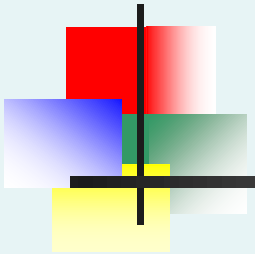
Statistics



Chapter 1

Introduction and Data Collection

Statistics

Statistics

is the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data

Population

- **Population**

  the complete collection of all individuals (scores, people, measurements, and so on) to be studied; the collection is complete in the sense that it includes *all* of the individuals to be studied

Data

- **Data**

  collections of observations (such as measurements, genders, survey responses)

Census versus Sample

Census

Collection of data from *every* member of a population

Sample

Subcollection of members selected from a population

- Sample data must be collected in an appropriate way, such as through a process of *random* selection.

- If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.

Types of Data

The subject of statistics is largely about using sample data to make inferences (or generalizations) about an entire population. It is essential to know and understand the definitions that follow.

Parameter

- **Parameter**

  a numerical measurement describing some characteristic of a population.

population

parameter

Statistic

- **Statistic**

  a numerical measurement describing some characteristic of a **sample**.

sample

statistic

Quantitative Data

- **Quantitative (or numerical) data**

  consists of *numbers* representing counts or measurements.

  Example: The weights of supermodels

  Example: The ages of respondents

Categorical Data

- **Categorical (or qualitative or attribute) data**

  consists of names or labels (representing categories)

Example: The genders (male/female) of professional athletes

Example: Shirt numbers on professional athletes uniforms - substitutes for names.

Working with Quantitative Data

Quantitative data can further be described by distinguishing between discrete and continuous types.

Discrete Data

Discrete data

result when the number of possible values is either a finite number or a ‘countable’ number

(i.e. the number of possible values is

0, 1, 2, 3, . . .)

Example: The number of eggs that a hen lays

Continuous Data

- **Continuous (numerical) data**

  result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps

  Example: The amount of milk that a cow produces; e.g. 2.343115 gallons per day

Levels of Measurement

Another way to classify data is to use levels of measurement. Four of these levels are discussed in the following slides.

Nominal Level

- **Nominal level of measurement**

  characterized by data that consist of names, labels, or categories only, and the data cannot be arranged in an ordering scheme (such as low to high)

  Example: Survey responses yes, no, undecided

Ordinal Level

- **Ordinal level of measurement**

  involves data that can be arranged in some order, but differences between data values either cannot be determined or are meaningless

  Example: Course grades A, B, C, D, or F

Interval Level

- **Interval level of measurement**

  like the ordinal level, with the additional property that the difference between any two data values is meaningful, however, there is no natural zero starting point (where none of the quantity is present)

  Example: Years 1000, 2000, 1776, and 1492

Ratio Level

❖ Ratio level of measurement

the interval level with the additional property that there is also a natural zero starting point (where zero indicates that none of the quantity is present); for values at this level, differences and ratios are meaningful

Example: Prices of college textbooks ($0 represents no cost, a $100 book costs twice as much as a $50 book)

Summary - Levels of Measurement

- **Nominal** - categories only
- **Ordinal** - categories with some order
- **Interval** - differences but no natural starting point
- **Ratio** - differences and a natural starting point

Basics of Collecting Data

Statistical methods are driven by the data that we collect. We typically obtain data from two distinct sources: *observational studies* and *experiment*.

Observational Study

- **Observational study**

  observing and measuring specific characteristics without attempting to modify the subjects being studied

Experiment

- **Experiment**

  apply some **treatment** and then observe its effects on the subjects; (subjects in experiments are called **experimental units**)

Simple Random Sample

- **Simple Random Sample**

  of n subjects selected in such a way that every possible sample of the same size n has the same chance of being chosen

Random & Probability Samples

- **Random Sample**

  members from the population are selected in such a way that each individual member in the population has an equal chance of being selected

