UTM SCHOOL OF COMPUTING
Faculty of Engineering

SECI 2143 / SCSI 2143
PROBABILITY & STATISTICAL DATA ANALYSIS

innovative • entrepreneurial • global

UTM

CHAPTER 1
1.1 INTRODUCTION TO STATISTIC

Introduction to Statistic

Introduction to Statistic

- Example:
- Statement 1
  - "I sleep for about eight hours per night on average".
- Statement 2:
  - "You are more likely to pass the exam if you start preparing earlier" are actually statistical in nature.

Introduction to Statistic

- Statistics:
  - Scientific discipline that provides methods to help us make sense of data.
  - Scientific application of mathematical principles to the collection, analysis, and presentation of numerical data.

- Statistical methods are used in business, medicine, agriculture, social sciences, natural sciences, and applied sciences, such as engineering.

Introduction to Statistic

- Statistics teaches us how to make **intelligent judgments** and **informed decisions** in the presence of uncertainty and variation.

- Statistics is a discipline which is concerned with:
  1. **Designing** experiments and other data collection,
  2. **Summarizing** information to aid understanding,
  3. **Drawing** conclusions from data, and
  4. **Estimating** the present or predicting the future.

Introduction to Statistic

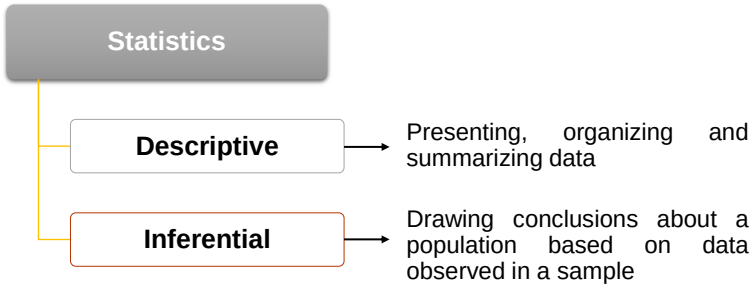
- There are 2 main branches of statistics:
  1. **Descriptive**
  2. **Inferential**

CHAPTER 1

1.1.1 Descriptive and Inferential Statistics

innovative • entrepreneurial • global | www.utm.my

Descriptive and Inferential Statistics



Descriptive Statistics

- **Descriptive Statistics:**
  - Describe the basic features of the data gathered from an experimental study in various ways.
  - Construction of graphs, charts, and tables and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

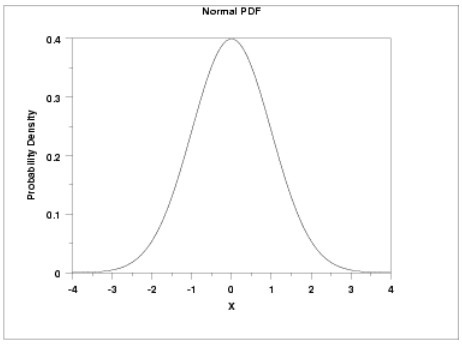
Descriptive Statistics

- The techniques are commonly classified as:
  1. Graphical description - use graphs to summarize data.
  2. Tabular description - use tables to summarize data.
  3. Parametric description - estimate the values of certain parameters which we assume to complete the description of the set of data.

Descriptive Statistics

Parametric description

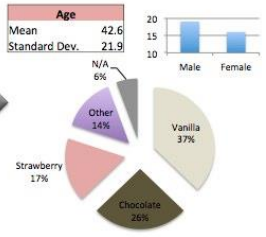
Mean	μ
Median	μ
Mode	μ
Range	Infinity in both directions.
Standard Deviation	σ
Skewness	0
Kurtosis	3



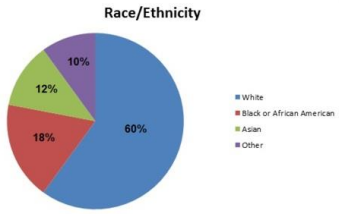
Descriptive Statistics

	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15	14	55	m	Vanilla
16	15	42	f	Chocolate
17	16	18	f	Vanilla

