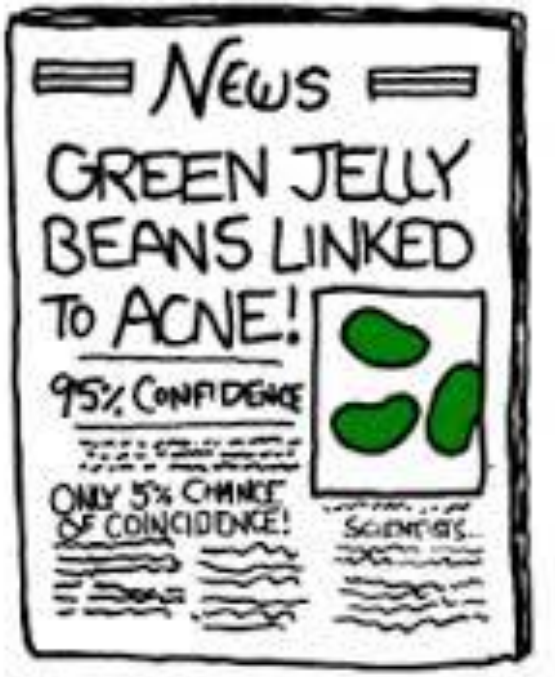
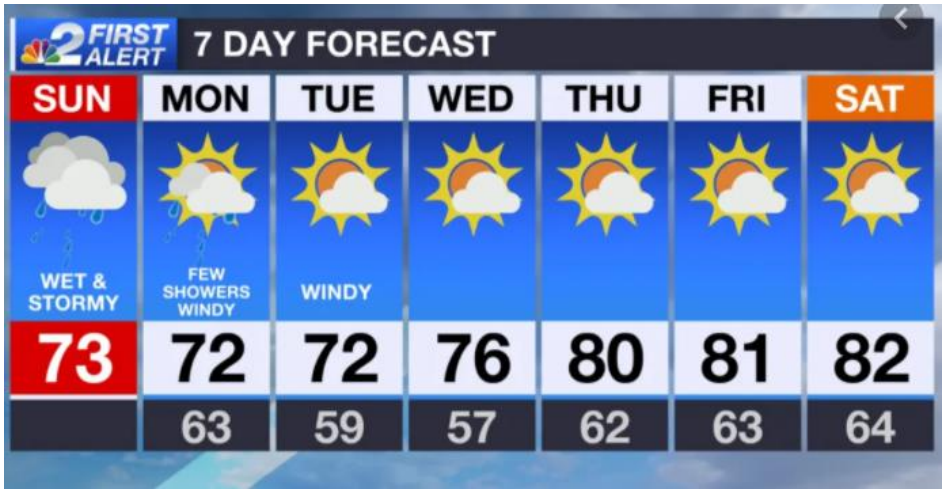
SECI2143: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 1

Introduction to Statistics

Why Study Statistics?





Further reading:
https://medium.com/@john_marsh7/10-awesome-reasons-why-statistics-are-important-96b87e283640

- Even though you may not have realized it, you probably have made some statistical statements in your everyday conversation or thinking.

- Statements like "I sleep for about eight hours per night on average" and "You are more likely to pass the exam if you start preparing earlier" are actually statistical in nature.

- We encounter data and conclusions based on data every day.
- Statistics is the scientific discipline that provides methods to help us make sense of data.
- Statistical methods are used in business, medicine, agriculture, social sciences, natural sciences, and applied sciences, such as engineering.
- The field of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

- Statistics is the scientific application of mathematical principles to the collection, analysis, and presentation of numerical data.
- Statistics is a discipline which is concerned with:
  - designing experiments and other data collection,
  - summarizing information to aid understanding,
  - drawing conclusions from data, and
  - estimating the present or predicting the future.
- There are 2 main branches of statistics:
  - Descriptive
  - Inferential

Descriptive Statistics

- **Descriptive statistics** are used to describe the basic features of the data gathered from an experimental study in various ways.
- The techniques are commonly classified as:
  - Graphical description in which we use graphs to summarize data.
  - Tabular description in which we use tables to summarize data.
  - Parametric description in which we estimate the values of certain parameters which we assume to complete the description of the set of data.

- Graphical description:

Example:

- Line chart
- Bar chart
- Pie chart

- Tabular description:

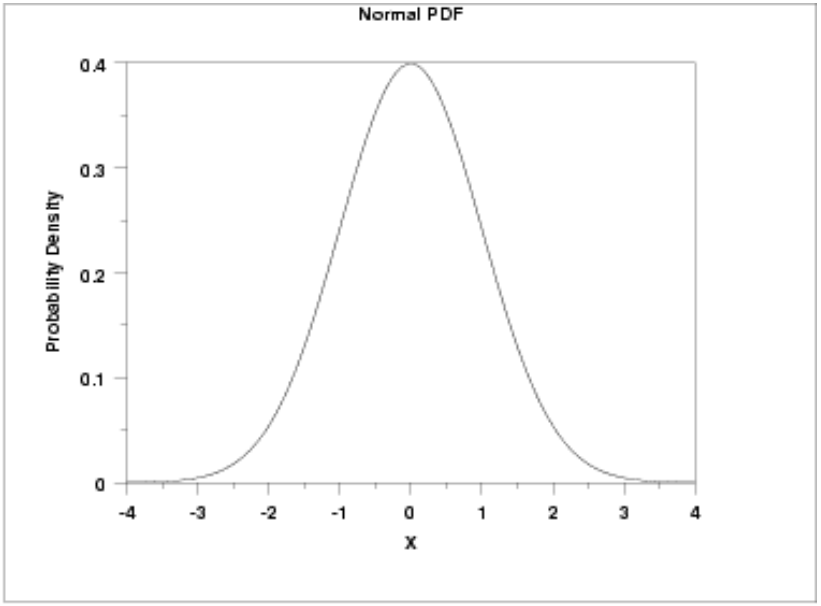
Example:

- Frequency Table

Score	Frequency
0	2
1	5
2	8
3	6
4	4
5	3

❑ Parametric description:

Mean	μ
Median	μ
Mode	μ
Range	Infinity in both directions.
Standard Deviation	σ
Skewness	0
Kurtosis	3

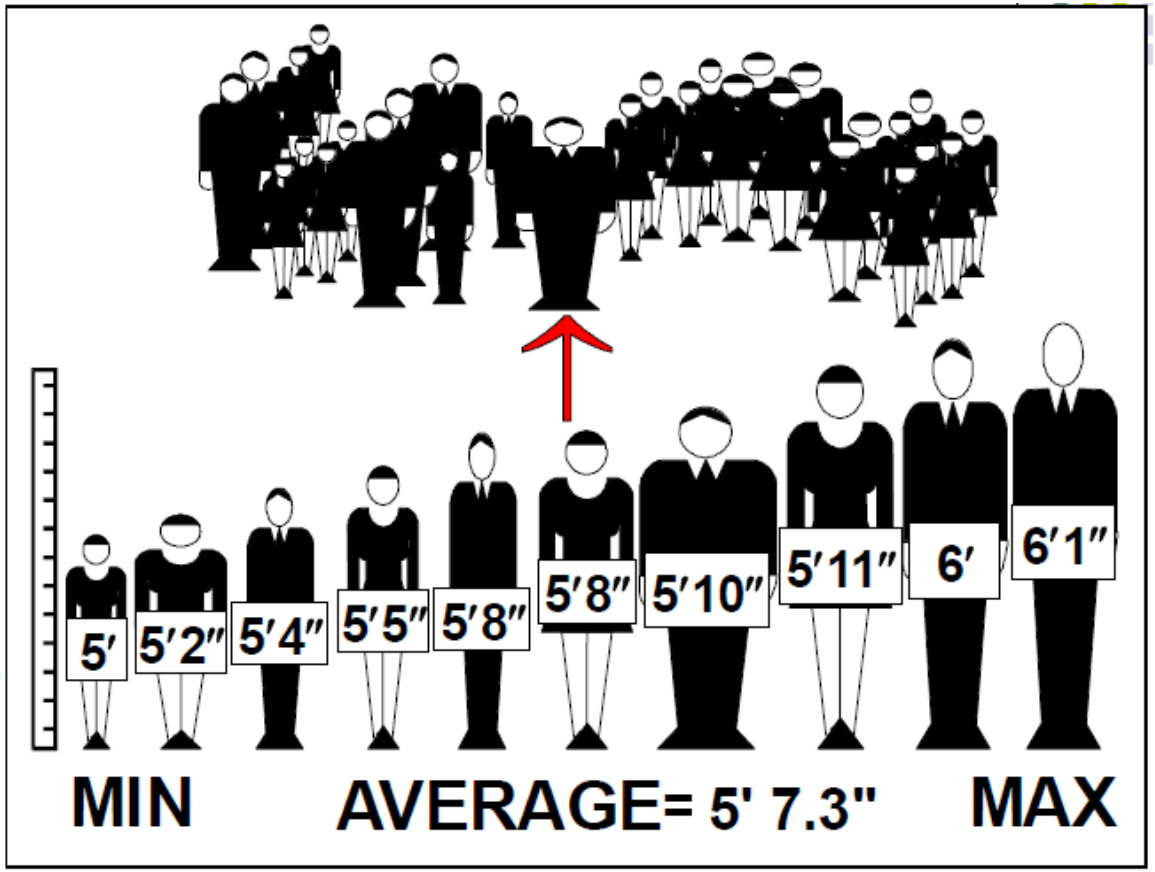


Inferential Statistics

- **Inferential statistics** are used to draw inferences about a population from a sample.
- It includes:
  - point estimation
  - interval estimation
  - hypothesis testing (or significant testing)
  - prediction

Inferential

Descriptive

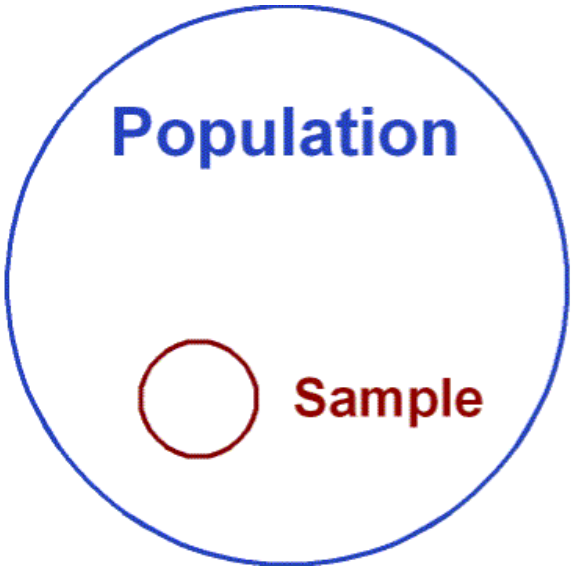


Population

Sample

Population & Sample

- The entire collection of individuals or object about which information is desired is called the **population** of interest.
- A **sample** is a subset of the population, selected for study in some prescribed manner.



Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

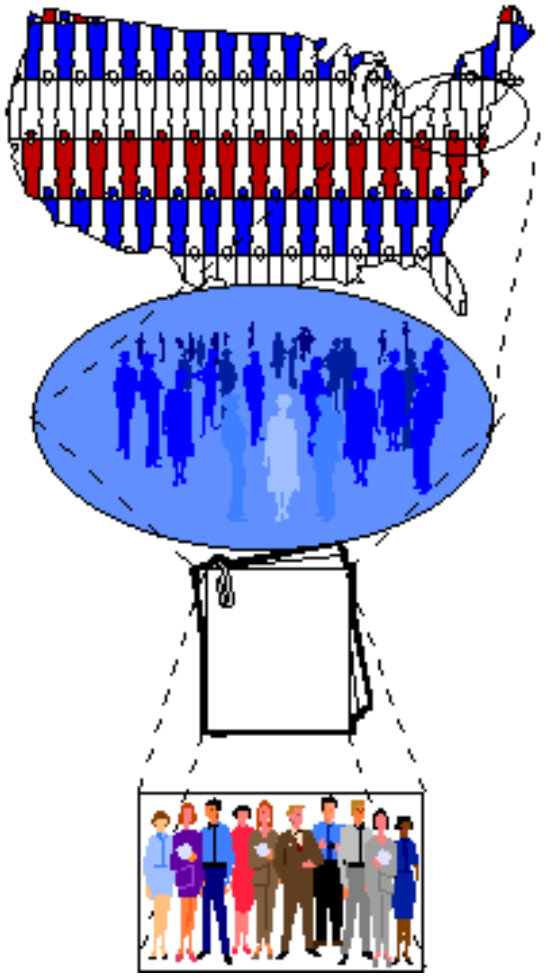
The Study Population

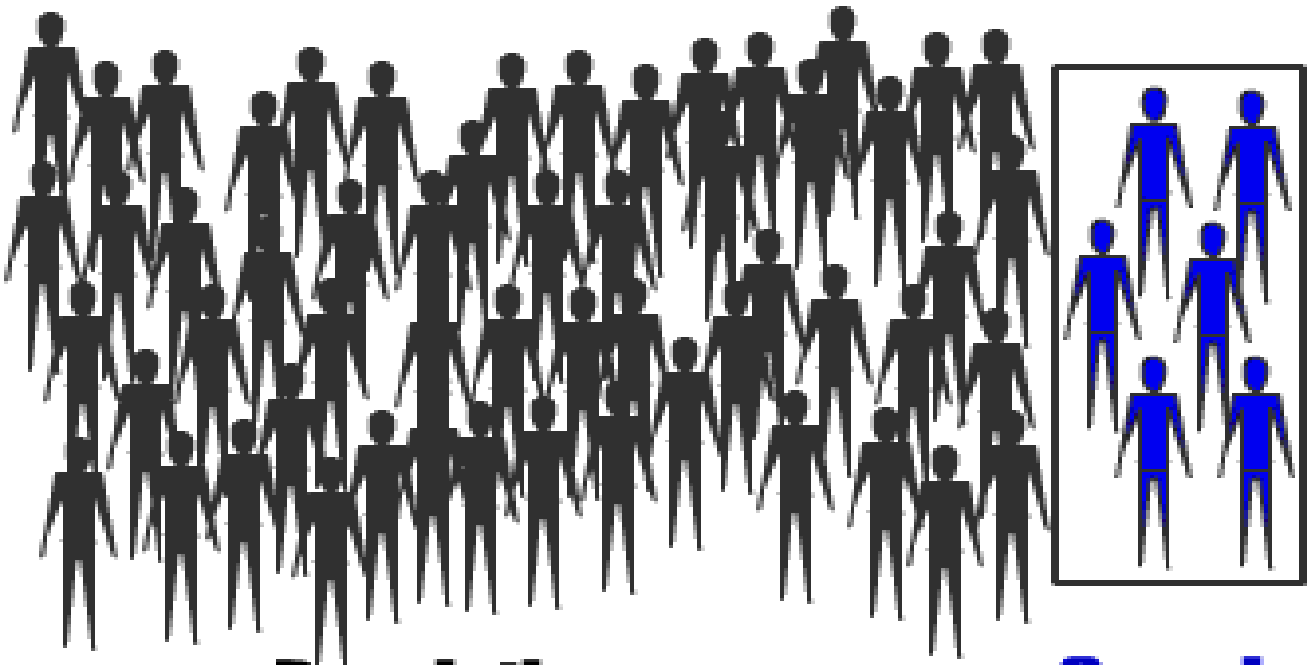
How can you get access to them?

The Sampling Frame

Who is in your study?