Random & Probability Samples

- **Random Sample**

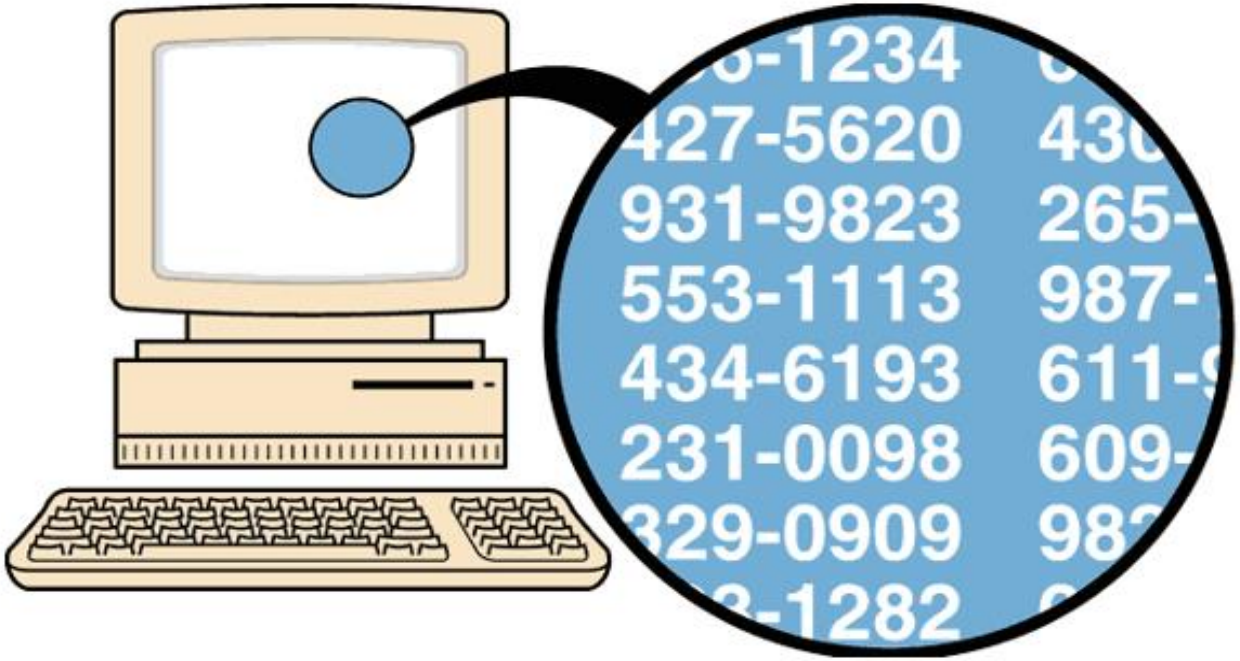
  members from the population are selected in such a way that each individual member in the population has an equal chance of being selected

