UCLA College | Physical Sciences Statistics

STATS 10 - Chapter 1: Introduction to Data

Bingling Wang
Department of Statistics

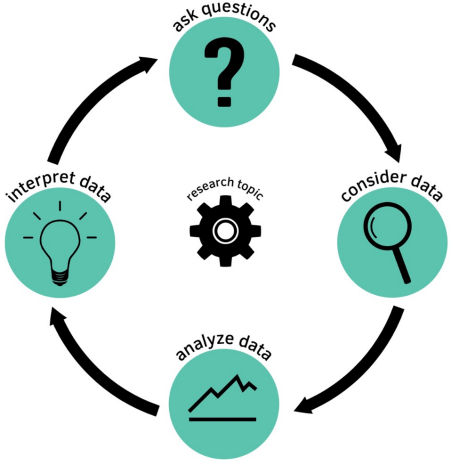
What is Statistics?

What is Statistics?

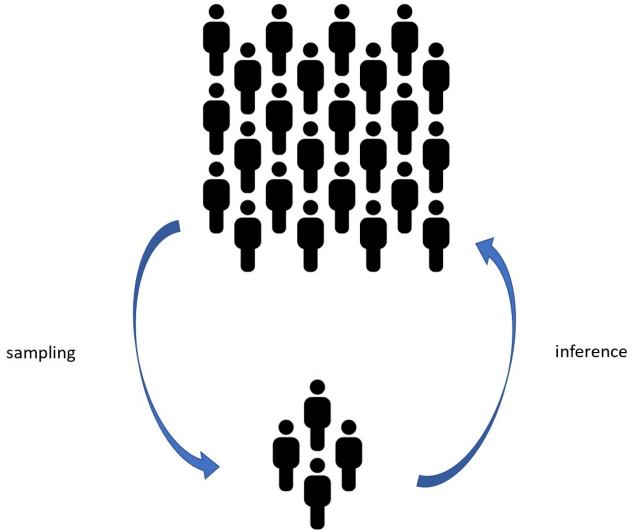
The cartoon guide to statistics -- by Larry Gonick

Statistics is the science (and art!) of data.

- A tool to understanding what data can (and cannot) tell us about the world.
- A systematic framework for quantifying uncertainty.



Population vs. Sample



Population

The entire collection of objects of interest

-- Often difficult to obtain

Sample

A portion/subset of a population of interest

-- Easier to obtain

The population is really what we want to learn about, and we learn about it by studying the data in our sample

Example

Suppose you want to find out the predominant hair color in a country. You randomly surveyed 2500 people in that country, asking them about their hair color.

- What is the population?
- What is the sample?
- What is the data collected?

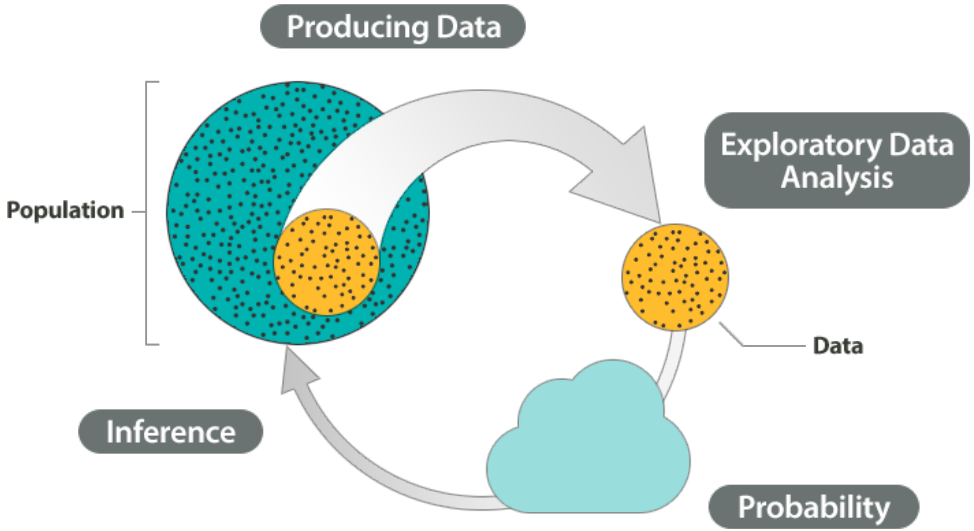
Example

Suppose you want to find out the predominant hair color in a country. You randomly surveyed 2500 people in that country, asking them about their hair color.

- What is the population? -- All the people in the country
- What is the sample? -- The 2500 people surveyed
- What is the data collected? -- The hair color of each person

The Statistical Process

1. Identify population of interest
2. Collecting data from the population
3. Exploratory data analysis
4. Make inference about the population



Exercise

Researchers want to find the opinions of U.S. adults about the death penalty. A poll was conducted (by *ABC News* and the *Washington Post*) in April 2005. A (representative) sample of 1,082 U.S. adults was chosen, and each adult was asked whether he or she favored or opposed the death penalty. The collected data were then summarized, and it was found that 65% of the sampled adults favor the death penalty for persons convicted of murder. Based on the sample result (of 65% favoring the death penalty) and knowledge of probability, it was concluded (with 95% confidence) that the percentage of those who favor the death penalty in the population is between 62% and 68%.

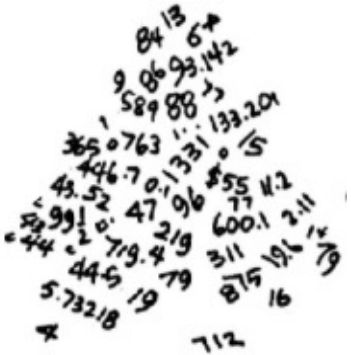
Identify each step in the Big Picture graph.

What is Data?