Raw Data



Descriptive Statistics



Inferential Statistics

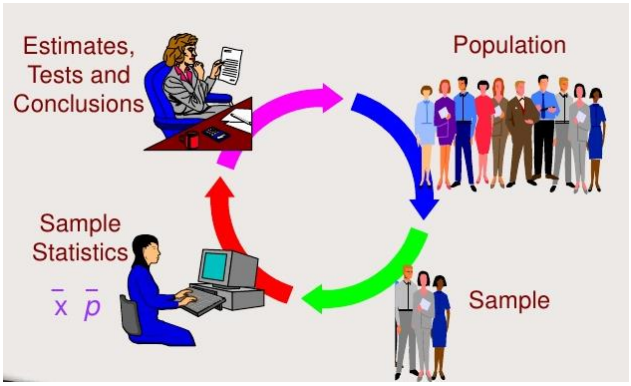
- **Inferential Statistics:**
  - Used to make inferences (draw conclusions) about a population from a sample.
  - Consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

Inferential Statistics

- It includes:
  1. Point estimation
  2. Interval estimation
  3. Hypothesis testing (or significance testing)
  4. Prediction

Inferential Statistics

- Process:

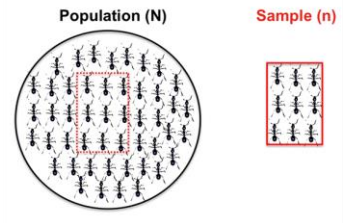


CHAPTER 1

1.1.2 Population and Sample

Population and Samples

- **Population & Samples:**
  1. **Population:** The entire collection of individuals or object about which information is desired.
  2. **Sample:** That part of the population, selected for study in some prescribed manner.



innovative • entrepreneurial • global | www.utm.my

Inferential Statistics

Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

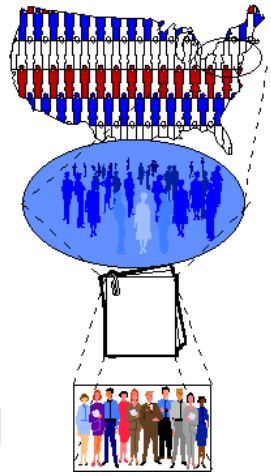
The Study Population

How can you get access to them?

The Sampling Frame

Who is in your study?

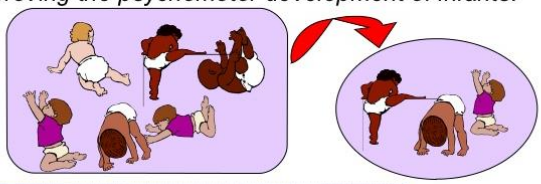
The Sample



Inferential Statistics

Example of population & samples

A new milk formulation designed to improve the psychomotor development of infants was tested on randomly selected infants. Based on the results, it was concluded that the new milk formulation is effective in improving the psychomotor development of infants.



CHAPTER 1

1.2 DATA

Data

- **Data:**
  - Statistics involves the **collection** and **analysis** of data.
  - 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.

innovative • entrepreneurial • global | www.utm.my

CHAPTER 1

1.2.1 Data Analysis Process

Data Analysis Process

- 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:**
  1. Understanding the nature of the problem
  2. Deciding what to measure and how to measure it
  3. Data collection
  4. Data summarization and preliminary analysis
  5. Formal data analysis
  6. Interpretation of results

innovative • entrepreneurial • global | www.utm.my

Data Analysis Process

1. **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.

Data Analysis Process

2. **Deciding what to measure and how to measure it (Choice)**

   Example 1:

   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.

   Example 2:

   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?

Data Analysis Process

3. 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.

Data Analysis Process

4. 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.

Data Analysis Process

5. **Formal data analysis**
   - Select and apply the appropriate inferential statistical methods.

6. **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?

CHAPTER 1

1.2.2 Data Sources (Primary and Secondary Data)

innovative • entrepreneurial • global | www.utm.my

Primary and Secondary Data

- **Secondary data:**
  - Data which has been collected by individuals or agencies for purposes other than those of particular research study.