- **Probability Sample**

  selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected

Random Sampling

selection so that each individual member has an equal chance of being selected

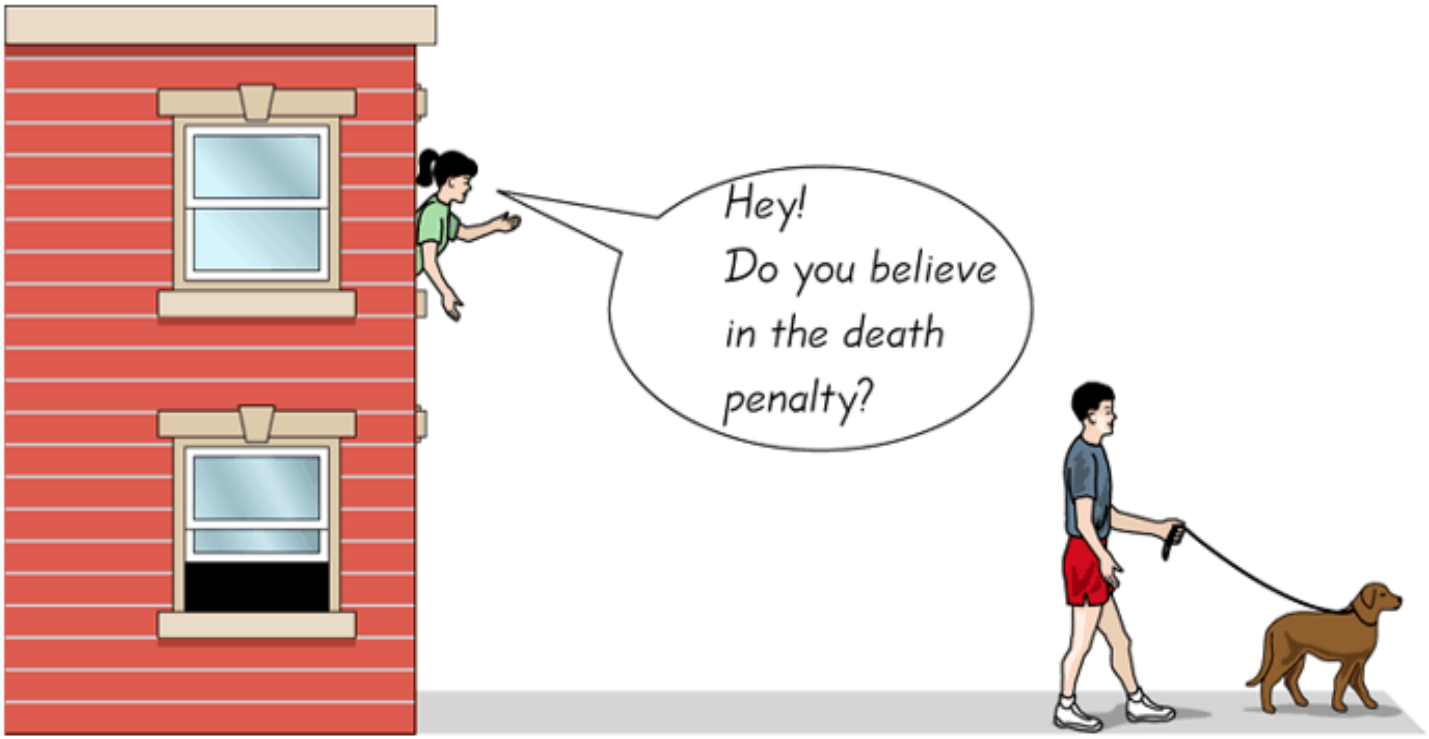


Systematic Sampling

Select some starting point and then select every *k*th element in the population

Convenience Sampling

use results that are easy to get

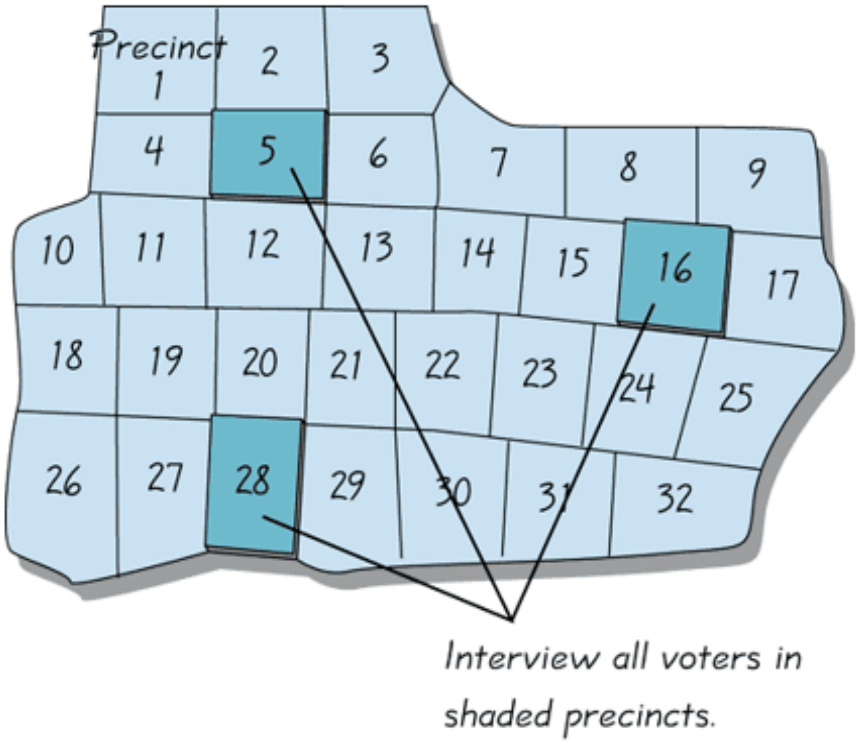


Stratified Sampling

subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum)

Cluster Sampling

divide the population area into sections (or clusters); randomly select some of those clusters; choose all members from selected clusters



Beyond the Basics of Collecting Data

Different types of observational studies and experiment design

Types of Studies

- **Cross sectional study**

  data are observed, measured, and collected at one point in time

- **Retrospective (or case control) study**

  data are collected from the past by going back in time (examine records, interviews, …)

- **Prospective (or longitudinal or cohort) study**

  data are collected in the future from groups sharing common factors (called cohorts)

Types of Statistics

- **Statistics**
  - The branch of mathematics that transforms data into useful information for decision makers.

