



# STATISTICS FUNDAMENTAL

DR. NURULHUDA FIRDAUS  
(HUDA)

# CONTENTS

- Know & Understand the Key Definition
- Descriptive Statistics
- Inferential Statistics
- Probability
- Linear Algebra



# HOW THEY ARE CONNECTED TO EACH OTHER?



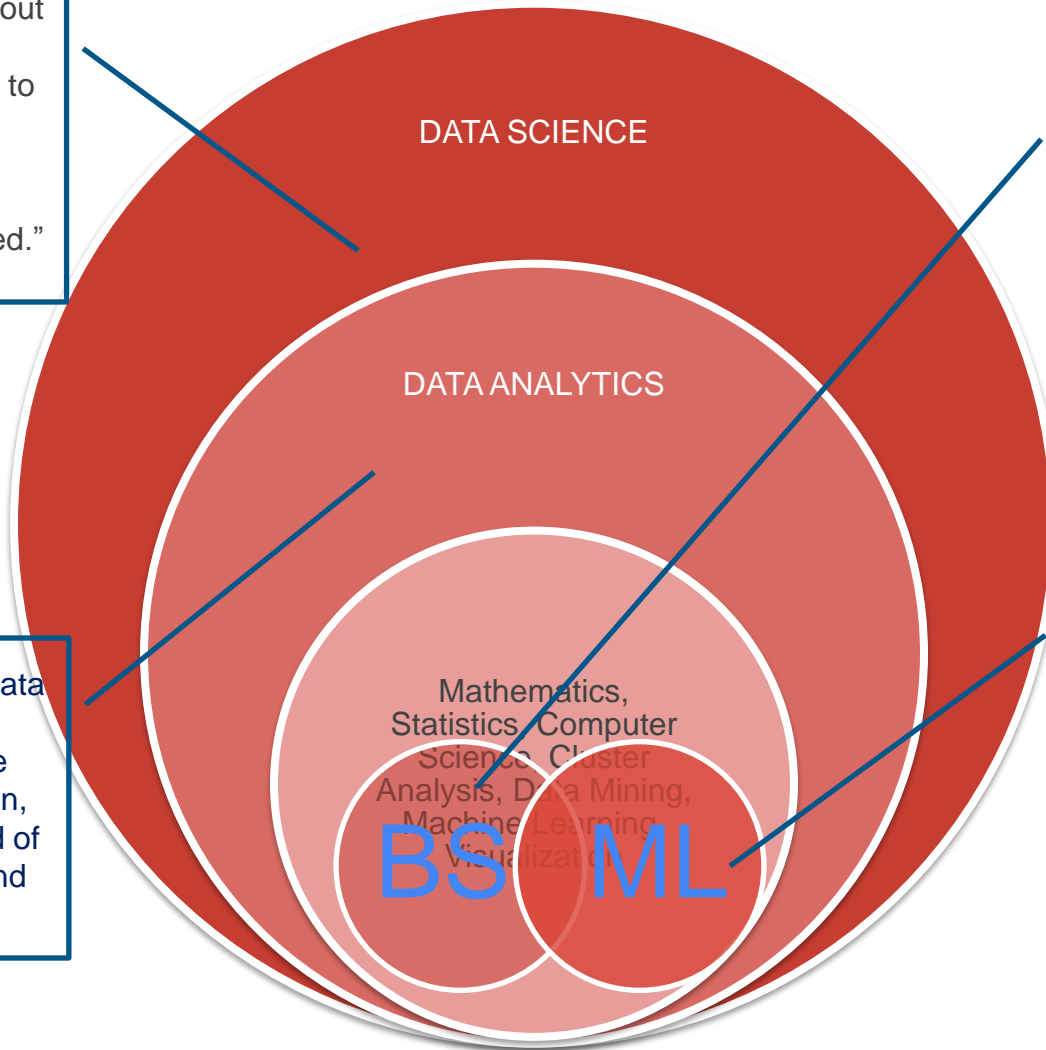
- ❑ Data Science
- ❑ Data Analytics
- ❑ Machine Learning (ML)
- ❑ Business Statistics (BS)

"interdisciplinary field about **scientific** methods, processes and systems to extract knowledge or insights from **data** in various forms, either structured or unstructured."

"collection of statistical procedures and techniques that are used to convert data into meaningful information in a business environment"

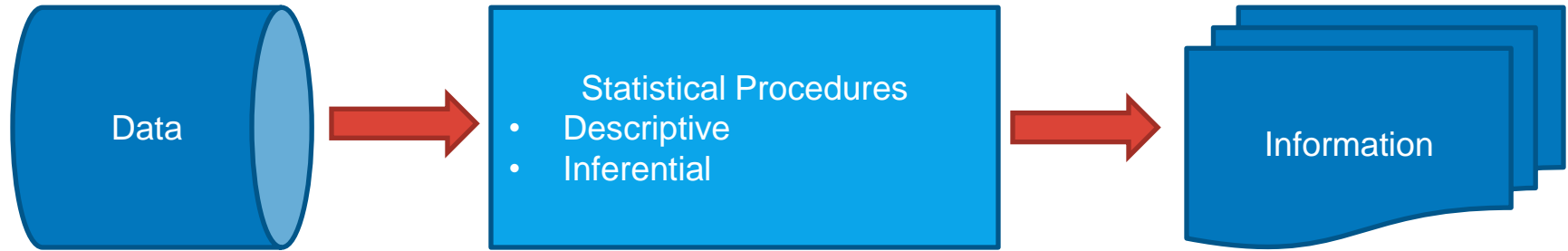
"process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software."