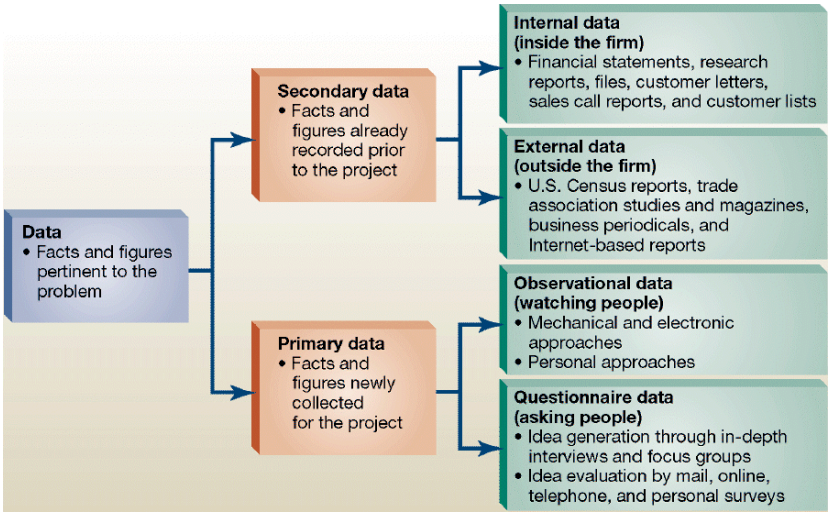
  - Example:

  If a government department has conducted a survey of, say, family food expenditures, then a food manufacturer might use this data in the organization's evaluations of the total potential market for a new product.

Primary and Secondary Data

- **Primary data:**
  - By contrast, are collected by the investigator conducting the research.

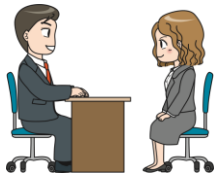
Primary and Secondary Data



Data Sources

Primary data: Experiment

Primary data: Survey (Questionnaire, Interview)



Data Sources

Secondary data: existing databases, record review





CHAPTER 1

1.2.3 Types of Data (Qualitative, Quantitative, Discrete and Continuous Data)

innovative • entrepreneurial • global | www.utm.my

Qualitative & Quantitative Data

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

Qualitative & Quantitative Data

Example 1:

- *Oil Painting*

Qualitative 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

Quantitative data:
- picture is 10" by 14"
- with frame 14" by 18"
- weighs 8.5 pounds
- surface area of painting is 140 sq. in.
- cost RM300

Qualitative & Quantitative Data

Example 2:

- ***Latte***

Qualitative data:	Quantitative data:
• Robust aroma • Frothy appearance • Strong taste • Burgundy cup	• 12 ounces of latte • Serving temperature 150º F. • Serving cup 7 inches in height • Cost RM4.95

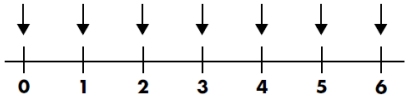
Qualitative & Quantitative Data

Example 3:

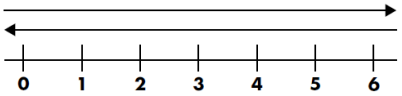
- ***A girl***

Discrete & Continuous Data

Discrete data can only take on certain individual values.



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



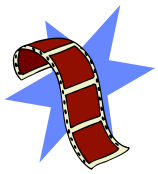
Discrete & Continuous Data

Example 1

Number of pages in a book is a **discrete variable.**

Example 2

Length of a film is a **continuous variable.**

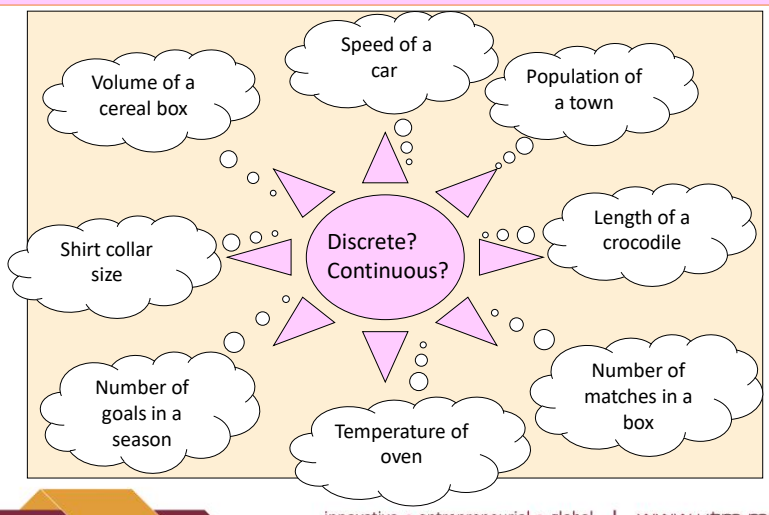


Example 3

Shoe size is a **Discrete variable.** E.g. 5, 5½, 6, 6½ etc. Not in between.