The Sample





Population

quantity (count) = N

mean = μ

variance = σ^2

standard deviation = σ

Sample

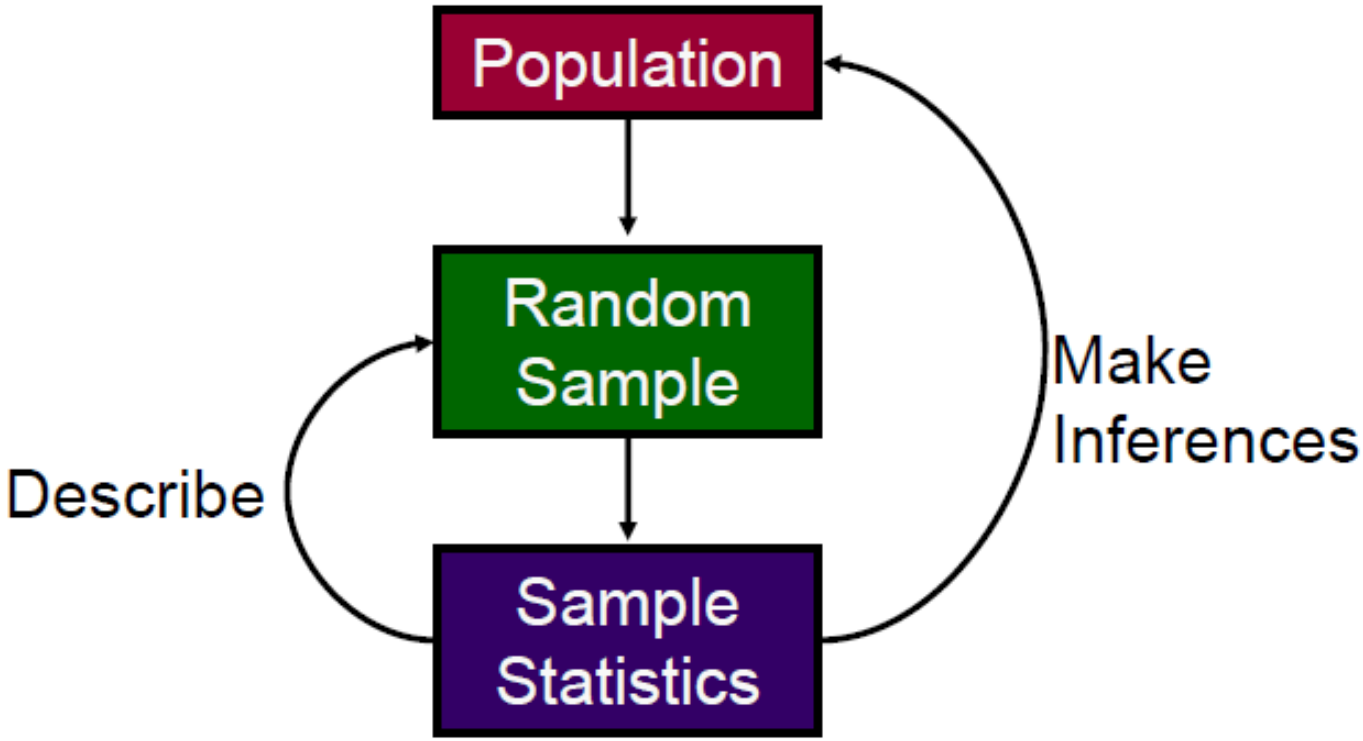
quantity (count) = n

mean = $\bar{x}$

variance = s^2

standard deviation = s

Data Analysis Process



- Statistics involves the **collection** and **analysis** of data
- Both task are critical.
- Raw data without analysis are of little value, and even a sophisticated analysis cannot extract meaningful information from data that were not collected in a sensible way.
- The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to informed conclusions based on the resulting data.

Steps in Data Analysis Process

6 steps:

i. Understanding the nature of the problem
ii. Deciding what to measure and how to measure it
iii. Data collection
iv. Data summarization and preliminary analysis
v. Formal data analysis
vi. Interpretation of results

i) Understanding the nature of the problem.

- An understanding of the research problem.
- Know the goal of the research and what questions we hope to answer.
- Have a clear direction before gathering data to lessen the chance of being unable to answer the questions of interest using the data collected.

ii) Deciding what to measure and how to measure it.

- In some cases, the choice is obvious, e.g. in a study of the relationship between the weight of a football player and position played, you would need to collect data on player weight and position.

- But in other cases, the choice of information is not as straightforward.

  Example: In a study of the relationship between preferred learning style and intelligence, how would you define learning style and measure it and what measure of intelligence would you use?

- It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.

iii) Data collection

- Decide whether an existing data source is adequate or whether new data must be collected.
- If a decision is made to use existing data (secondary data), it is important to understand how the data were collected and for what purpose.

- If new data are to be collected (primary data), a careful plan must be developed.
- The type of analysis that is appropriate and subsequent conclusions that can be drawn depend on how the data are collected.

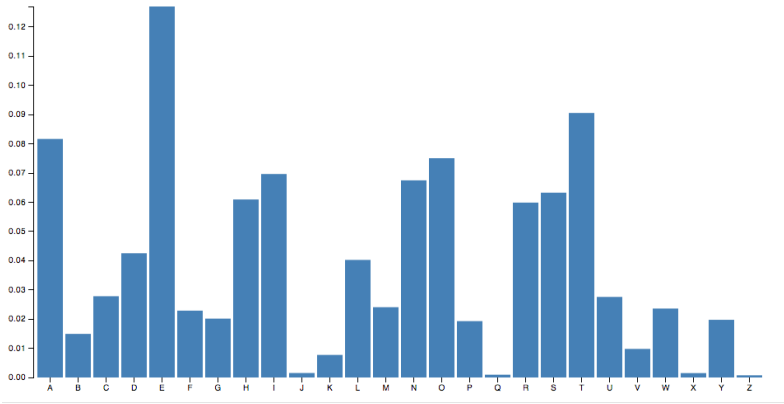
iv) Data summarization and preliminary analysis

- Summarizing the data graphically and numerically.
- This initial analysis provides insight into important characteristics of the data and can provide guidance in selecting appropriate methods for further analysis.

Microsoft Excel - Book3

	A	B	C	D	E
1		October	November	December	
2	Refrigerator	84770	79941	76741	
3	TV	195179	236378	233712	
4	Air-condition	124244	115395	99554	
5	Cellphone	198673	206041	222230	
6	Computer	230263	236253	235127	
7	Total	833129	874008	867364	
8					

Sheet1



v) Formal data analysis

- Select and apply the appropriate inferential statistical methods.

