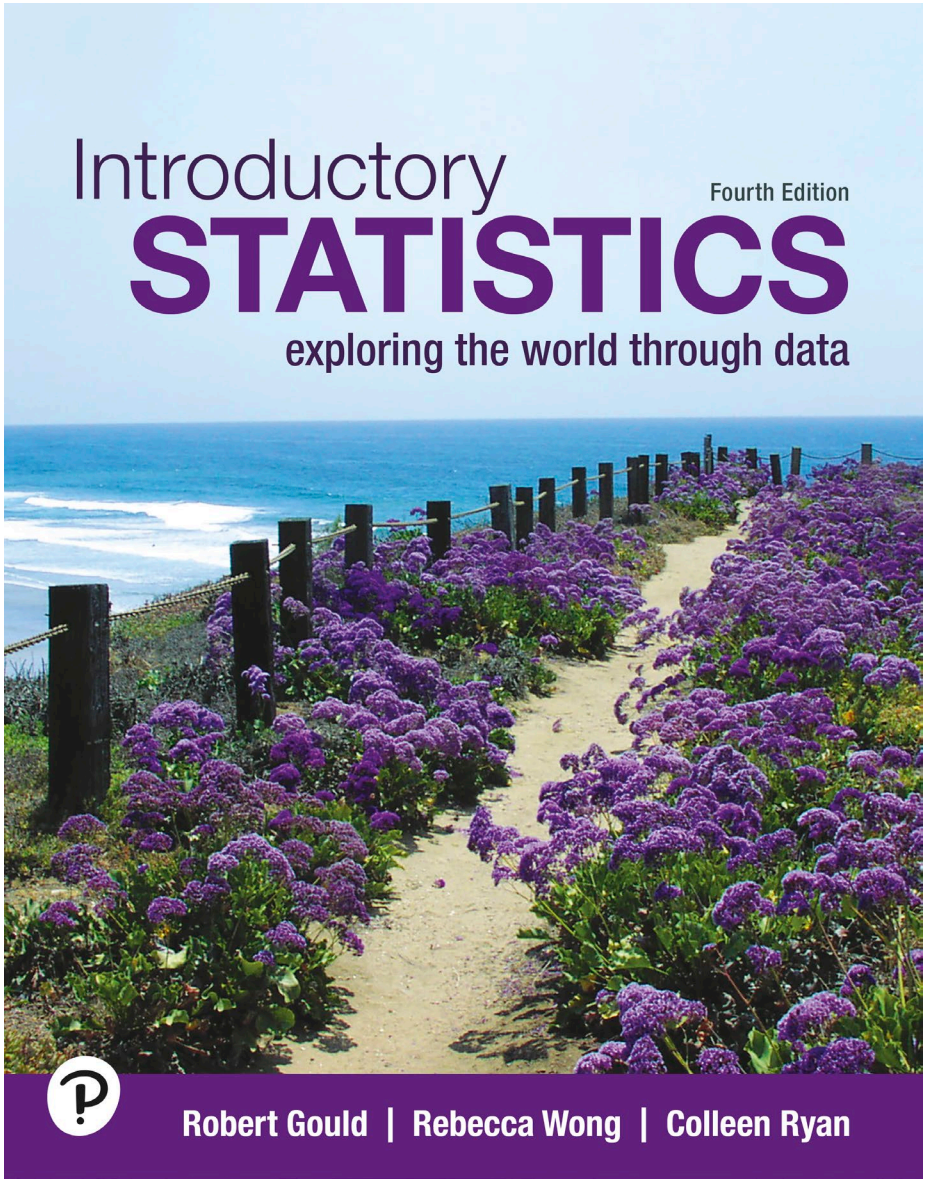
Chapter 1

Introduction to Data



What Is Statistics?

Statistics is the science of collecting, organizing, summarizing, and analyzing ***data*** to answer questions and/or draw conclusions.

Data are numbers in context: *observations, measurements, classifications*. Example:

98.6 98.4 98.6 98.4 98.0 98.4 98.0 98.4 99.0 98.6

These numbers represent the body temperature of 10 randomly selected normal and healthy adults.

How can Statistics help us?

Statistics allows us to:

- Explore the world around us.
- Use evidence to check whether our beliefs are true.
- Find patterns to lead to discoveries.
- Share new discoveries with others.

