trees in the $500m^2$ grid).
- Poportion of defective phones in a random sample of 50 phones produced by a manufacturing plant today. (Population: Phones produced by the plant; Sample: 50 phones selected).

Examples

1. The average height of all adults in the world.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.
   - Statistic. Population is students who wrote the standardized test.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.
   - Statistic. Population is students who wrote the standardized test.
3. The median age of all people living in a country.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.
   - Statistic. Population is students who wrote the standardized test.
3. The median age of all people living in a country.
   - Parameter. Population is all residents of the country.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.
   - Statistic. Population is students who wrote the standardized test.
3. The median age of all people living in a country.
   - Parameter. Population is all residents of the country.
4. The mean score of your friends on a recent quiz.

Examples

1. The average height of all adults in the world.
   - Parameter. Population is all adult humans.
2. The proportion of students who passed a standardized test in a specific school.
   - Statistic. Population is students who wrote the standardized test.
3. The median age of all people living in a country.
   - Parameter. Population is all residents of the country.
4. The mean score of your friends on a recent quiz.
   - Statistic. Population is the class in question.

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.
   - Statistic. Population is the set of all cars (in a country, region, etc.)

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.
   - Statistic. Population is the set of all cars (in a country, region, etc.)
6. The average number of students per classroom in a school district.

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.
   - Statistic. Population is the set of all cars (in a country, region, etc.)
6. The average number of students per classroom in a school district.
   - Parameter. Population is the classrooms in the school district.

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.
   - Statistic. Population is the set of all cars (in a country, region, etc.)
6. The average number of students per classroom in a school district.
   - Parameter. Population is the classrooms in the school district.
7. The range of salaries in a particular company.

Examples (Continued)

5. The percentage of cars in a parking lot that are electric vehicles.
   - Statistic. Population is the set of all cars (in a country, region, etc.)
6. The average number of students per classroom in a school district.
   - Parameter. Population is the classrooms in the school district.
7. The range of salaries in a particular company.
   - Parameter. Population is the employees of the company.

Researchers are interested in the physical activity levels of elementary aged children in New Brunswick. As a part of this research, they measure the percentage of students who participate in extracurricular activities in a specific school district.

This is best described as a statistic.

0%

This is best described as a parameter.

0%

Common Types of Variables

- Categorical versus quantitative:

Common Types of Variables

- Categorical versus quantitative:
  - Occupation (categorical) compared to years of experience (quantitative)

Common Types of Variables

- Categorical versus quantitative:
  - Occupation (categorical) compared to years of experience (quantitative)
- Discrete versus continuous:

Common Types of Variables

- Categorical versus quantitative:
  - Occupation (categorical) compared to years of experience (quantitative)
- Discrete versus continuous:
  - Age in years (discrete) compared to actual age

Common Types of Variables

- Categorical versus quantitative:
  - Occupation (categorical) compared to years of experience (quantitative)
- Discrete versus continuous:
  - Age in years (discrete) compared to actual age
- Is **education** categorical or quantitative?

Common Types of Variables

- Categorical versus quantitative:
  - Occupation (categorical) compared to years of experience (quantitative)
- Discrete versus continuous:
  - Age in years (discrete) compared to actual age
- Is **education** categorical or quantitative?
- We care since our analysis will depend on variable types!

Is "the colour of cars in a parking lot" best treated as :

A categorical variable.

0%

A discrete numeric variable.

0%

A continuous numeric variable.

0%

Is "height in centimeters" best treated as :

A categorical variable. 0%

A discrete numeric variable. 0%

A continuous numeric variable. 0%

Is "the number of books a person has read" best treated as :

A categorical variable.

0%

A discrete numeric variable.

0%

A continuous numeric variable.

0%

The Four Roles the Statistics Can Play

1. **Descriptive Statistics:** “What has happened?”

The Four Roles the Statistics Can Play

1. **Descriptive Statistics:** “What has happened?”

2. **Inferential Statistics:** “What conclusions can be drawn?”

The Four Roles the Statistics Can Play

1. **Descriptive Statistics:** "What has happened?"
2. **Inferential Statistics:** "What conclusions can be drawn?"
3. **Predictive Statistics:** "What is going to happen in the future?"

The Four Roles the Statistics Can Play

1. **Descriptive Statistics:** "What has happened?"
2. **Inferential Statistics:** "What conclusions can be drawn?"
3. **Predictive Statistics:** "What is going to happen in the future?"
4. **Prescriptive Statistics:** "What should be done?"

