

Statistics

the study of information

Data = information

Background – History

- Statistics was first developed in the 17th and 18th Centuries as a way of solving problems
 - Why were people dying of cholera?
 - How could the king of France be more successful at cards?
 - Can the weather be predicted?
- Statistics grew rapidly and became a formal branch of mathematics during the Industrial Revolution of the 19th Century.
 - Factories wanted to produce quality products at a minimum cost.
 - The insurance industry developed by using statistics to assess risk.
- With the information age (late 20th Century), Statistics became essential for interpreting vast amounts of data.
 - Public opinion polls
 - Forensic science
 - Need for accountability
- In the 21st Century, technology is making statistics easier and more accessible to everyone.

2 main branches of statistics

Descriptive Statistics

describes data

- organizes in a meaningful way
- tells what is “average”
- tells how spread out the data is
- tells how an individual piece of data compares to the others

Inferential Statistics

interprets data

- makes generalizations and predictions based on the data
- uses data to estimate overall trends
- determines whether a result is “significant”

Significant

A result is SIGNIFICANT in statistics if it is ***unlikely to have happened by chance***

- “Significant” is the single most important term in statistics.
- **IMPORTANT**: “significant” does **NOT** necessarily mean big or important

Population

a large group we want to find out about

Parameter

a piece of information about a population

- It is often impossible or impractical to find parameters.

Sample

a small group that we use to represent a population

Statistic (singular)

a piece of information about a sample

- We use **statistics** (information about samples) to estimate **parameters** (information about populations).

REMEMBER:

<u>P</u> opulation	→	<u>P</u> arameter
<u>S</u> ample	→	<u>S</u> tatistic

4 main ways to gather data:

Sampling

- Using a small group to represent the population
- Most common method used in statistics, and the method we will focus on
- **If done carefully**, sampling can give useful information about a population.

Census

- includes the **entire population**
- usually not practical, and often impossible

Experiment

1. Divide the sample into two (or more groups)
2. **Treat the groups differently.**
3. Check for differences.

Terms involved in experiments ...

- Control group vs. experimental group
- Observation vs. treatment
- Placebo
- Double-blind

Simulation

- A small-scale model of a real-world situation
- In modern times, most often done on computer
- Could also be a constructed model, like a science fair project

Ideally, a sample should be **representative** of the population.

- characteristics of the sample should be similar to those of the population
- A “good” sample is representative; a “bad” sample isn’t.

Types of samples:

Convenience Sample

- A sample chosen because it is **EASY** to obtain
- Often a “bad” sample
- However, for casual purposes, it’s often the best we can obtain.

Systematic Sample

- Ahead of time you decide on a **SYSTEM** for how you will choose the sample
 - Ex.: Numbering off, and picking every 7th person.
- **IMPORTANT:** Before you begin, you’ve decided who will and won’t be part of the sample.

Stratified Sample

1. Divide the population into groups, based on some characteristic (race, sex, etc.)
 2. Choose samples from each group (usually in the same proportion the groups come up in the population).
- Makes a formal effort to include a variety, ...
 - BUT ... May not be representative of the population on any other characteristic besides the one you chose

Cluster Sample

- Chose one (or a few) **CLUSTERS**
 - complete sub-groups
 - close to each other in either time or space
 - Usually clusters are places (town, zip code, school, etc.)
- Take a **CENSUS** of the clusters (include everybody).
- The clusters are your sample.

Random Sample

- IMPORTANT: Random does **NOT** mean “haphazard”.
- Everyone in the population has an equal chance of being selected.
- You find an organized way ahead of time that guarantees the selection is fair.
- You have no idea ahead of time who will or won't be in your population.
- Random samples are the “best” samples.
 - Have the best chance of being truly representative of the population

BUT ...

- Random samples are often difficult to gather.

To get random samples, you can use:

- Games of chance
- Random number tables
- Random number generators (on calculators or computers)

In most real-world situations, various types of sampling are combined.

- For example, **stratified random** samples are often used in market research and by the government.
- It's rare to have only one of the types of sample completely by itself (but in book problems they will assume there is just one)

4 levels of data:

Nominal Data

- Divide the sample into **named categories**
 - No category is considered better or worse than another
- Count up how many are in each category
- Provides the least information of any level of measurement

Ordinal Data

- Divide into categories that can be arranged in a definite **ORDER** from bottom to top.
 - There is a definite continuum, with one category above another.
- Count up how many are in each category.
- Still just a count, but the order provides more information.

Nominal and ordinal data are also called **qualitative data**, because they describe qualities (named characteristics).

Interval data

- The data is **NUMBERS**
 - but ...
- there is **NO** definite **ZERO**.
 - You can compare the numbers by subtracting (how many more one is than another).
 - You can't divide (can't tell how many times one number is of another).
- Most common examples are **time** and **temperature**
- Things that can involve **negative numbers** are also often interval data.
- In interval data, it doesn't make sense to say “twice” or “half” when you compare the numbers.

Ratio data

- **NUMBERS**, with a **DEFINITE ZERO** (or bottom of some sort).
 - You can both subtract and divide the numbers
 - Terms like “twice” and “half” make sense with ratio data.
- Most numerical data is ratio data.
- Considered the “best” kind of data
- Usually the hardest kind of data to gather

Interval and ratio data are also called **quantitative data**, because they describe quantities (numbers).