Data

- "Number in context" by David Moore.
  - 10.00, 9.88, 9.81, 9.75, 9.69, 9.5, 9.44, 9.31..
    Price? Distances? Birth weights?

  (Without context, we do not know what these numbers mean.)
- Information/measurements/observations that is recorded/collected.
  - Text, images, sound files etc
- A collection of data is called a **data set** (or **dataset**).



Data comes from surveys, experiments, automated data collection devices.

You leave a "data trail": Google search, credit card transactions, social media posts, etc.

We have reached a historical moment where almost everything can be thought of as data.

Variability

There is an inherent element of **randomness** in each situation that caused possibly different outcomes of something.

- Same medical treatments have different effects on different patients
- Price fluctuates for the same stock
- Draw a circle by hand.

Variability and **data** are the two major concepts underlying the study of statistics

Variables and Observations

Data are usually organized into variables and observational units.

Variables

- Any piece of information we record. Any characteristics, number or quantity that can be measured or counted.

Observations

- A set of data collected from an object/unit of interest.

Data Table

Data sets are usually stored as a data table

Sleep	Exercise	Happy
6	14	3
7	10	5
7	18	3
6.5	3	3
6.5	5	4

- Each **observational unit** is a row in the data table.
- Each **variable** is a column in the data table.

Classifying Data

Two Types of Data

Numerical data (Quantitative data)

- Tell us how much or how many.
  - Discrete
  - Continuous
- Example: number of siblings, weight, temperature, …

Categorical data (Qualitative data)

- Tell us what type or what kind.
  - Nominal
  - Ordinal
- Example: eye color, zip code, major in school, …

Variable Coding

Caution!

It is not always obvious if a variable is numeric or categorical just by whether the values are numerical or not. It is important to consider what the values represent in context.

Categorical variables can be coded as numerical values

- Area codes (e.g. 310, 626, 800)
- Weekday / Weekend → 1 / 0
- Yes / No → 1 / 0

Numerical variables can be coded into categories.

- Income 10K, 30K, 60K, 90K … → low, middle, high
- Age 1m, 3m, 9m, 20m, … → newborn, infant, toddler, etc

Numerical or Categorical?

Name	Male	Age	Height	Hair Color
Leslie	1	34	62	Blonde
Ben	0	35	70	Brown
Ron	1	49	71	Brown
April	0	20	66	Black
Andy	1	28	74	Brown

Numerical or Categorical?

Name	Male	Age	Height	Hair Color
Leslie	1	34	62	Blonde
Ben	0	35	70	Brown
Ron	1	49	71	Brown
April	0	20	66	Black
Andy	1	28	74	Brown
Categorical	Categorical	Numerical	Numerical	Categorical

Organizing Categorical Data

Frequency Tables

- **Frequencies (or counts):** the number of times a value is observed in a data set.
- **Relative frequency:** the proportion/percentage of times a value is observed in a data set.

Eye color	Frequency	Relative frequency
Brown	30	30/100 = 30%
Blue	20	20/100 = 20%
Black	25	25/100 = 25%
Green	15	15/100 = 15%
Yellow	10	10/100 = 10%

Summarize Two Categorical Variables

Example: A youth Behavior Risk Survey was conducted to study the relationship between gender and whether the respondent always wears a seat belt when riding in or driving a car.

Male	Not Always
1	1
1	1
1	0
1	0
1	0
0	1
0	1
0	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Two-way Tables

Summarize two categorical variables, display frequency of combinations of categories.

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Male

Not Always	1	0	Total
1	2	3	5
0	3	7	10
Total	5	10	15

Two-way Tables

Summarize two categorical variables, display frequency of combinations of categories.

1. How many observations are in the sample?
2. How many people do not always wear seat belts?
3. How many of those who always wear seat belts are males?
4. What percent always wear seat belts?
5. Are males in the sample more likely to take the risk of not wearing a seat belt?
6. Being a man results in not wearing seat belts.

   Yes or No?

Male

Not Always	1	0	Total
1	2	3	5
0	3	7	10
Total	5	10	15

1. 15
2. 5
3. 3
4. 2/3, 66.7%
5. 40% > 30%
   Yes
6. NO!

Comparing Data

Data summary: number of sports-related injuries that were treated in U.S emergency rooms in 2009

Which team sports is the most dangerous?

Sport	Injuries
Baseball	165,842
Basketball	501,251
Bowling	20,878
Football	451,961
Ice hockey	19,035
Soccer	208,214
Softball	121,175
Tennis	23,611
Volleyball	60,159

Comparing Data

Data summary: number of sports-related injuries that were treated in U.S emergency rooms in 2009

Which team sports is the most dangerous?

Sport	Injuries	Participants
Baseball	165,842	11,500,000
Basketball	501,251	24,400,000
Bowling	20,878	45,000,000
Football	451,961	8,900,000
Ice hockey	19,035	3,100,000
Soccer	208,214	13,600,000
Softball	121,175	11,800,000
Tennis	23,611	10,800,000
Volleyball	60,159	10,700,000

Comparing Data

- The groups need to be similar.
- Percentages or rates are often better for comparisons.

Sport	Participants	Injuries	Rate of Injury per Participant	Rate of Injury per Thousand Participants
Baseball	11,500,000	165,842	0.01442	14.42
Basketball	24,400,000	501,251	0.02054	20.54
Bowling	45,000,000	20,878	0.00046	0.46
Football	8,900,000	451,961	0.05078	50.78
Ice hockey	3,100,000	19,035	0.00614	6.14
Soccer	13,600,000	208,214	0.01531	15.31
Softball	11,800,000	121,175	0.01027	10.27
Tennis	10,800,000	23,611	0.00219	2.19
Volleyball	10,700,000	60,159	0.00562	5.62