Introduction

ANA 500 – Foundations of Data Analytics

Module 1: Lecture 1

Outline

- Basic topics of statistics
- Statistical analysis steps

Introduction

What are **data**?

- Pieces of information
- Can be qualitative or quantitative

What is **data analytics**?

- Using data analysis to inform decisions

Introduction

What **are statistics**?

- Calculations derived from a dataset used to convey important features of the data in a concise way.

What is **statistics**?

- The science focused on collecting, analyzing, interpreting, and presenting data

We never truly know the 'exact' answer to anything.

- There is always some measurement error
- Some processes may have more measurement error than others

Descriptive vs Inferential Statistics

Descriptive statistics

Descriptive statistics describe the characteristics of a particular dataset.

- Presenting, organizing, and summarizing data

inferential statistics

- Inferential statistics uses a sample of data to draw conclusions about an underlying population of interest

Example: 500 U.S. college students' GPA

- **Descriptive statistics** the average students' GPA by those 500 students.
- **Inferential statistics:** using the sample of 500 U.S. college students' GPA to draw conclusions about all college students' average GPA in the U.S.

Descriptive statistics-- “Getting to know your data”

- Describe the characteristics of a dataset (Numerical, graphical, tabular)
- **Frequency distribution**
- **Measure of Central Tendency**
  - Mode, Median, Mean
- **Measures of Variability (**tell you how spread out the values in a data set are)
  - Range, Standard deviation, Variance