Descriptive Statistics

Collecting, summarizing, and describing data

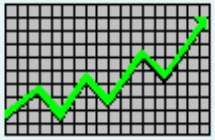
Inferential Statistics

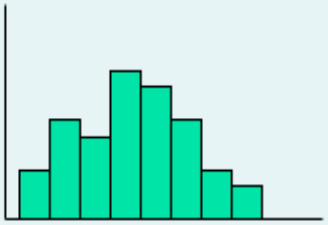
Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





- Characterize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 120 pounds

Drawing conclusions about a large group of individuals based on a subset of the large group.

Basic Vocabulary of Statistics

VARIABLE

A **variable** is a characteristic of an item or individual.

DATA

Data are the different values associated with a variable.

OPERATIONAL DEFINITIONS

Data values are meaningless unless their variables have **operational definitions,** universally accepted meanings that are clear to all associated with an analysis.

Basic Vocabulary of Statistics

POPULATION

A **population** consists of all the items or individuals about which you want to draw a conclusion.

SAMPLE

A **sample** is the portion of a population selected for analysis.

PARAMETER

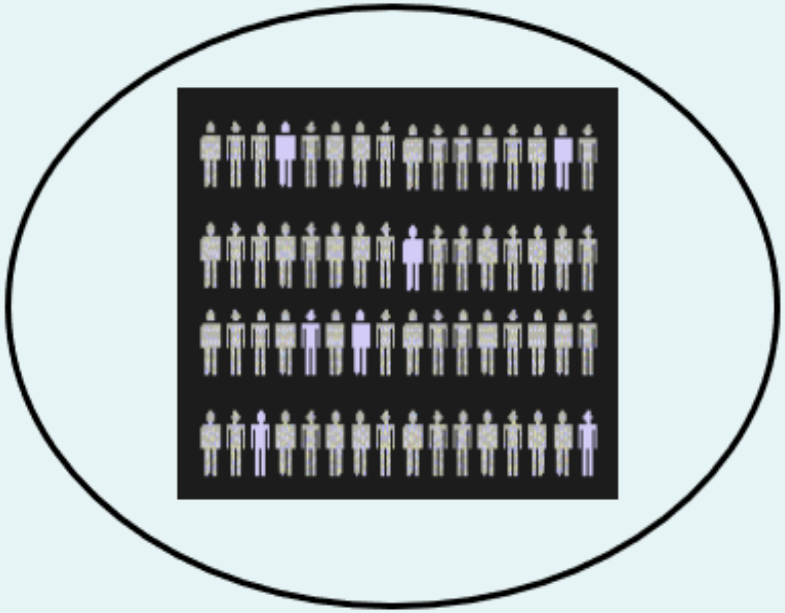
A **parameter** is a numerical measure that describes a characteristic of a population.

STATISTIC

A **statistic** is a numerical measure that describes a characteristic of a sample.