"explores the study and construction of algorithms that can learn from and make predictions on data."



**MLDS**

**STATISTICS:** The branch of **mathematics** that deals with the collection, analysis, interpretation, presentation and organization of data



Descriptive:

Describe about the data either through visual tools (chart or graph ) and/or numerical measures.

Inferential:

Making estimation & hypothesis testing to help in decision making process

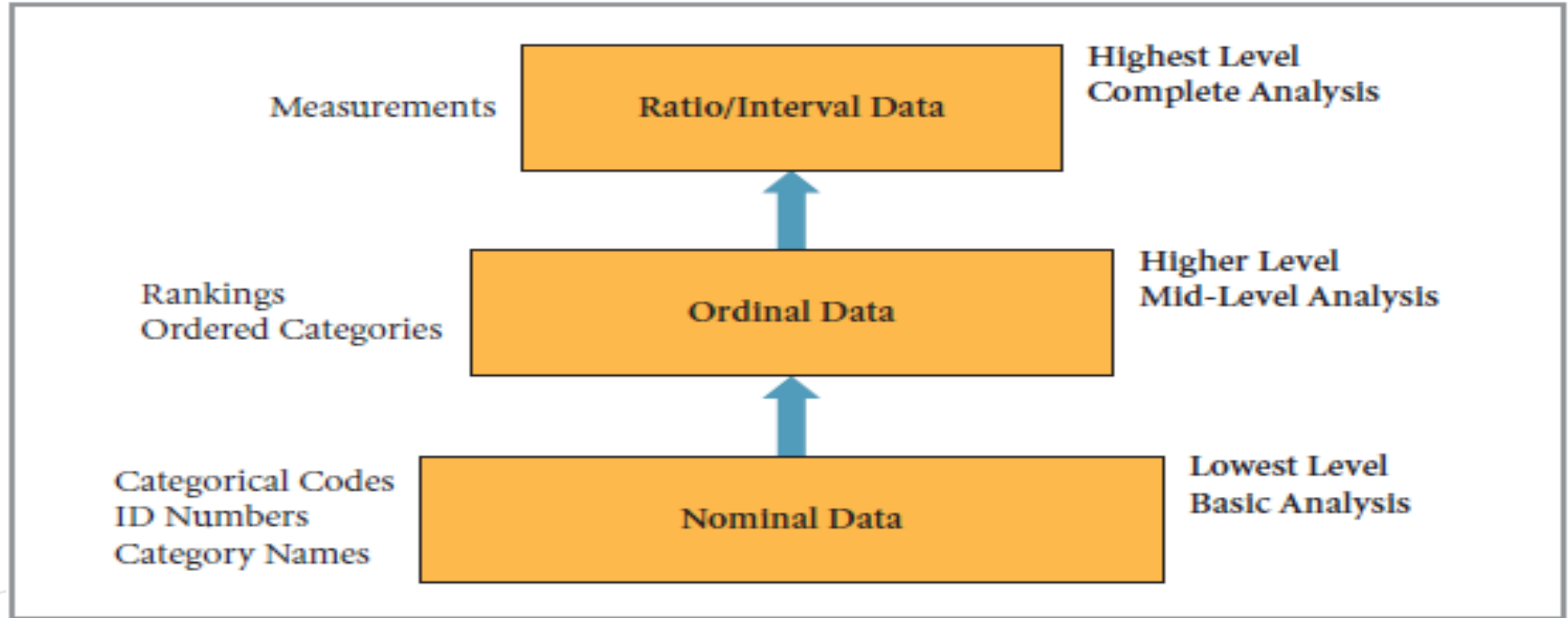
**MLDS**

# DATA TYPE

Why need to know the type of data and the measurement level of data?: **to analyze the data are partially dependent on the level and type of data you have available**

1. Quantitative Data – data whose measurement scale are inherently **numerical**.  
Example: percentage, inches, accuracy...
2. Qualitative Data – data whose measurement scale are inherently **categorical**.  
Example: Scale in income status [1 = very poor], [2 = poor], [3 = neutral], [4 = rich], [5 = very rich]
3. Time-series Data – Set of consecutive data values observed at **successive points in time**.
4. Cross-sectional Data – Set of data values observed at a **fixed point in time**.

# DATA MEASUREMENT LEVEL (DATA HIERARCHY)



# VARIABLES

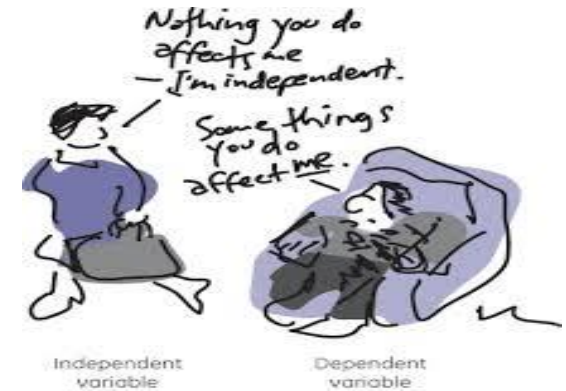
Category:

**Independent variable** – represent inputs or causes

**Dependent variable** – the output or outcome whose variation is being studied

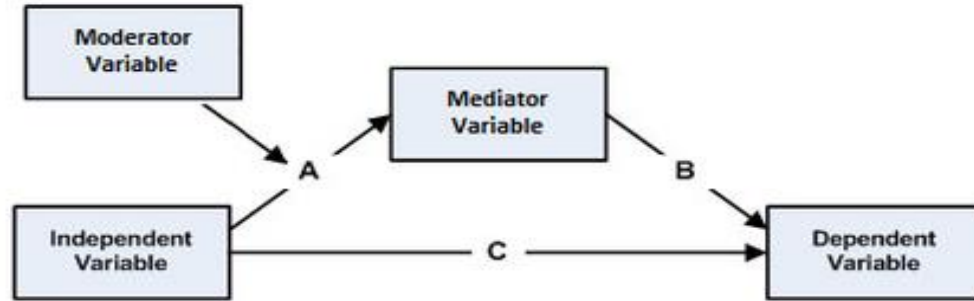
**Moderator (intervening) variable** – influenced by the independent variable, which in turn influences the dependent variable

**Mediator variable** – influence the nature and strength of the relationship (dependent and independent). May reduce or increase cause-and-effect between variables

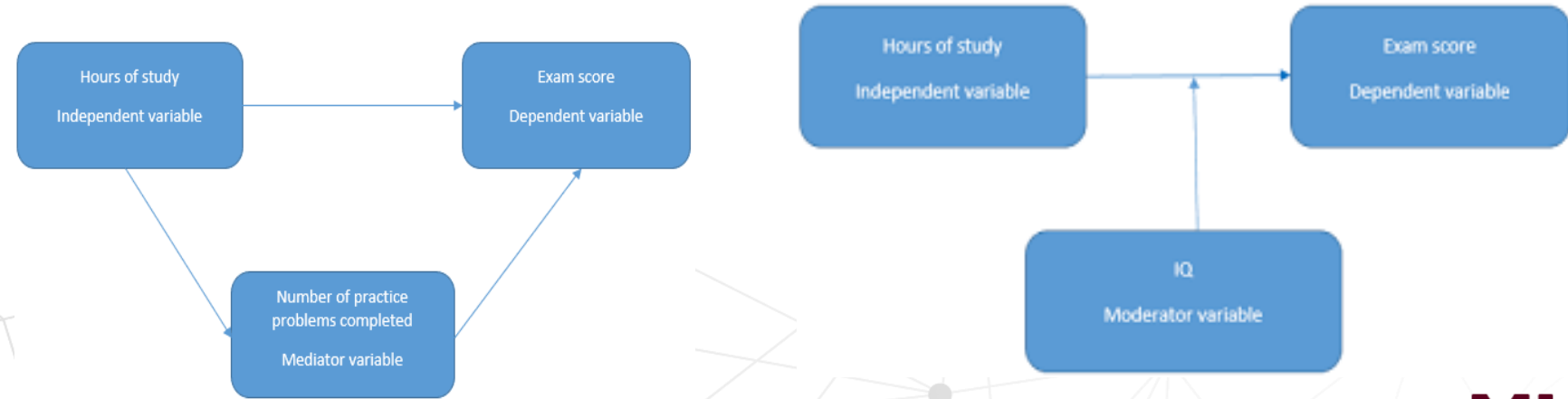




# MODERATOR AND MEDIATOR VARIABLES



## EXAMPLE





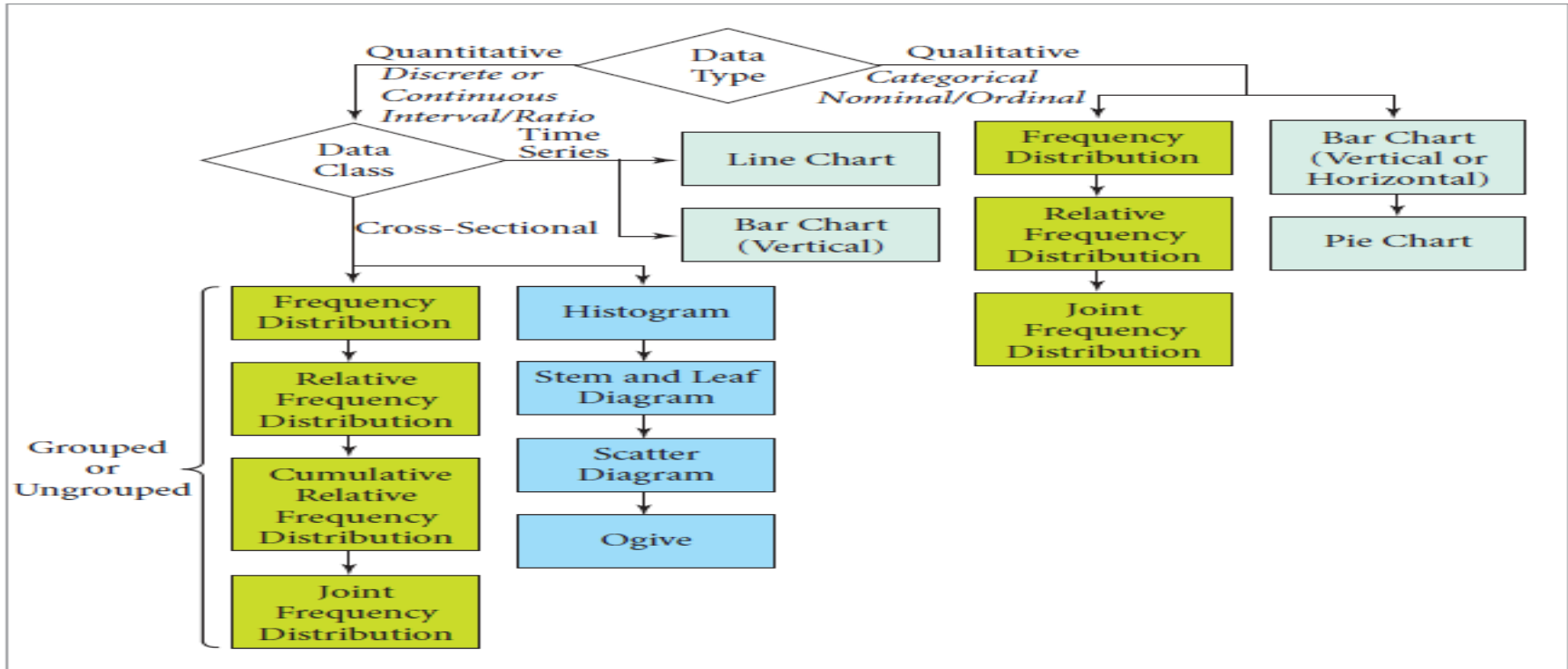
# DESCRIPTIVE STATISTICS

Describing Data through:

- visual tools (chart or graph ) and/or
- numerical measures.

**MLDS**

# DESCRIBING DATA IN A VISUAL WAY



# EXAMPLE: DESCRIBING DATA IN A VISUAL WAY

## – TABLE REPRESENTATION

### PROBLEM: EMERGENCY RESPONSE COMMUNICATION LINKS

“One of the major efforts of the Homeland Security has been to improve the communication between emergency responders, like the police and fire departments. The communications have been hampered by problems involving linking divergent radio and computer systems, as well as communication protocols. While most cities have recognized the problem and made efforts to solve it, Homeland Security recently funded practice exercises in 72 cities of different sizes throughout the United States. The resulting data, already sorted but representing seconds before the systems were linked.” **What can you conclude how many seconds which most cities took to link their communications systems?.**

35	339	650	864	1,025	1,261
38	340	655	883	1,028	1,280
48	395	669	883	1,036	1,290
53	457	703	890	1,044	1,312
70	478	730	934	1,087	1,341
99	501	763	951	1,091	1,355
138	521	788	969	1,126	1,357
164	556	789	985	1,176	1,360
220	583	789	993	1,199	1,414
265	595	802	997	1,199	1,436
272	596	822	999	1,237	1,479
312	604	851	1,018	1,242	1,492

**MLDS**

# SOLUTION

- 1.  $n = 72$  data items, thus, number of classes,  $k = 7$  where  $(2^7 = 128 \geq 72)$
- 2. minimum class width using:  $w = \frac{1492-35}{7} = 208.143$  rounded the class width up from the minimum required value of 208.1429 to the more convenient value of 225

3. Define the class boundaries.

0	and under	225
225	and under	450
450	and under	675
675	and under	900
900	and under	1,125
1,125	and under	1,350
1,350	and under	1,575

4. Determine the class frequency for each class.

Time to Link Systems (in second)	Frequency
0 and under 225	9
225 and under 450	6
450 and under 675	12
675 and under 900	13
900 and under 1,125	14
1,125 and under 1,350	11
1,350 and under 1,575	7

**Summary:** This frequency distribution shows that most cities took between 450 and 1,350 seconds (7.5 and 22.5 minutes) to link their communications systems.

# DESCRIPTIVE STATISTICS

## – NUMERICAL MEASURES

- ❑ Measures of Center and Location
- ❑ Measures of Variation



# 1

## MEASURE OF CENTER LOCATION

**MLDS**

# THE MEASUREMENTS

1. Mean (Population, Sample) - The mean measure can be affected by **extreme values** (for example: income/salary data)
2. Median (Population, Sample)
3. Mode - Mode is the value in a data set that occurs most frequent. Occasionally used as a measure of central location. **HOWEVER**, useful in describing the central location value for sizes (clothes, shoes and etc)



# OTHER MEASURES OF LOCATION : WEIGHTED MEAN

Weighted Mean (population & sample) - The mean value of data values that have been weighted according to their relative importance

**Weighted Mean for a Population**

$$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

**Weighted Mean for a Sample**

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

where:

$w_i$  = The weight of the  $i$ th data value

$x_i$  = The  $i$ th data value

# OTHER MEASURES OF LOCATION: PERCENTILES

The  $p$ th percentile in a data array is a value that divides the data set into **two** parts.

The lower segment contains at least  $p\%$  and the upper segment contains at least  $(100 - p)\%$  of the data.

The **50th percentile is the median**.

## Percentile Location Index

$$i = \frac{p}{100}(n)$$

where:

$p$  = Desired percent

$n$  = Number of values in the data set

# OTHER MEASURES OF LOCATION : PERCENTILES

## Problem:

The Henson Trucking Company is a small company in the business of moving people from one home to another within the Dallas, Texas, area. Historically, the owners have charged the customers on an hourly basis, regardless of the distance of the move within the Dallas city limits. However, they are now considering adding a surcharge for moves over a certain distance. They have decided to base this charge on the 80th percentile. They have a sample of travel-distance data for 30 moves. These data are as follows:

13.5	8.6	16.2	21.4	21.0	23.7	4.1	13.8	20.5	9.6
11.5	6.5	5.8	10.1	11.1	4.4	12.2	13.0	15.7	13.2
13.4	13.1	21.7	14.6	14.1	12.4	24.9	19.3	26.9	11.7

## Solution:

1. Sort the data from lowest to highest
2. Determine the percentile location index,  $i$

$$i = \frac{p}{100}(n) = \frac{80}{100}(30) = 24$$

3. Locate the appropriate percentile

The 80th percentile is found by averaging the values in the 24th and 25th positions. These are 20.5 and 21.0. Thus, the 80th percentile is  $(20.5+21.0)/2 = 20.75$