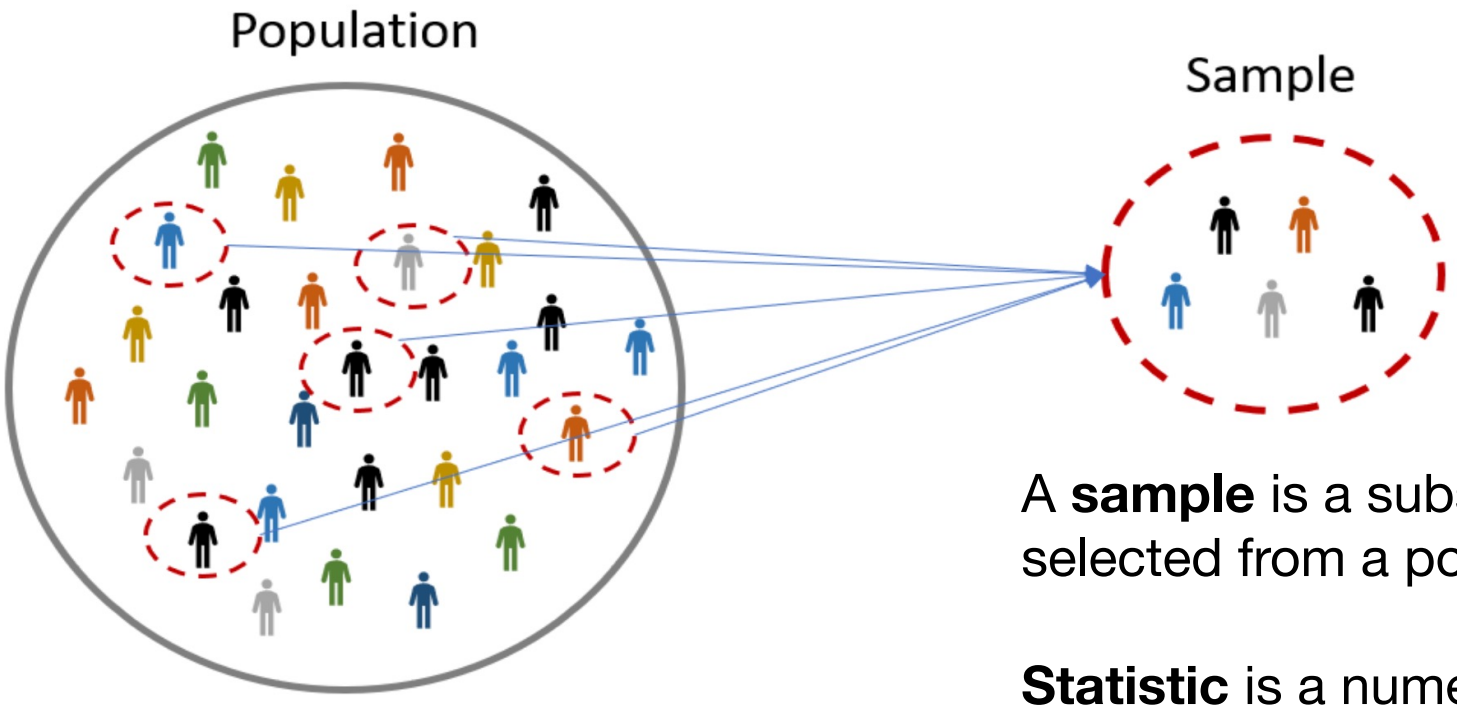
Inferential Statistics

- The goal of inferential statistics is to draw conclusions from a sample and generalize them to a population.

- Two main uses:
  - Making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
  - Testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).

- Draw a representative sample from that population.

Population and Sample



A **population** is the entire group that you want to draw conclusions about.

Parameter is a numerical value describing a characteristic in the population. (unknown)

A **sample** is a subset of data selected from a population.

Statistic is a numerical value calculated using a sample of data. (Known)

Parameter and Statistics

Statistics

$\bar{x}$: **sample** mean

s : **sample** standard deviation

s^2 : **sample** variance

$\hat{p}$: **sample** proportion

n : **sample** size

Parameter

μ : **population** mean

σ : **population** standard deviation

σ^2 : **population** variance

p : **population** proportion

N : **population** size

Sampling

We cannot often collect data from a whole population.

- Too expensive
- Unrealistic to survey all of the population
- Too time-consuming

Sampling is the process of selecting units (e.g., people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen.

Sampling

The purpose of sampling is to achieve *generalizability.*

To ensure a high level of *generalizability*, the sample should be a good representation of the population. (Valid statistical inference requires a representative sample of data)

Probability sampling methods

- is any method of sampling that utilizes some form of random selection (e.g., simple random sampling, stratified random sampling
- reduces the risk of sampling bias and enhances both internal and external validity.

Non-probability sampling

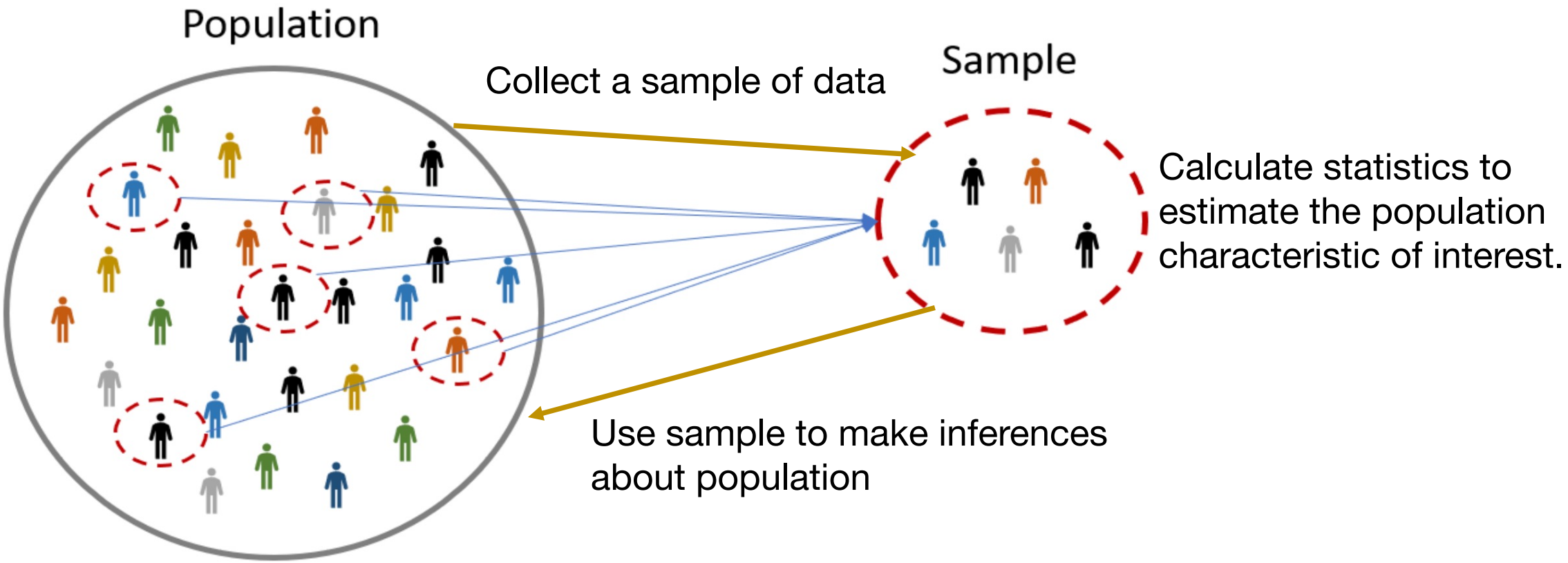
- Non-probability samples are chosen for specific criteria; they may be more convenient or cheaper to access.
- researchers use it widely for qualitative research.