Discrete & Continuous Data

EXERCISE: Group the following as either discrete or continuous data.



CHAPTER 1

1.2.4 Data Scale and Measurement (Nominal, Ordinal, Interval, Ratio)

Levels of Measurement

- **4 levels of measurements,** ranked from top to bottom in order of complexity.

- Each level of measurement is characterized by its properties.
  1. Nominal scale (CLASSIFICATION).
  2. Ordinal scale (CLASSIFICATION and ORDER).
  3. Interval scale (CLASSIFICATION, ORDER and EQUAL INTERVALS).
  4. Ratio scale (CLASSIFICATION, ORDER, EQUAL INTERVALS and TRUE ZERO)

Levels of Measurement – Nominal Scales

- Properties: classification
- Observations reflect: differences in kind
- Examples: gender, ethnic background, major in college

- Nominal measurement is simply concerned with sorting observations into categories.

- Because the single property of nominal data is classification it tells us nothing about differences in degree or amount.

Levels of Measurement – Nominal Scales

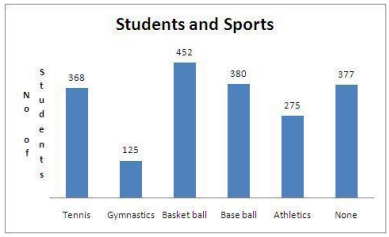
- Numbers assigned to categories (as identification codes) have no numeric value (we **cannot add, subtract, divide or multiply nominal data**) and any ordering of categories is arbitrary.
- This is the most primitive form of measurement. The presence vs. absence of something is a form of nominal measurement ("do you smoke?" YES, NO).
- Although it is considered a form of measurement the collection of nominal data is more easily thought of as a sorting method.

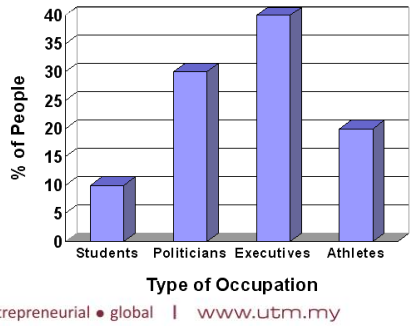
Levels of Measurement – Nominal Scales

Examples:

Nominal Data

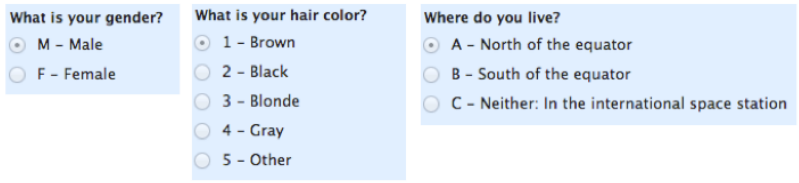
Point	airport	town	mine	capital
Line	river	road	boundary	pipeline
Area	orchard	desert	forest	water





Levels of Measurement – Nominal Scales

Examples:



Sometimes numbers are used to designate category membership

Example:
Country of Origin
1 = United States
2 = Mexico
3 = Canada
4 = Other

However, in this case, it is important to keep in mind that the numbers do not have intrinsic meaning

Levels of Measurement – 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.

Levels of Measurement – Ordinal Scales

Example: The 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

Levels of Measurement – Ordinal Scales

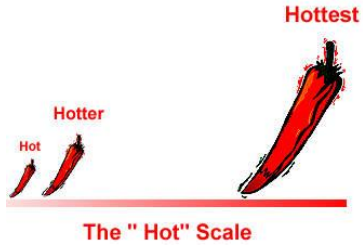
Example:

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



Levels of Measurement – 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 - the ability to meaningfully **add** and **subtract** measurements.

Levels of Measurement – Interval Scales

- 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. "X is five inches longer than Y" has meaning regardless of the values of X and Y.

innovative • entrepreneurial • global | www.utm.my

Levels of Measurement – Interval Scales

- However, ratios are meaningless on an interval scale because an interval scale has no true zero.

- Examples:
  1. Temperature scales - are decibel scales.
  2. Zero degrees Fahrenheit does not mean the total absence of temperature.
  3. Zero decibels does not mean there is no sound.