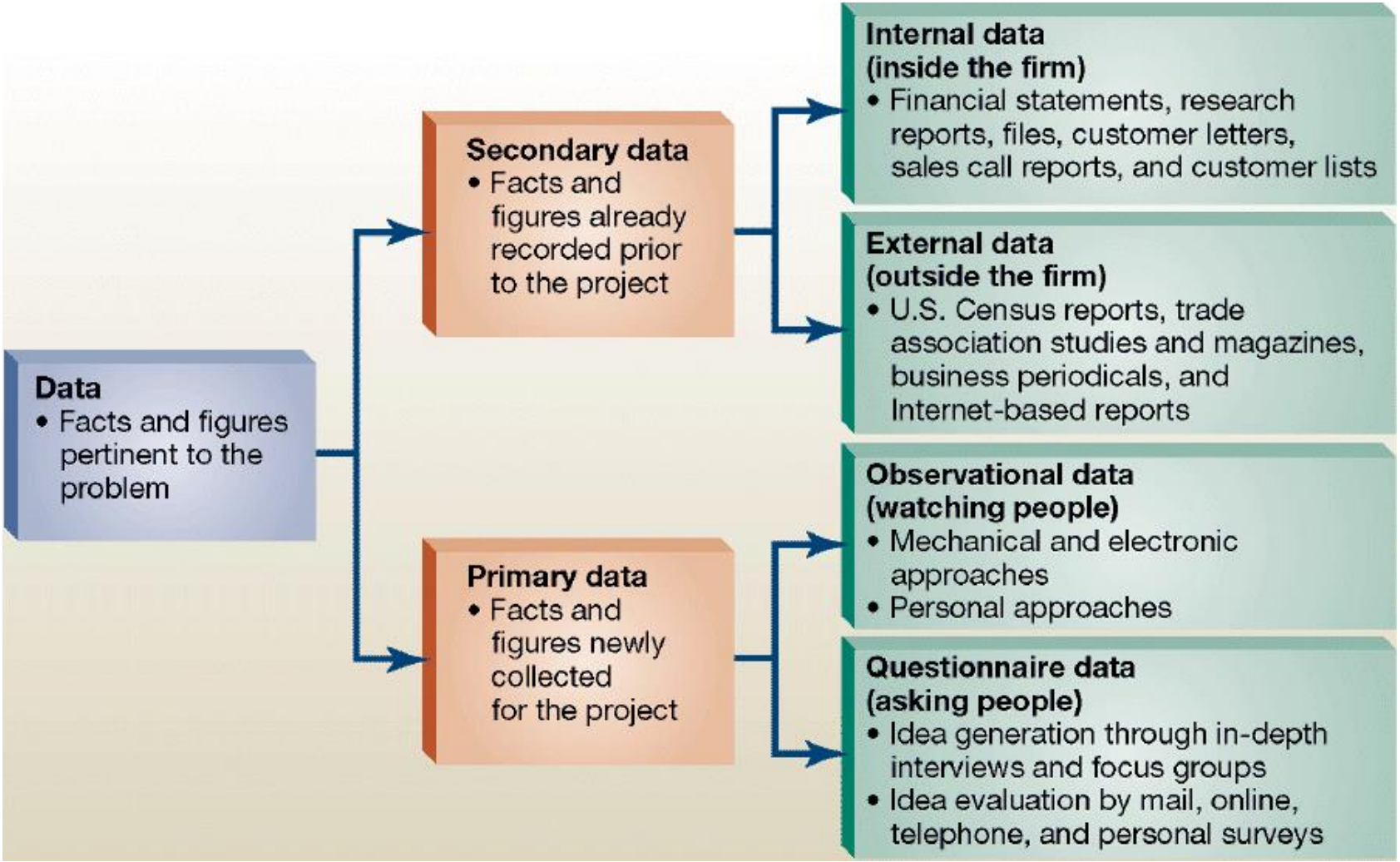
Example


Sales by region
Region
West
East
Central
South
$800,000
$600,000
$400,000
$200,000
$0
Sales
$725,458
$678,781
$501,240
$391,722
West
East
Central
South
Avg. profit distribution by region
Region
$200
$150
$100
$50
$0
($50)
Avg. Profit
Central
East
South
West
Sales trend
$120,000
$100,000
$80,000
$60,000
$40,000
$20,000
$0
Sales
Month: November 2016
Sales: $118,448
2013
2014
2015
2016
2017
Month of Order Date
Top sub categories with maximum sales
Sub-Category
$300,000
$200,000
$100,000
$0
Sales
14.37%
28.66%
38.41%
47.42%
56.27%
64.51%
71.80%
78.31%
83.31%
87.99%
91.98%
95.40%
97.43%
98.61%
99.32%
99.87%
100.00%
100%
80%
60%
40%
20%
0%
% of Total Running Sum of Sales
Phones
Chairs
Storage
Tables
Binders
Machines
Accessories
Copiers
Bookcases
Appliances
Furnishings
Paper
Supplies
Art
Envelopes
Labels
Fasteners

vi) Interpretation of results

- What conclusions can be drawn from the analysis?
- How do the result of the analysis inform us about the stated research problem or question?
- How can our results guide future research?

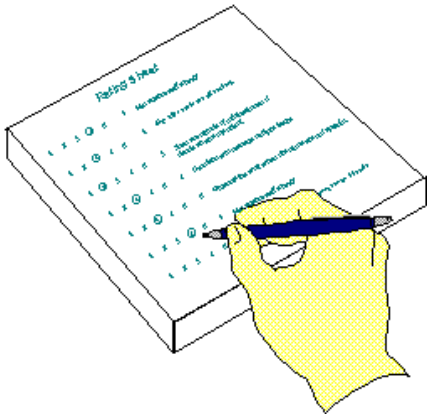
Data Sources:
Primary & Secondary Data



Primary Data

- EXPERIMENT

- SURVEY (Questionnaire/Interview)



Secondary Data

- EXISTING DATABASE

- RECORD REVIEW

Data Types:
Qualitative & Quantitative Data

Qualitative Data	Quantitative Data
▪ Deals with descriptions. ▪ Data can be observed but not measured. ▪ Example: Colors, textures, smells, tastes, appearance, beauty, etc. ▪ Qualitative → Quality	▪ Deals with numbers. ▪ Data which can be measured. ▪ Example: Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc. ▪ Quantitative → Quantity