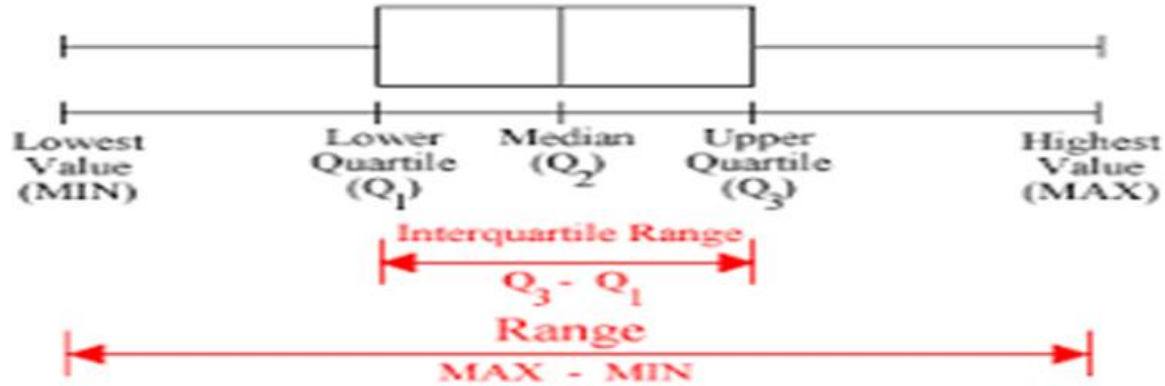
Summary: Any distance exceeding 20.75 miles will be subject to a surcharge.

# OTHER MEASURES OF LOCATION : QUARTILES

Values that divide the data set into four equal-sized groups.

The 1<sup>st</sup> quartile ( $Q_1$ ) = 25<sup>th</sup> percentile, 2<sup>nd</sup> quartile ( $Q_2$ ) = median, 3<sup>rd</sup> quartile ( $Q_3$ ) = 75<sup>th</sup> percentile and 4<sup>th</sup> quartile = 100<sup>th</sup> percentile

# OTHER MEASURES OF LOCATION : BOX AND WHISKERS



Descriptive tool that many decision maker use

Used to identify outliers.

# DATA-LEVEL ISSUES TO MEASURE CENTER OF LOCATION

Descriptive Measure	Computation Method	Data Level	Advantages/ Disadvantages
Mean	Sum of values divided by the number of values	Ratio Interval	<ul style="list-style-type: none"><li>• Numerical center of the data</li><li>• Sum of deviations from the mean is zero</li><li>• Sensitive to extreme values</li></ul>
Median	Middle value for data that have been sorted	Ratio Interval Ordinal	<ul style="list-style-type: none"><li>• Not sensitive to extreme values</li><li>• Computed only from the center values</li><li>• Does not use information from all the data</li></ul>
Mode	Value(s) that occur most frequently in the data	Ratio Interval Ordinal Nominal	<ul style="list-style-type: none"><li>• May not reflect the center</li><li>• May not exist</li><li>• Might have multiple modes</li></ul>

# 2

## MEASURE OF VARIATION

A decorative network diagram at the bottom of the slide, consisting of a series of interconnected nodes (small circles) and lines (edges) forming a complex web-like structure.

**MLDS**

# WHY NEED TO MEASURE VARIATION?

Example:

Plant A	Plant B		Plant A	Plant B
15 units	23 units			
25 units	26 units	→	$\bar{x} = 25$ units	$\bar{x} = 25$ units
35 units	25 units		$M_d = 25$ units	$M_d = 25$ units
20 units	24 units			
30 units	27 units			

Table: Manufacturing output for ABX

Summary: The descriptive statistics (mean and median) are **equal**, the **distribution** of production output at the two plants is **symmetrical**. Therefore, the two plants are equal in terms of their production output.

**HOWEVER!!!:** There is a **HUGE** variation in terms of day by day of the production output. Plant B is very stable, producing almost the same amount every day. Plant A varies considerably, with some high-output days and some low-output days.