Levels of Measurement – Interval Scales

- Furthermore, if it is 80 degrees outside today and it was only 40 degrees outside yesterday we cannot say that today is twice as hot as yesterday.

- Similarly a sound level of 80 dB is not twice as loud as a sound level of 40 dB.

- In short, if the data can be ordered and the arithmetic difference is meaningful, then the data are at least interval data.

Levels of Measurement – Ratio Scales

- Properties: classification, order, equal intervals, true zero.
- Observations reflect: measurable differences in total amount.
- Examples: weight, income, family size, number of cows in a field.

- Ratio data are the highest form of data measurement and the form we are most familiar with.

Levels of Measurement – Ratio Scales

- For ratio data both differences and ratios are interpretable.

- Ratio data have a natural zero.

Levels of Measurement – Ratio Scales

- Examples:
  - Number of computers you own, weight, height, a bank balance, number of people watching a movie, goals scored by brazil in the world cup, etc.
- 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.

Levels of Measurement – Ratio Scales

- Ratio data always have the flavor of counting: 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.

Levels of Measurement

	NOMINAL	ORDINAL	INTERVAL	RATIO
Indicates Difference	X	X	X	X
Indicates Difference & Direction		X	X	X
Indicates Amount of Difference			X	X
Absolute Zero				X

OK to compute....	Nominal	Ordinal	Interval	Ratio
frequency distribution.	Yes	Yes	Yes	Yes
median and percentiles.	No	Yes	Yes	Yes
add or subtract.	No	No	Yes	Yes
mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
ratio, or coefficient of variation.	No	No	No	Yes

Level of Measurement	Properties	Examples	Descriptive statistics	Graphs
Nominal / Categorical	Discrete Arbitrary (no order)	Dichotomous • Yes / No • Gender Types / Categories • colour • shape	Frequencies Percentage Mode	Bar Pie
Ordinal / Rank	Ordered categories Ranks	Ranking of favourites Academic grades	Frequencies Mode Median Percentiles	Bar Pie Stem & leaf
Interval	Equal distances between values Discrete (e.g., Likert scale) Metric (e.g., deg. F) Interval scales >5 can usually be treated as ratio	Discrete - Thoughts, behaviours, feelings, etc. on a Likert scale Metric - Deg. C or F	Frequencies (if discrete) Mode (if discrete) Median Mean SD Skewness Kurtosis	Bar (if discrete) Pie (if discrete) Stem & Leaf Boxplot Histogram (if metric)
Ratio	Continuous / Metric / Meaningful 0 allows ratio statements (e.g., A is twice as large as B)	Age Weight VO_2 max Deg. Kelvin	Mean SD Skewness Kurtosis	Histogram Boxplot Stem&Leaf (may need to round leafs)

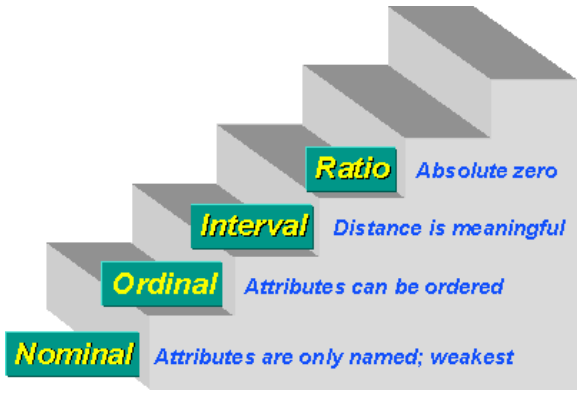
innovative • entrepreneurial • global | www.utm.my

Levels of Measurement

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	√

Levels of Measurement



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

Levels of Measurement

Scales to classify different measurements

	Nominal	Ordinal	Interval	Ratio
Sex	x			
Hair colour	x			
Pulse				x
Temp. °C			x	
Team number	x			
Shoe size		x		

Exercises

Identify the following as **Nominal, Ordinal, Interval or Ratio Level data**:

1. Flavors of frozen yogurt.
2. Amount of money in saving accounts.
3. Students classified by their reading ability: Above average, Below average, Normal
4. Letter grades on an English essay
5. Religions
6. Commuting times to work
7. Ages (in years) of art students
8. Ice cream flavour preference
9. Years of important historical events
10. Instructors classified as: Easy, Difficult or Impossible.