Example

- ***Oil Painting***

Qualitative data:	Quantitative data:
•blue/green color, gold frame •smells old and musty •texture shows brush strokes of oil paint •peaceful scene of the country •masterful brush strokes	•picture is 10" by 14" •with frame 14" by 18" •weighs 8.5 pounds •surface area of painting is 140 sq. in. •cost $300

Example

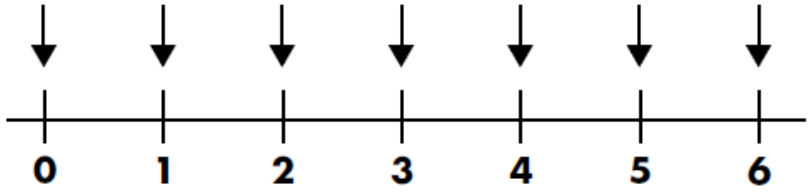
- ***Latte***

Qualitative data:
robust aroma
frothy appearance
strong taste
burgundy cup

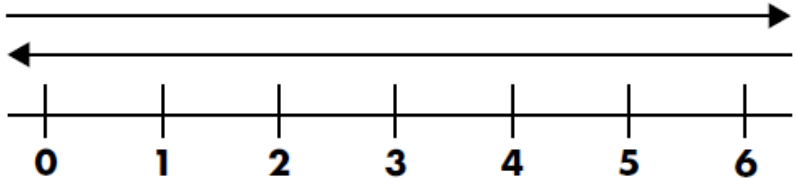
Quantitative data:
12 ounces of latte
serving temperature 150º F.
serving cup 7 inches in height
cost $4.95

Quantitative Data:
Discrete or Continuous

Discrete data can only take on certain individual values.



Continuous data can take on any value in a certain range.



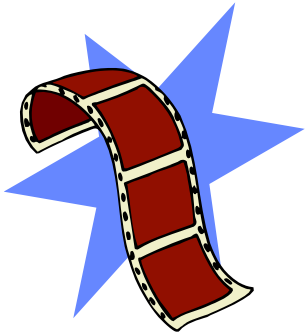
Example

Example 1

Number of pages in a book is a discrete variable.

Example 2

Length of a film is a continuous variable.



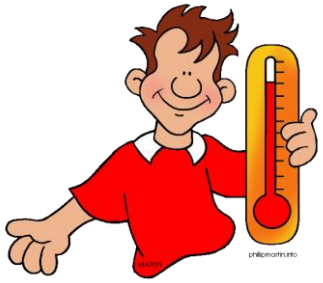
Example 3

Shoe size is a Discrete variable. E.g. $5, 5\frac{1}{2}, 6, 6\frac{1}{2}$ etc. Not in between.

Example

Example 4

Temperature is a continuous variable.

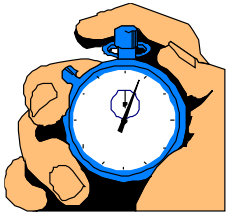


Example 5

Number of people in a race is a discrete variable.

Example 6

Time taken to run a race is a continuous variable.



Exercise #1

State the type of data for the following cases as either Discrete or Continuous.

Case	Discrete or Continuous?
Number of matches in a box	
Speed of a car	
Population of a town	
Length of crocodile	
Temperature of oven	
T-Shirt size	

Levels of Data Measurement

- There are four levels of data measurement.
- Ranked from top to bottom in order of complexity and information content.
  - Ratio scale
  - Interval scale
  - Ordinal scale
  - Nominal/categorical scale

Each level of measurement is characterized by its properties:

- Nominal/categorical measurement has just one property: CLASSIFICATION.
- Ordinal measurement has two properties: CLASSIFICATION and ORDER.
- Interval measurement has three properties: CLASSIFICATION, ORDER and EQUAL INTERVALS.
- Ratio data has four properties: CLASSIFICATION, ORDER, EQUAL INTERVALS and TRUE ZERO.

Nominal Scales

Properties: Classification.

Observations reflect: Differences in kind.

Examples: gender, ethnic background.

- Nominal scales are used to labeling variables without any quantitative value.
- Numbers assigned to categories (as identification codes) have no numeric value (we cannot add, subtract, divide or multiply nominal data) and any ordering of categories is arbitrary.

- Since the single property of nominal data is classification, thus it doesn't tells us about differences in degree or amount.
- This is the most primitive form of measurement. The presence vs. absence of something is a form of nominal measurement (e.g., "do you smoke?" YES, NO).
- Collection of nominal data is easy.

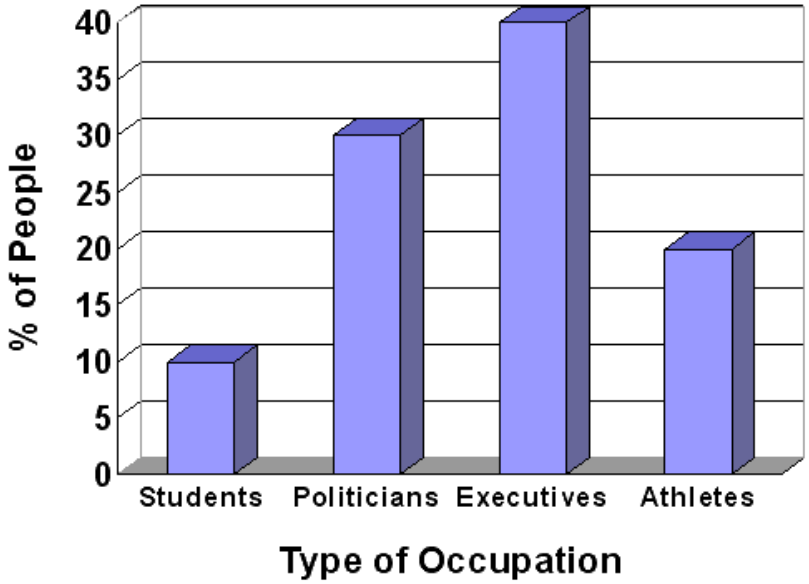
Example

What is your gender?
- M – Male
- F – Female

What is your hair color?
- 1 – Brown
- 2 – Black
- 3 – Blonde
- 4 – Gray
- 5 – Other

Where do you live?
- A – North of the equator
- B – South of the equator
- C – Neither: In the international space station

Examples of Nominal Scales



Ordinal Scales

Properties: Classification, Order.