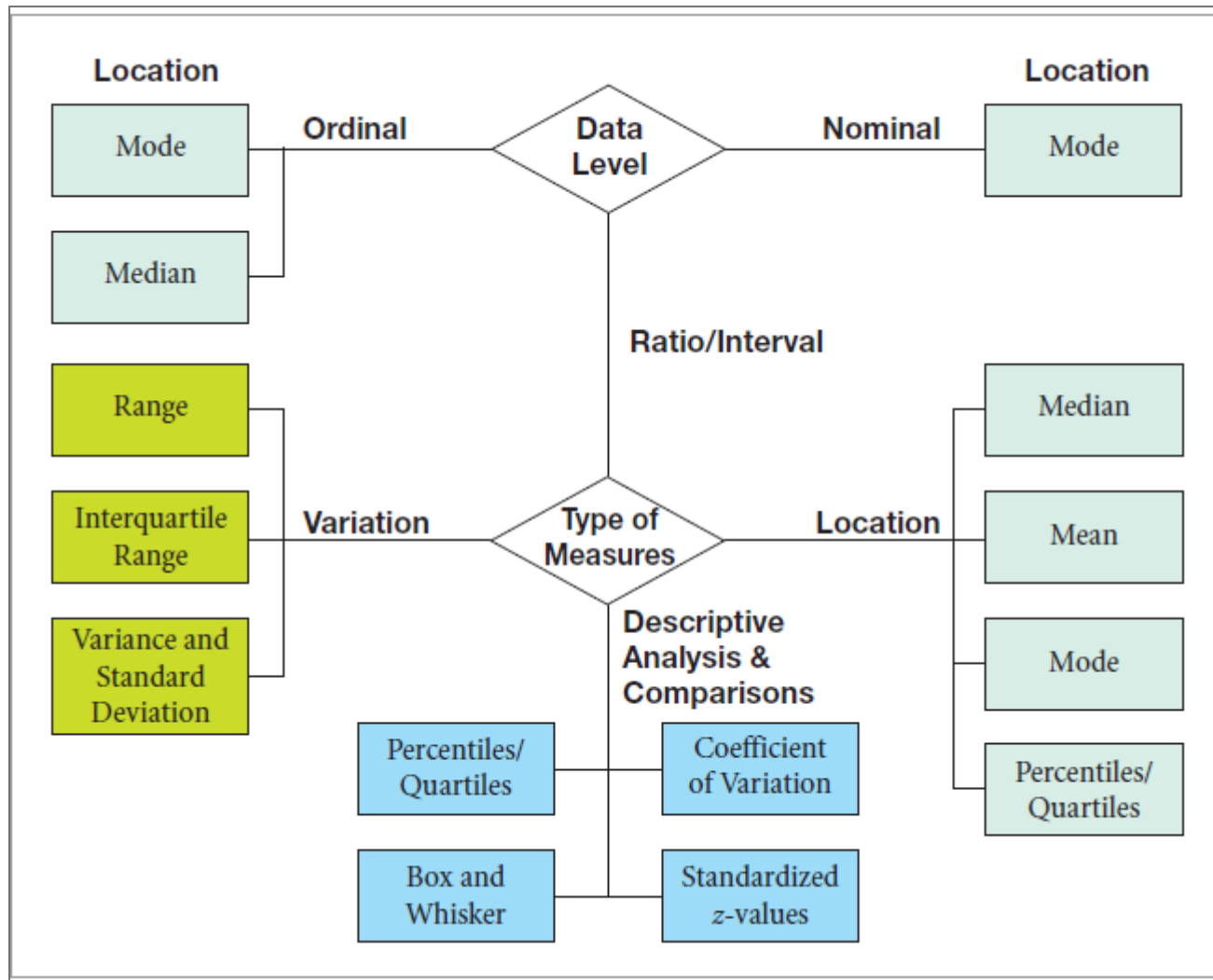
**THUS:** looking at **only** measures of the data's central location can be misleading. We need MEASURES OF VARIATION



# MEASURE OF VARIATION

Several different measures of variation that are used in business decision making are:

- Range
- Interquartile Range
- Variance
- Standard Deviation



# INFERENCEAL STATISTICS

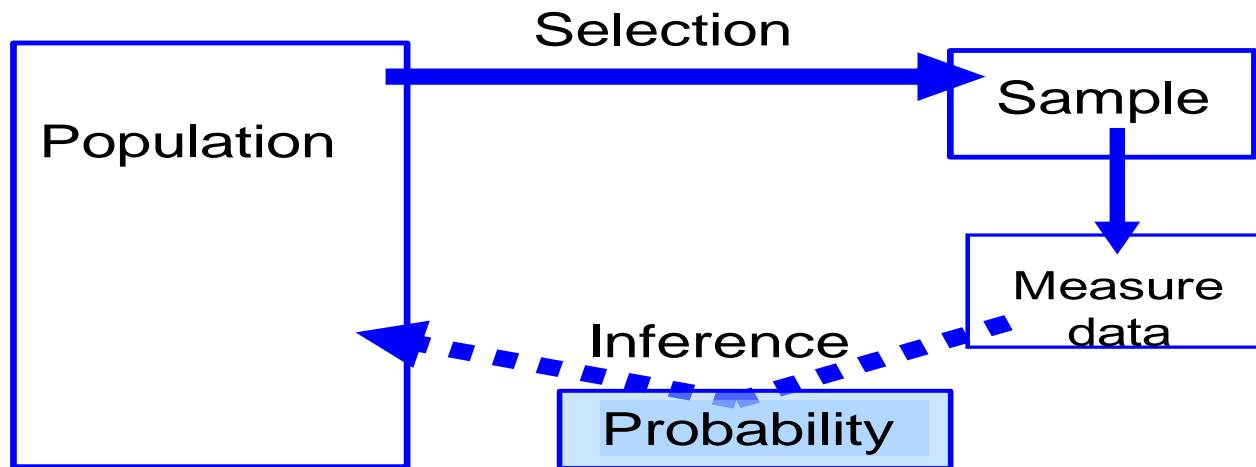
Making estimation & hypothesis testing to help in decision making process

**MLDS**



"I can prove it or disprove it! What do you want me to do?"

# Chain of Reasoning for Inferential Statistics



*Are our inferences valid? ...Best we can do is to calculate probability about inferences*

# INFERENCE STATISTICS



## ESTIMATION

- “guessing” the value of the parameter
- provide the measure of the quality (reliability) of the guess



## HYPOTHESIS TESTING

- making a “yes-no” decision regarding the parameter
- understand the chances of making incorrect decision



**MLDS**



# HYPOTHESIS TESTING

- **Direct our observations**

Identifies the variables examined and data to be collected

- **Describe a relationship among variables**

Can state that as one variable increases, the other will decrease; as one variables increases, the other will increase, and so on.