Example

1. What is the average income of employees in our company?

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.
3. How confident are we that the recent increase in website traffic will lead to higher sales?

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.
3. How confident are we that the recent increase in website traffic will lead to higher sales?
   - Inferential. You are trying to quantify your conclusions.

Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.
3. How confident are we that the recent increase in website traffic will lead to higher sales?
   - Inferential. You are trying to quantify your conclusions.
   - Predictive? This would be more appropriate if it was "Will an increase in traffic lead to higher sales?"

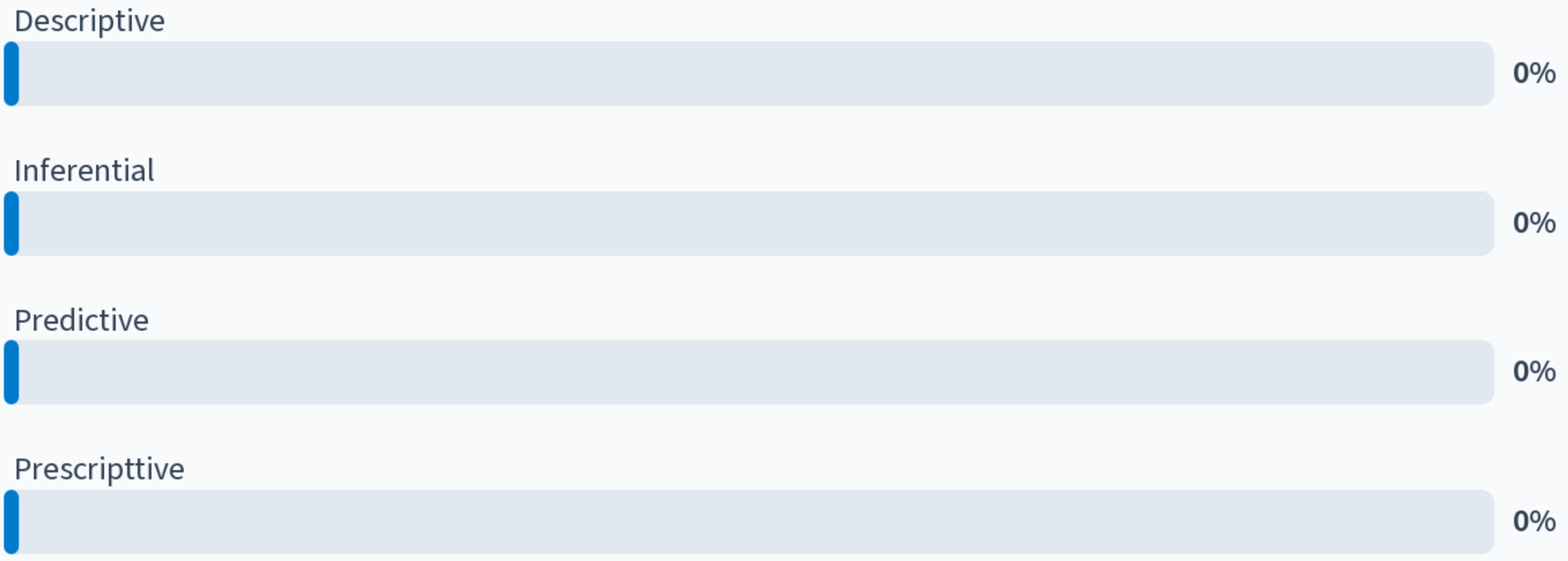
Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.
3. How confident are we that the recent increase in website traffic will lead to higher sales?
   - Inferential. You are trying to quantify your conclusions.
   - Predictive? This would be more appropriate if it was "Will an increase in traffic lead to higher sales?"
4. Should we implement a new marketing strategy based on customer feedback and buying patterns?