Observations reflect: differences in degree.

Examples: Likert scale categories, rankings, academic letter grade, stages in development.

- The distinctive property of ordinal measurement is order.
- On a typical Likert Scale “strongly agree” represents more agreement than “agree”. However, we do not know how much more.

Example

Ordinal Data

Point	**Airports** international national regional	**Oil well production** high medium low	**Populated places** large medium small
Line	**Roads** expressway major local	**Drainage** river stream creek	**Boundaries** international provincial county
Area	**Soil quality** good fair poor	**Cost of living** high medium low	**Industrial regions** major minor

An ordinal data example

How often do you eat cheese for breakfast?

		Code
always		6
usually		5
often		4
sometimes		3
occasionally		2
rarely		1
never		0

"always" is clearly more frequent than "sometimes" but not necessarily twice as frequent, even though 6 = twice 3

Example

The example of questionnaire that used Likert Scale:

	Strongly Agree	Agree	Neither	Disagree	Strongly Disagree
If the price of raw materials fell firms would reduce the price of their food products.	1	2	3	4	5
Without government regulation the firms would exploit the consumer.	1	2	3	4	5
Most food companies are so concerned about making profits they do not care about quality.	1	2	3	4	5
The food industry spends a great deal of money making sure that its manufacturing is hygienic.	1	2	3	4	5
Food companies should charge the same price for their products throughout the country	1	2	3	4	5

Interval Scales

Properties: Classification, Order, Equal Intervals.

Observations reflect: measurable differences in amount.
Examples: IQ scores, degrees of temperature.

- Essentially, interval data are ordinal, but they have an extra property, that is the ability to meaningfully add and subtract measurements.
- In interval-scaled data, the gaps between the numbers are comparable, unlike with ordinal data.
- Any interval has the same meaning regardless of its location on the scale. Example: " *x* is five inches longer than *y*" has meaning regardless of the values of *x* and *y*.

- However, ratios are meaningless on an interval scale because an interval scale has no true zero.
- Example:
  - Temperature scales: Zero degrees Fahrenheit does not mean the total absence of temperature.
  - Decibel scales: Zero decibels does not mean there is no sound.

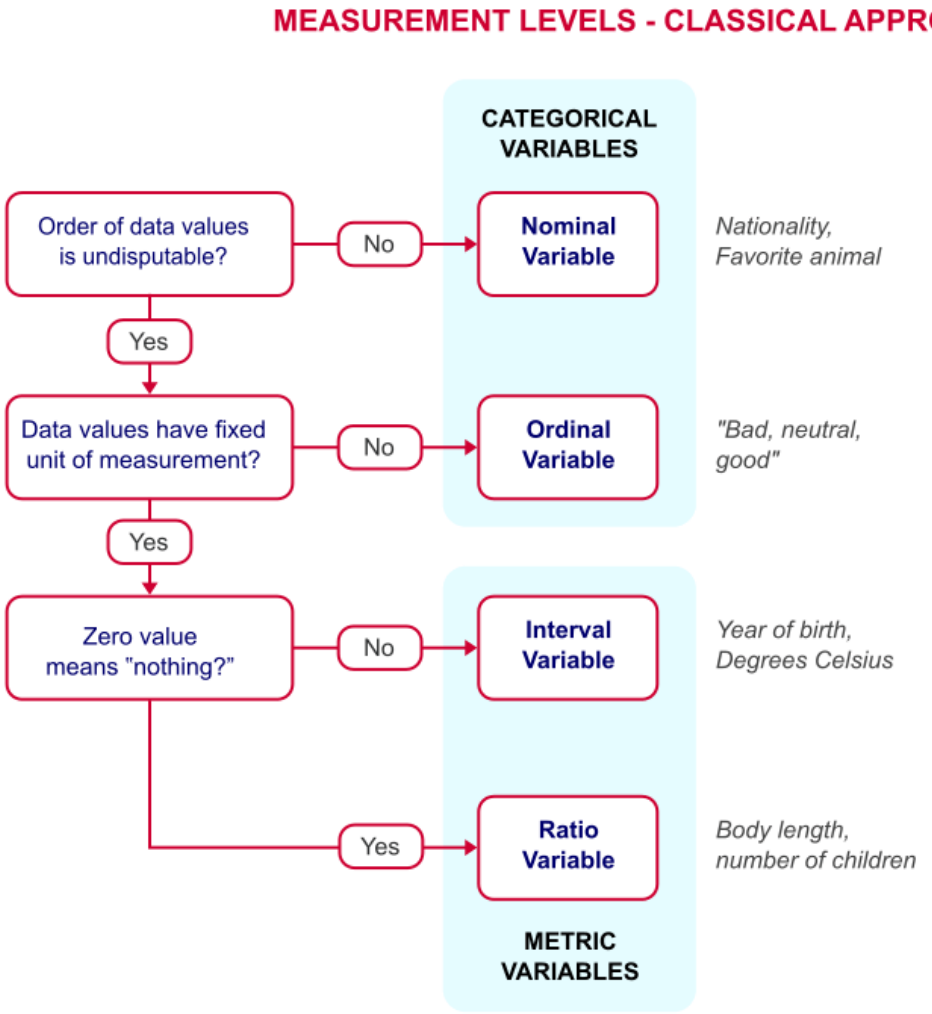
Ratio Scales

Properties: Classification, Order, Equal Intervals, True Zero
Observations reflect: measurable differences in total amount
Examples: weight, income, family size.