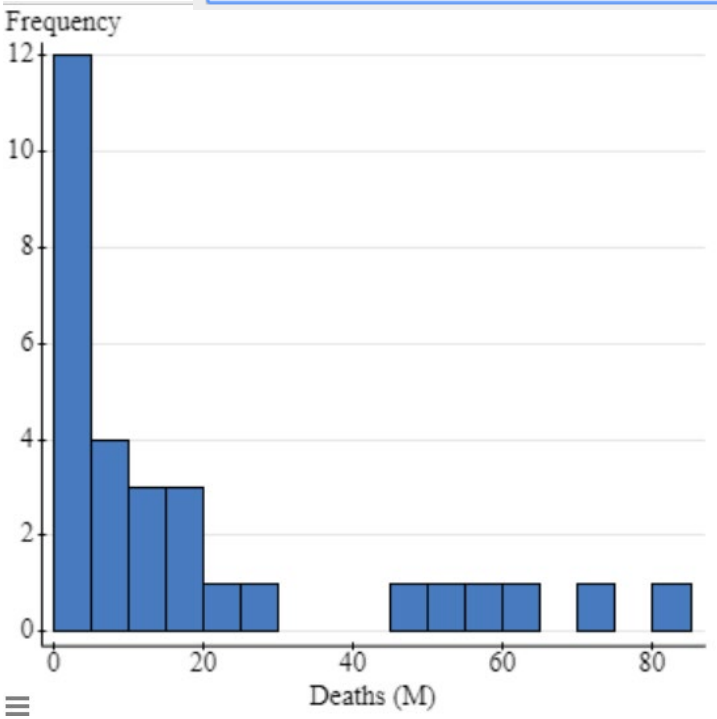
Keep in Mind:

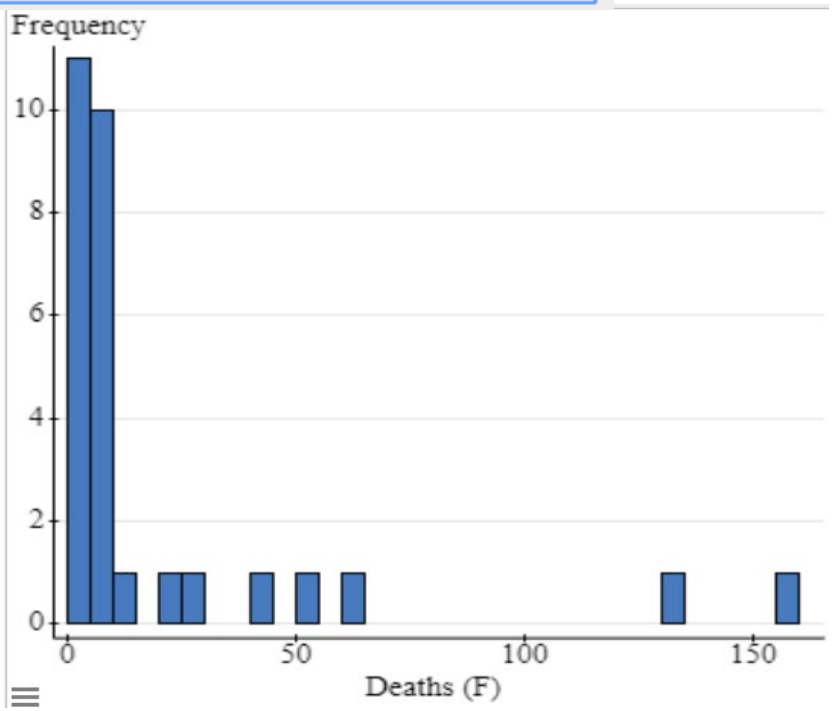
- Statistics must be used carefully.
- Inappropriate use will result in inaccurate beliefs.
- Results are always uncertain.

Are female hurricanes more deadly than male hurricanes?