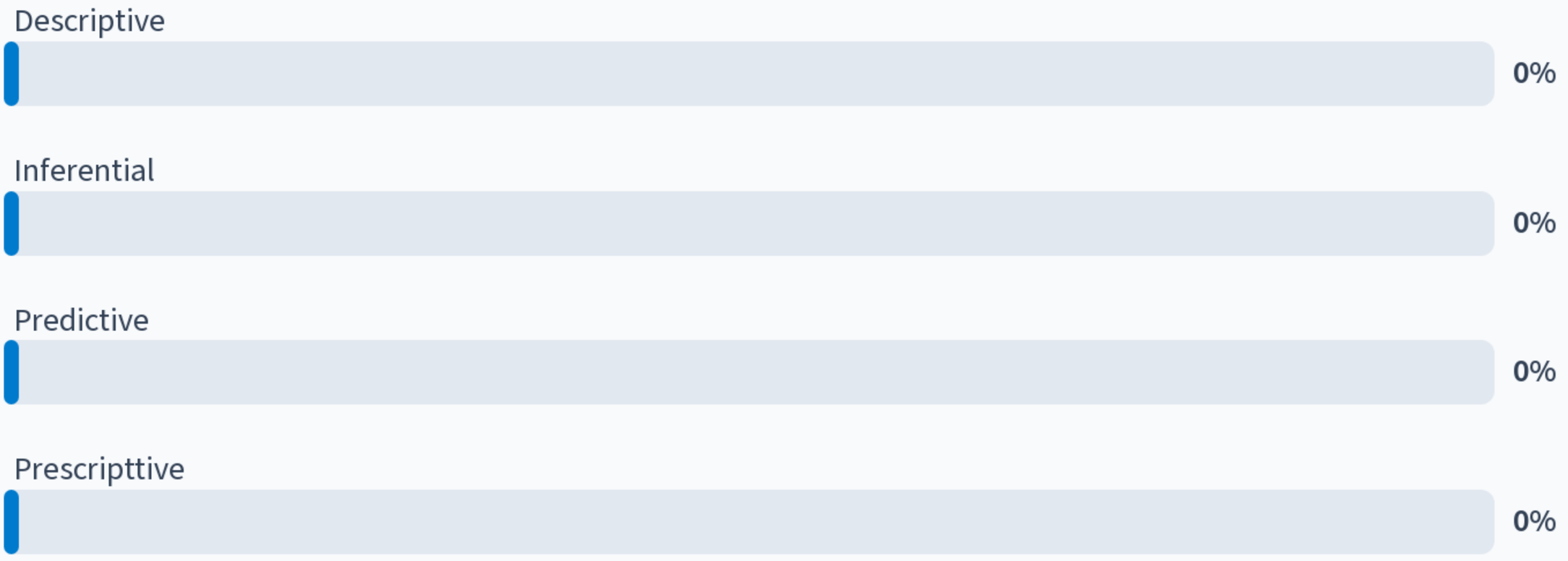
Example

1. What is the average income of employees in our company?
   - Descriptive. You are describing the way that things are.
2. What percentage of customers are likely to purchase our new product based on market research?
   - Predictive. You are attempting to infer what will happen, based on what you know/what has happened.
3. How confident are we that the recent increase in website traffic will lead to higher sales?
   - Inferential. You are trying to quantify your conclusions.
   - Predictive? This would be more appropriate if it was "Will an increase in traffic lead to higher sales?"
4. Should we implement a new marketing strategy based on customer feedback and buying patterns?
   - Prescriptive. You are trying to decide what to do, based on data.

What role of statistics answers: "Can we recommend changes to the manufacturing process to reduce defects in our products?"



What role of statistics answers: "What are the key factors influencing employee job satisfaction in our organization?"



Statistical Studies

- Studies in statistics can be divided into **observational studies** versus **designed experiments**.

Statistical Studies

- Studies in statistics can be divided into **observational studies** versus **designed experiments**.

- **Observational studies** occur when the data being collected have not been directly manipulated by the researchers.

Statistical Studies

- Studies in statistics can be divided into **observational studies** versus **designed experiments**.
- **Observational studies** occur when the data being collected have not been directly manipulated by the researchers.
- **Designed experiments** use researcher intervention, and the process of randomization, to compare populations or processes.

Observational Studies

- Harder to discern causation (but not impossible!).

Observational Studies

- Harder to discern causation (but not impossible!).
- Have to watch-out for **confounding**.

Observational Studies

- Harder to discern causation (but not impossible!).
- Have to watch-out for **confounding**.
- Samples need to be **representative**, typically.

Sample Techniques

- **Convenience sampling:** select those individuals who are available.

Sample Techniques

- **Convenience sampling:** select those individuals who are available.
- **Simple random sampling:** any observation is equally likely to be included.

Sample Techniques

- **Convenience sampling:** select those individuals who are available.
- **Simple random sampling:** any observation is equally likely to be included.
- **Stratified random sampling:** A simple random sample is performed in different groups (or strata).

Designed Experiments

- **Randomization** helps to eliminate confounding.

Designed Experiments

- **Randomization** helps to eliminate confounding.
- Normally focused on a very specific intervention.

Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.

Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.
   - Designed experiment. There is direct intervention.

Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.
   - Designed experiment. There is direct intervention.
2. In an engineering research project, engineers observe the structural behavior of different materials while varying temperature conditions.

Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.
   - Designed experiment. There is direct intervention.
2. In an engineering research project, engineers observe the structural behavior of different materials while varying temperature conditions.
   - Designed experiment. There is direct intervention.

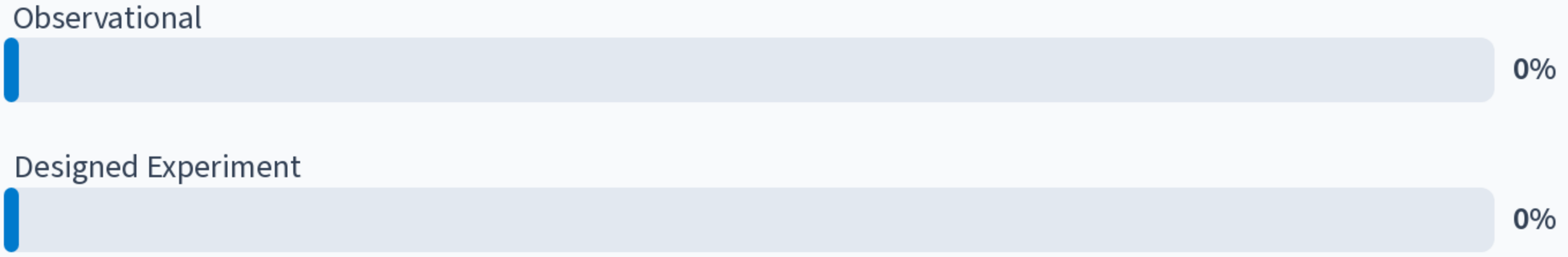
Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.
   - Designed experiment. There is direct intervention.
2. In an engineering research project, engineers observe the structural behavior of different materials while varying temperature conditions.
   - Designed experiment. There is direct intervention.
3. Researchers collect data on the smoking habits of individuals and their lung cancer incidence over a 10-year period.

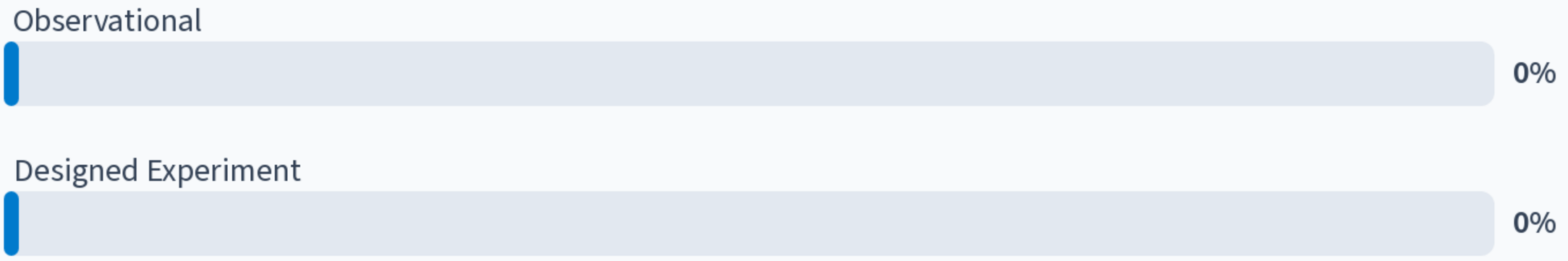
Practice

1. In a study to analyze the effects of a new drug on patient recovery, participants are randomly assigned to two groups: one receiving the new drug, and the other receiving a placebo.
   - Designed experiment. There is direct intervention.
2. In an engineering research project, engineers observe the structural behavior of different materials while varying temperature conditions.
   - Designed experiment. There is direct intervention.
3. Researchers collect data on the smoking habits of individuals and their lung cancer incidence over a 10-year period.
   - Observational study. There is no direct intervention.

What type of study is described: A study examines the impact of social media advertising on consumer purchasing behavior by randomly assigning participants to view different ads and measuring their subsequent purchases.



What type of study is described: A psychology study observes and records the behavior of children during free playtime at a daycare center to assess social interactions.



A Note on Probability

- **Probability** represents the opposite process of statistics.

- You assume that you know something about a population and then ask what is expected to be observed in samples.

Summary

- Statistics is the process of using data to make inferences about a population from a sample.
- Parameters represent population quantities of interests, statistics represent sample quantities of interest.
- Variables can be quantitative or categorical, discrete or continuous.
- Statistics can be used to describe, infer, predict, or prescribe.
- Studies in statistics are either observational or experimental.
- Probability performs the opposite process of statistics