Sampling Error

- **A sampling error** is the difference between a population parameter and a sample statistic.
  - For example, the sampling error is the difference between the mean of the 500 college students' GPA and all college students' average GPA in the entire United States.

- Sampling errors happen even when you use a randomly selected sample. This is because random samples are not identical to the population in terms of numerical measures like means and standard deviations.
- In general, the larger the sample size, the smaller the sampling error.

Inferential Statistics



- We want to know a numerical characteristic of the population

Data matrix

In a data matrix, each row represents an observation, and each column represents a variable.

A variable is a characteristic or measurement that can be determined for each object in the population.

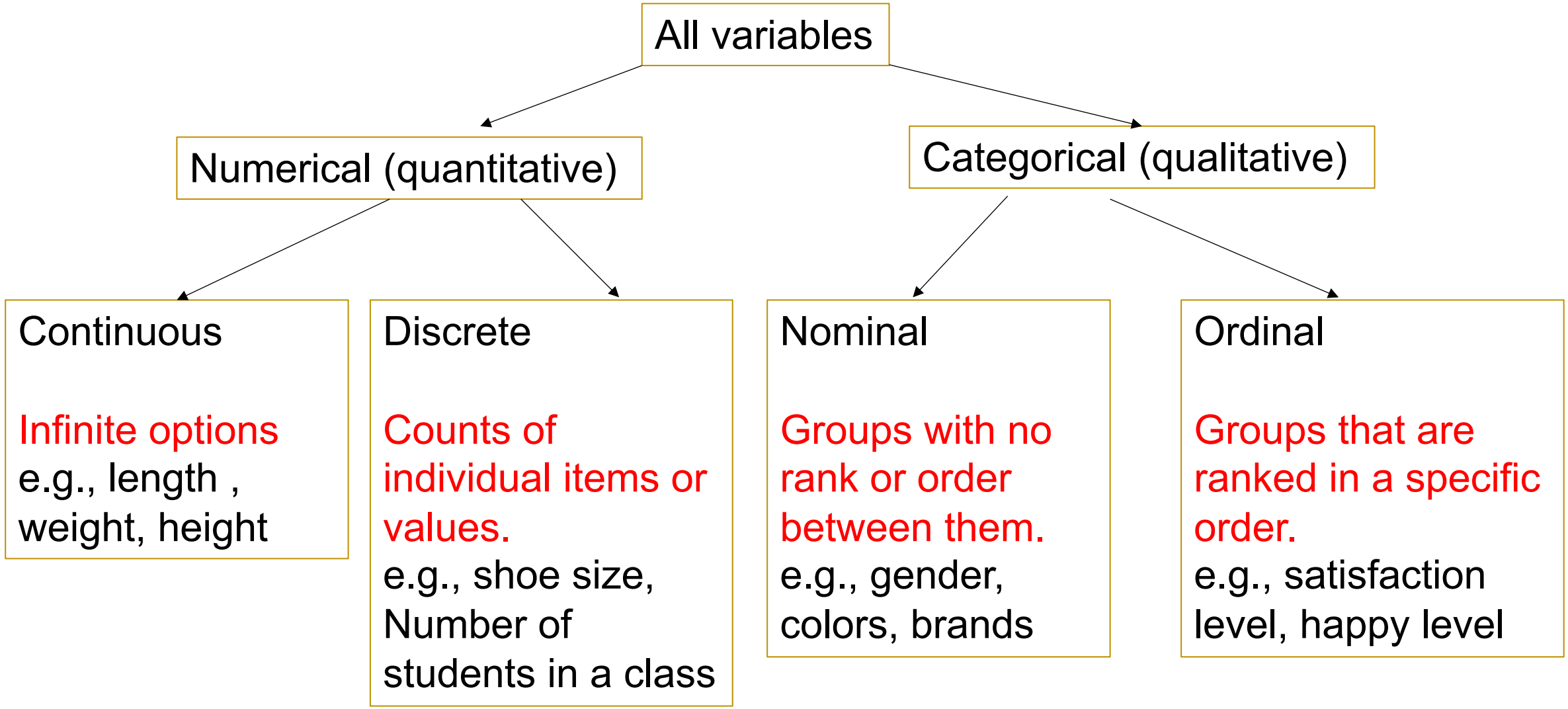
Example: Data collected on students in a statistics class on a variety of variables:

Variable

observation →

Student	Gender	Age	GPA
1	Male	23	3.2
2	Female	25	3.5
3	Male	22	2.9
4	Male	26	3.9
…	…	…	…
50	Female	24	3.0

Types of variables



Levels of measurement (Scales of Measurement)

Qualitative data

Nominal	**Ordinal**
Categories, No natural order or ranking	**Ordered categories**
Gender	Language ability (e.g., beginner, intermediate, fluent)
Ice cream flavor preference	Likert-type questions (e.g., strongly disagree to strongly agree)
Blood type	Education level (“high school”, “BS”, “MS”, “PhD”)
zip code	Happy level (“Very unhappy”, “Unhappy”, “Ok”, “Happy”, “Very happy”)

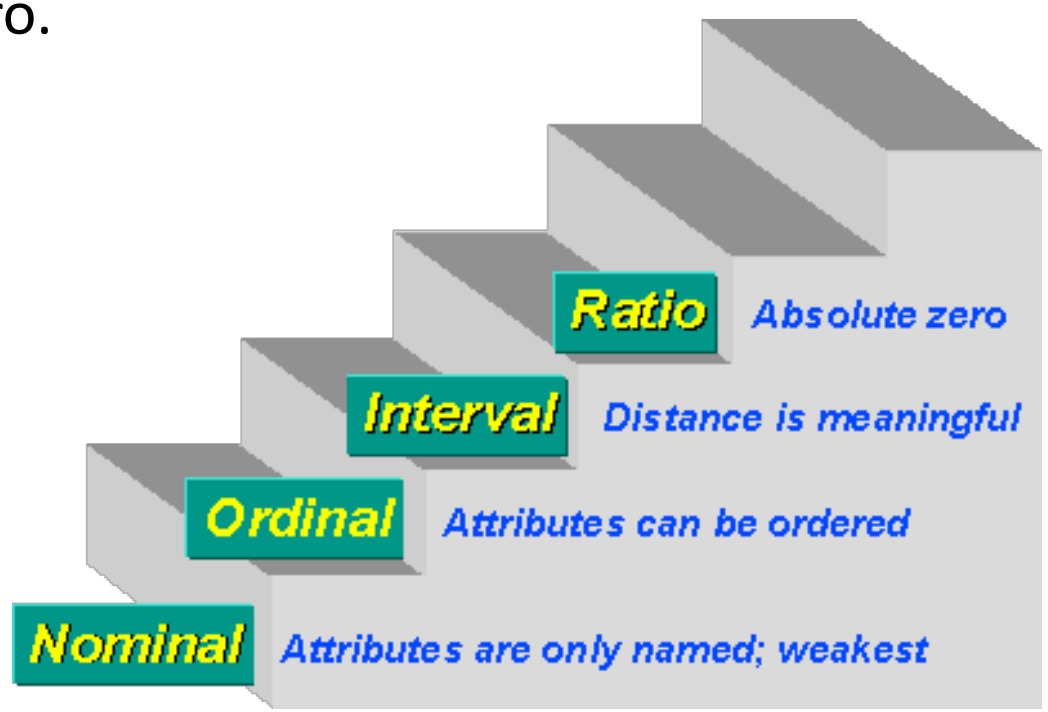
Levels of measurement (Scales of Measurement)

Quantitative data

Interval Data	Ratio Data
Equal spacing **No true zero starting point or or fixed beginning** **cannot calculate Ratios**	**True zero exists**
SAT score (200-800) Temperature (in Fahrenheit or Celsius) IQ test (intelligence scale)	Height, Weight Income earned in a year Number of children

Levels of measurement (Scales of Measurement)

Nominal: the data can only be categorized
Ordinal: the data can be categorized and ranked
Interval: the data can be categorized, ranked, and evenly spaced
Ratio: the data can be categorized, ranked, evenly spaced, and has an absolute zero.