Summary statistics:

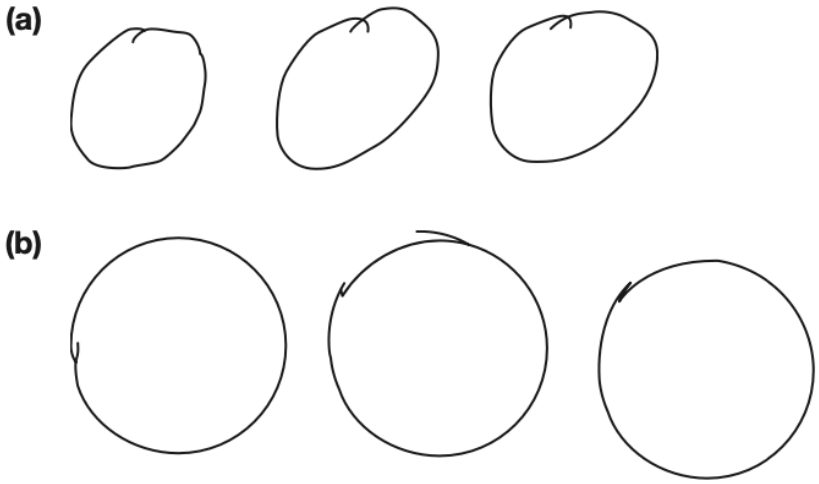
Column	Mean	Median	Range	Min	Max	Mode
Deaths (M)	17.7	6.5	84	0	84	1
Deaths (F)	177.03226	6	3056	1	3057	3





Statistics Rests on Two Major Concepts

Variation: differences or changes in an item



Draw circles by hand and by outlining a coin

Pay attention to letters u, n, o

Statistics Rests on Two Major Concepts

Data

- Observations gathered to draw conclusions
- Context is important

Example:

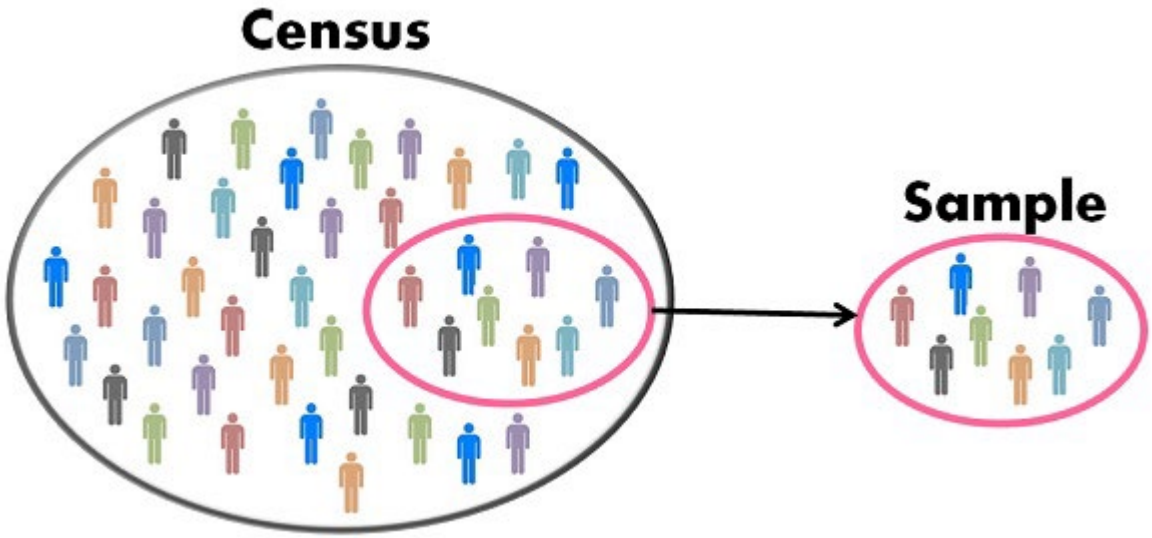
10.00, 9.88, 9.81, 9.81, 9.75, 9.69, 9.5, 9.44, 9.31

- Weights in pounds of the nine heaviest babies in a sample of babies born in North Carolina
- Units matter

Data Analysis

Data analysis examination of collected data to look for patterns/summaries (shape, center, spread, outliers, trends) to capture the essence of what the data is telling us about the real world.

Know the difference between Populations and Samples



Population vs Sample

- **Population:** the *complete* set of people or things being studied.
- **Sample:** a subset of the population from which the raw data are actually obtained.
- **Population parameter/variable:** characteristic of the population that is being studied.
- **Sample statistics:** numbers or observations that summarize the raw data.