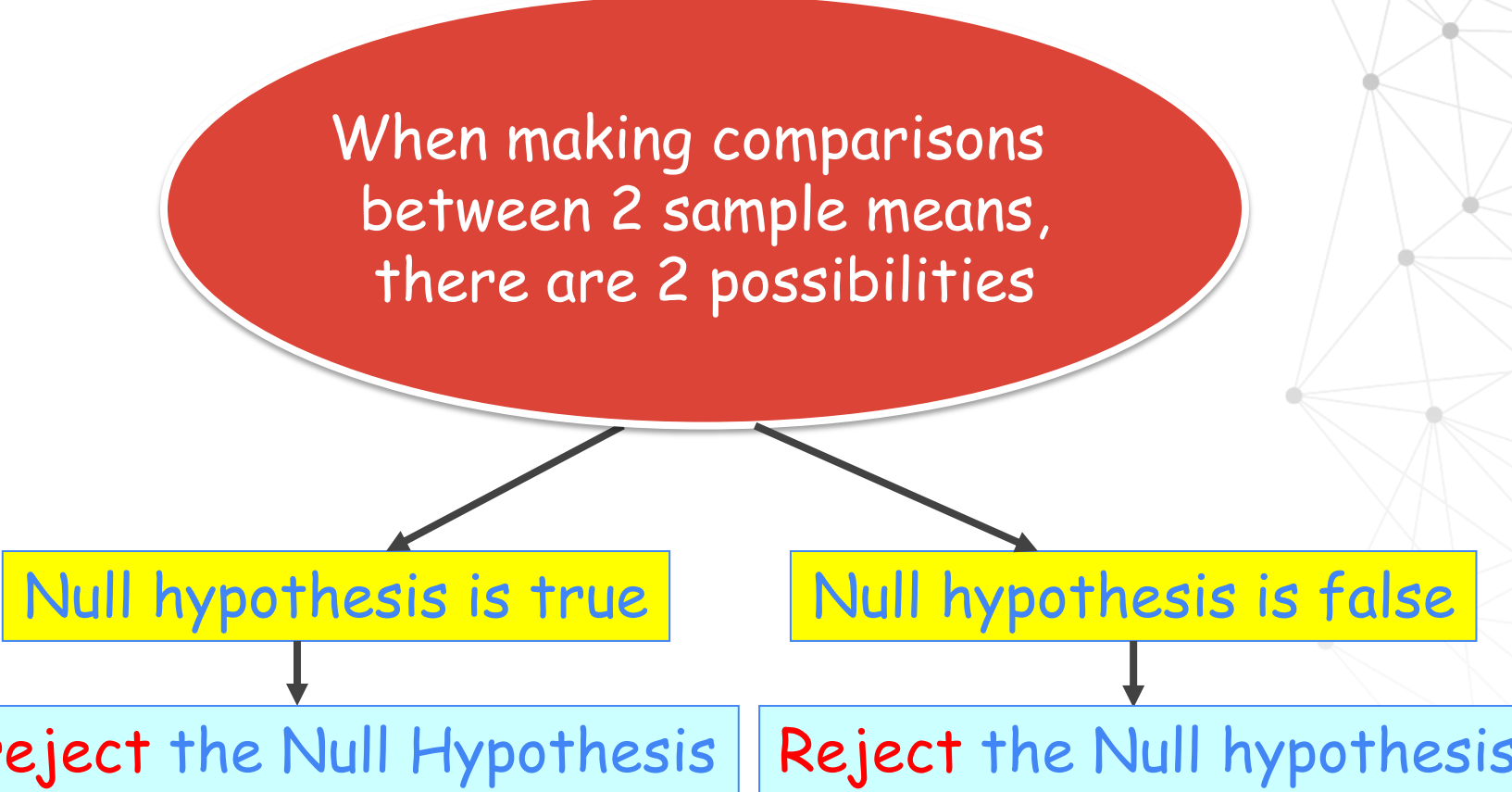
- **Refer to populations**

Hypotheses help researchers infer that results of a sample will translate to a population

# TYPES OF HYPOTHESIS

Type of Hypothesis Test	Null Hypothesis	Alternative Hypothesis
Comparison of samples	The samples are the same; the samples come from the same population; the characteristics of the samples differ only because of sampling randomness	The samples are fundamentally different
Correlation	There is no significant correlation between the variables	The correlation is statistically significant
Normality	The data are normally distributed	The data are not normally distributed

When making comparisons  
between 2 sample means,  
there are 2 possibilities



```
graph TD; A([When making comparisons between 2 sample means, there are 2 possibilities]) --> B[Null hypothesis is true]; A --> C[Null hypothesis is false]; B --> D[Not reject the Null Hypothesis]; C --> E[Reject the Null hypothesis];
```

The diagram is a flowchart illustrating the two possible outcomes of a hypothesis test when comparing two sample means. It starts with a red oval at the top containing the text 'When making comparisons between 2 sample means, there are 2 possibilities'. Two arrows point down from this oval to two yellow rectangular boxes. The left box contains 'Null hypothesis is true' and the right box contains 'Null hypothesis is false'. From each yellow box, an arrow points down to a light blue rectangular box. The left light blue box contains 'Not reject the Null Hypothesis' and the right light blue box contains 'Reject the Null hypothesis'. The words 'Not reject' and 'Reject' are in red, while 'the Null Hypothesis' and 'the Null hypothesis' are in blue. In the bottom right corner, the text 'MLDS' is written in a bold, dark red font. A faint, light gray network diagram is visible in the background on the right side of the slide.