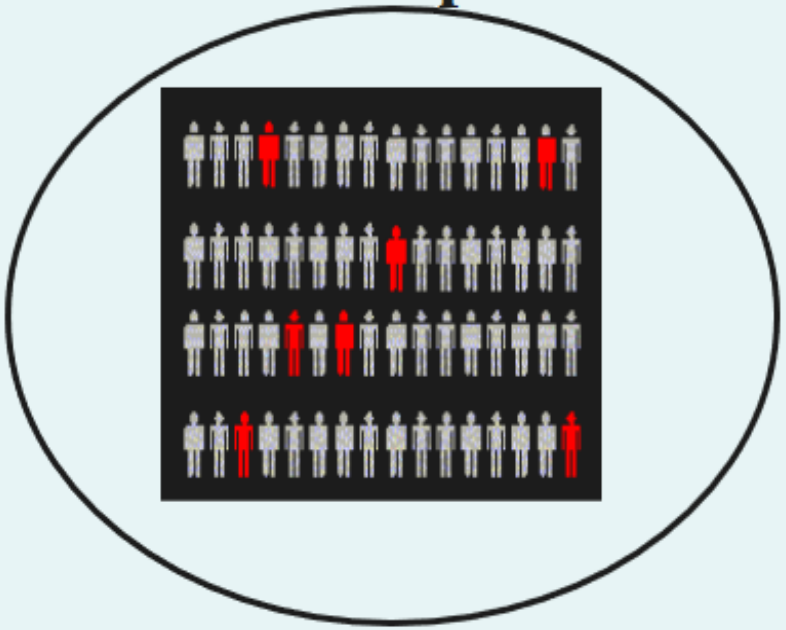
Population vs. Sample

Population



Measures used to describe the population are called **parameters**

Sample



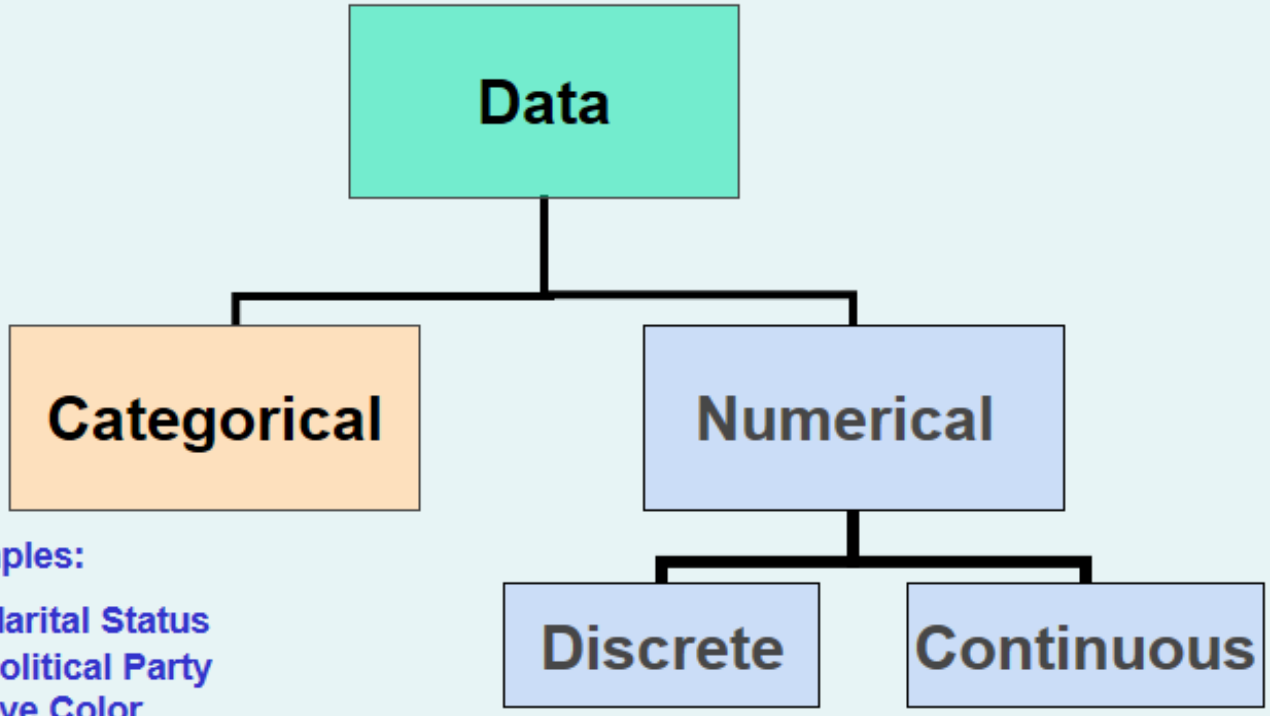
Measures computed from sample data are called **statistics**

Types of Variables

- **Categorical** (qualitative) variables have values that can only be placed into categories, such as “yes” and “no.”

- **Numerical** (quantitative) variables have values that represent quantities.

Types of Data



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)