Population vs Sample

A researcher wants to find out the average daily screen time of high school students in a city. They survey 200 students chosen randomly from all high schools in that city.

Population: all high school students in the city

Sample: the 200 students that were surveyed

Population parameter/variable: average daily screen time of all high school students in the city

Sample statistic: average daily screen time calculated from the 200 surveyed students

Classifying Data

- **Numerical data** describe quantity or measurement.

Examples: temperature, height, time, weight, distance.

- **Categorical data** describe quality or classification.

Arithmetic do not make sense on categorical data, even if the data are made of numbers.

Examples: letter grade, zip code, student ID, license plates.

Classifying Data

Classify the following variables as numerical or categorical:

Variable	Numerical	Categorical
Height of a bridge	X	
GPA	X	
Letter grade in class		X
Hours worked each week	X	
Type of pets owned (cat, dog, etc.)		X
Flower varieties planted in a garden		X
Dots on the sides of a die	Depends on what you are using it for!	

Storing Data: Coding

Coded Data

- Using numbers to record categorical data
- Can make reading the data easier
- You can sort it easily

Example: Gender is female?

0 = No, 1 = Yes

Female
1
0
1
1
1
1

Note: The data is still categorical, not numerical.

0 and 1 are labels, not values.

Storing Data: Stacked

Stacked Data

- Data values stored in a spreadsheet format
- Each row contains data for a single individual
- Can store many variables!

Example:

- Each row corresponds to one infant
- Variables:
  - Birth weight
  - Gender
  - Smoking status of the mother

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

Storing Data: Unstacked

Unstacked Data

- Data values are stored in two columns
- Each column is a variable from a different group
- Can only store data for two variables
- *Info in a row does NOT correspond to the same individual*

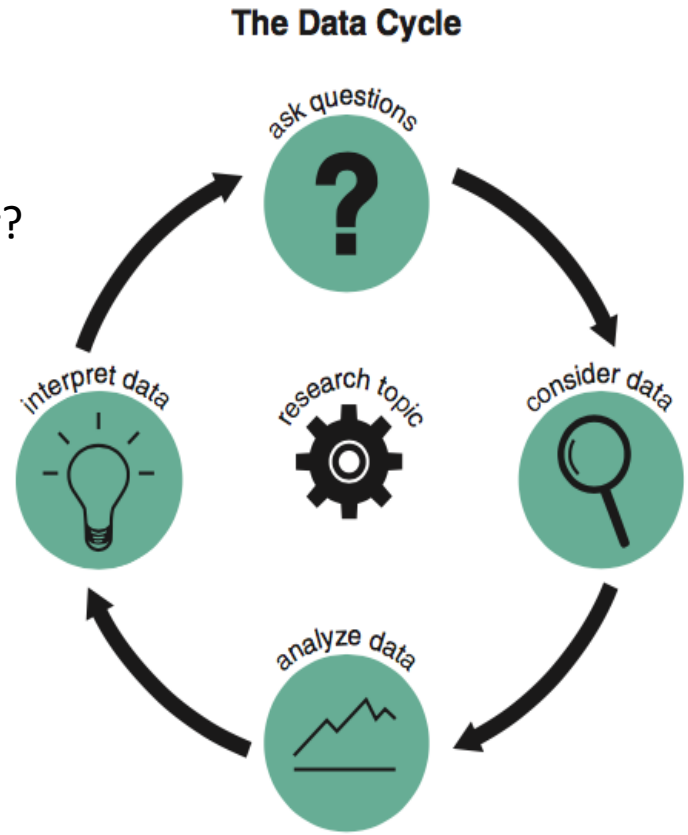
Men's Heights	Women's Heights
70	59
68	70
71	61
	62

Investigating Data

This diagram represents the cycle of a statistical investigation.

Ask questions, such as:

- What makes a runner faster or slower?
- Do cell phones cause cancer?



The Data Cycle

The following order is a useful way to plan the analysis:

1) **Ask Questions** – It is important to ask *good* questions
2) **Consider Data** – determine which data is available to answer the question
3) **Analyze Data** – Begin by visualizing the data
4) **Interpret Data** – interpret your analysis