- Ratio data are the highest form of data measurement and the form we are most familiar with.
- The ratios are interpretable and it has a natural zero.
- Ratio data look a lot like interval data. However, the zero point has a special meaning in ratio-scaled data. It indicates the absence of whatever property is being measured.

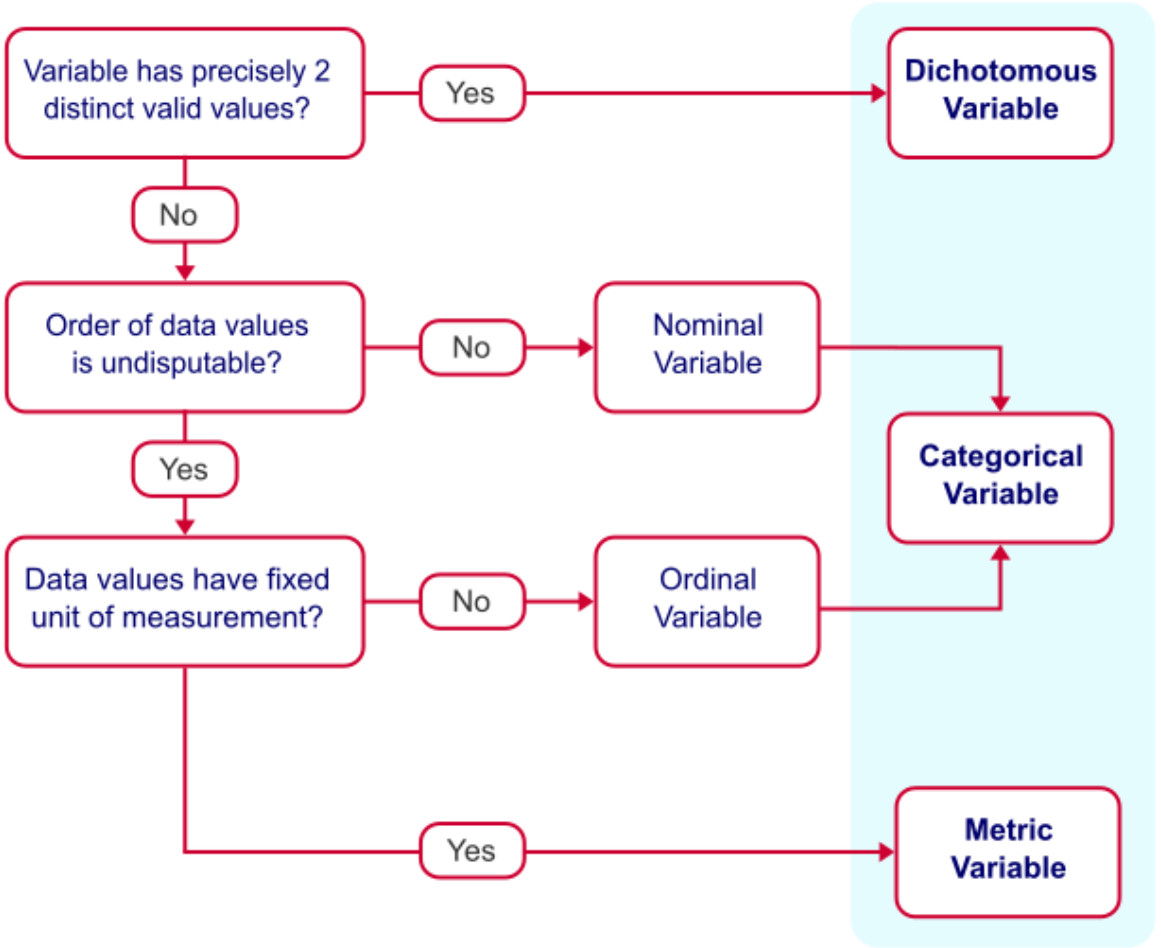
- Ratio data always have the flavor of counting. Example:
  - ✓When you measure the amount of money that you have, you are counting up coins and bills.
  - ✓When you are measuring your height, you are counting the number of inches off the ground to the top of your head.
- Both ratio and interval data make use of a wide range of statistical analysis tools.

Summary



Source: https://www.spss-tutorials.com/measurement-levels/

MEASUREMENT LEVELS - MODERN APPROACH



Source: https://www.spss-tutorials.com/measurement-levels/

Statistic	Nominal	Ordinal	Interval	Ratio
Mode	√	√	√	If meaningful
Median	X	√	√	√
Range, Min. Max	X	√	√	√
Mean	X	X	If metric	√
SD	X	X	If metric	√

Graph	Nominal	Ordinal	Interval	Ratio
Bar / Pie	√	√	If discrete	X
Stem & Leaf	X	√	√	√
Boxplot	X	√	√	√
Histogram	X	X	If metric	√

"You can have data without information, but you cannot have information without data." — Daniel K. Moran

Exercise #2

Determine which of the four levels of measurement (nominal, ordinal, interval and ratio) is most appropriate.

a) Heights of women basketball players in a tournament.

b) Ratings of superior, above average, average, below average or poor for blind dates.

c) Today temperatures (in degrees Celsius) in Kuala Lumpur.

d) A movie critic's classification of "drama, comedy, adventure".

e) The number of bugs made when a programmer develop a coding for a project.