Dependent and Independent variables

Dependent Variable:

(outcome, response)

This is the variable whose values we want to explain or predict.

Its values depend on something else.

We denote it as Y

Independent Variable:

(cause, treatment)

This is the variable that explains the other one.

Its values are independent.

We denote it as X

Dependent and Independent variables

Study *Examples:*	**Independent Variable** *Examples:*	**Dependent Variable** *Examples:*
A scientist studies the impact of a drug on cancer	Administration of the drug (such as dosage or timing)	The drug's impact on cancer
Whether education level impacts how much a person earns in their job	Highest level of educational attainment	Earnings (salary or wages)
Whether lack of sleep significantly affects learning in 10-year-old boys.	Sleep time	Learning outcome in 10-year-old boys
Whether stressful experiences significantly increase the likelihood of headaches.	Stressful experiences	The likelihood of headaches

Statistics Analysis Step

- Write your Research Questions/hypotheses and plan the research design
- Get a dataset
- Summarize the data with descriptive statistics
- Test hypotheses with inferential statistics
- Interpret the results

Statement of Research Questions

- What topic are you interested in researching?
  - What is happening around you?
  - Literature in your field (extend or refine previous studies)
  - Curiosity and imagination

- A research question guides and centers your research.
  - It should be clear and focused.
  - Be careful to avoid the “all-about” paper and questions that can be answered in a few factual statements.

Statement of Research Questions

- Choose the variables for answering research questions and determine their level of measurement.
  - Which variable is the dependent variable? Which variable/variables is/are the independent variable/variables?
  - Nominal, ordinal, interval, or ratio?
- Research question=Interrogative statements or questions
- Hypothesis=Prediction researcher holds about relationships among variables

Research Questions examples

Examples:

- Is there a relationship between parental income and college grade point average (GPA)?
- Will internet advertising increase the company A's revenue?
- How COVID-19 pandemic has changed the food consumption of adults?
- Do teacher support, conceptual teaching and procedural teaching influence students' mathematics achievement after controlling for family SES and student prior achievement?

Hypotheses examples

Examples:

- **Null hypothesis:** A 5-minute meditation exercise will have no effect on math test scores in teenagers.
- **Alternative hypothesis:** A 5-minute meditation exercise will improve math test scores in teenagers.

- **Null hypothesis:** Parental income and GPA have no relationship with each other in college students.
- **Alternative hypothesis:** Parental income and GPA are correlated in college students.

Collect data from a sample or choose a dataset from ANA500

- Describe your data source.
  - When were the data collected?
  - What is the unit of observation (i.e. element/entity)?
  - For what purposes were the data collected?
  - Are there any well-known general limitations/issues with the dataset?

Summarize your data with descriptive statistics

- Cleaning the data
  - e.g. missing observations, outliers
- Calculate descriptive statistics
  - Mean, mode, median, standard deviation, variance, etc. for numerical variables
  - The percentage/proportion for categorical variables
  - Correlation matrix
  - Appropriate charts and graphs.

Analyzing The Data

- Hypothesis testing
  - Using data from a sample, you can test hypotheses about relationships between variables in the population.
- **Comparison** tests
  - Comparison tests usually compare the means of groups
    - A t test is for exactly 1 or 2 groups when the sample is small (30 or less).
    - An ANOVA is for 3 or more groups.
- **regression** models:
  - A regression models the extent to which changes in a predictor variable results in changes in outcome variable(s).
    - A simple linear regression includes one predictor variable and one outcome variable.
    - A multiple linear regression includes two or more predictor variables and one outcome variable.

Interpreting The Results

- Interpret the results and drawn meaningful insights
- Create visualizations
- The importance of the study
- Discuss limitations of your findings

Next Lecture

- Descriptive statistics