Null hypothesis is true

Null hypothesis is false

Not reject the Null Hypothesis

Reject the Null hypothesis



# WHEN WE DO STATISTICAL ANALYSIS...

if alpha ( $p$  value- significance level) greater than 0.05



WE ACCEPT THE NULL HYPOTHESIS (no difference btwn means)

if alpha ( $p$  value- significance level) is equal to or less than 0.05 we



REJECT THE NULL (difference btwn means)

**MLDS**

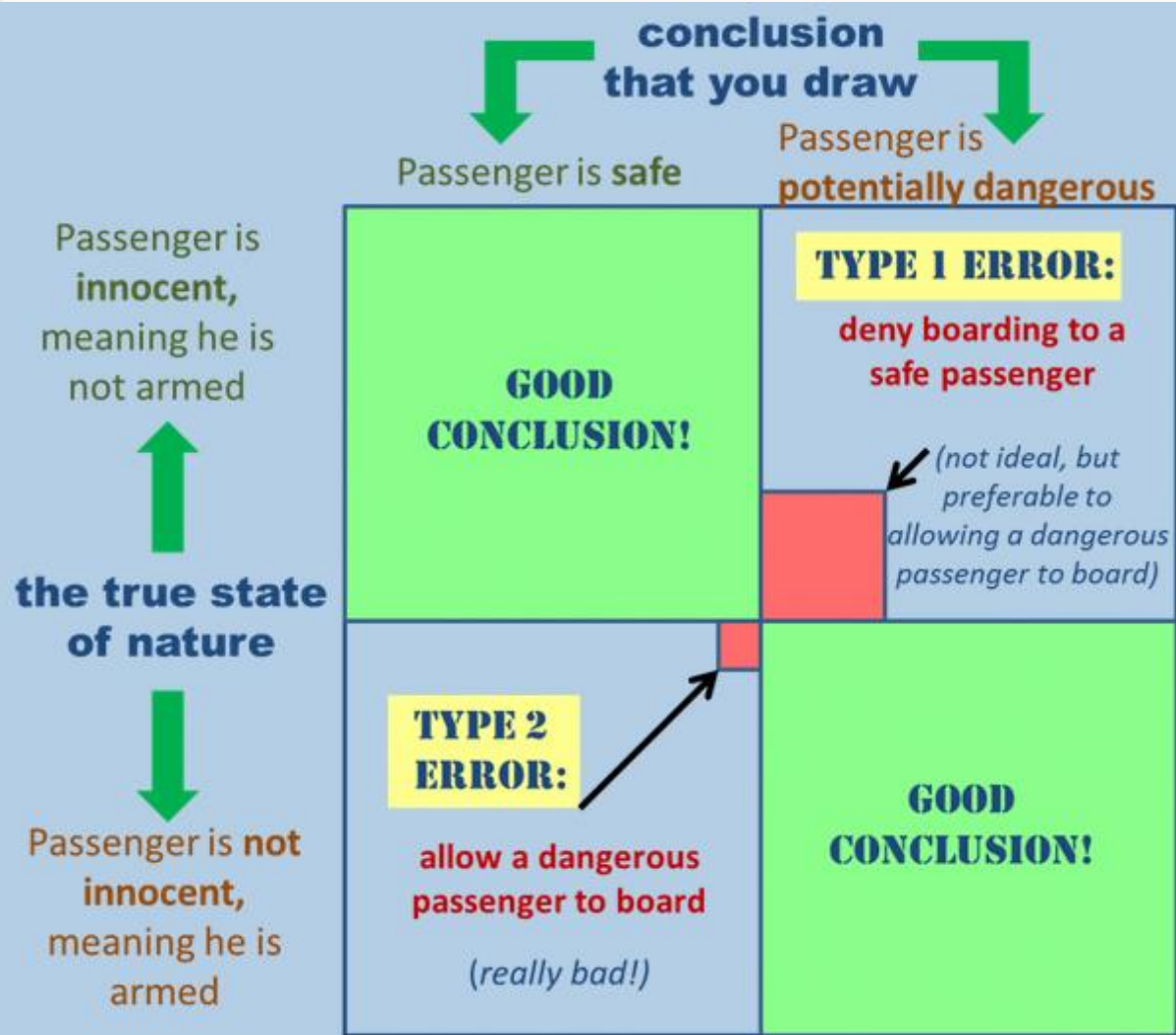
# POSSIBLE OUTCOMES IN HYPOTHESIS TESTING (DECISION)

	Null is True	Null is False
Accept	Correct Decision	Error Type II Error
Reject	Error Type I Error	Correct Decision

Type I Error ( $\alpha$ ): Rejecting a hypothesis which is actually true in nature

Type II Error ( $\beta$ ): Accepting a hypothesis which is actually false in nature

# POSSIBLE OUTCOME IN HYPOTHESIS TESTING: AIRLINE PASSENGER SCREENING



# ESTIMATION



- Estimation from sample are only guesses (of the parameter)
- Every estimate has a standard error, and it is measure of the variation in the estimates
- If you were to repeat the study, you would get a different answer
- With 2 answers:

# An Estimate of a population parameter may be expressed in:

## POINT ESTIMATE

- A point estimate of a population parameter is a single value of a statistics
- For example, the sample mean  $\bar{x}$  is a point estimate of the population mean,  $\mu$
- Similarly, the sample proportion  $p$  is a point estimate of the population proportion  $P$ .

## INTERVAL ESTIMATE

- Is defined by two numbers, between which parameter is said to lie
- For example,  $a < x < b$  is an interval estimate of the population mean  $\mu$ . It indicates that the population mean is greater than  $a$  but less than  $b$



**PROBABILITY**

**MLDS**

# WHAT IS PROBABILITY?

- The chance that a particular event will occur.
- The probability value will be in the range 0 to 1.

A value of 0 means the event will not occur.

A probability of 1 means the event will occur.

- Anything between 0 and 1 reflects the uncertainty of the event occurring.

# PROBABILITY: WHY YOU NEED TO KNOW?

1. In business decision making, there are many instances where chance is involved in determining the outcome of a decision.

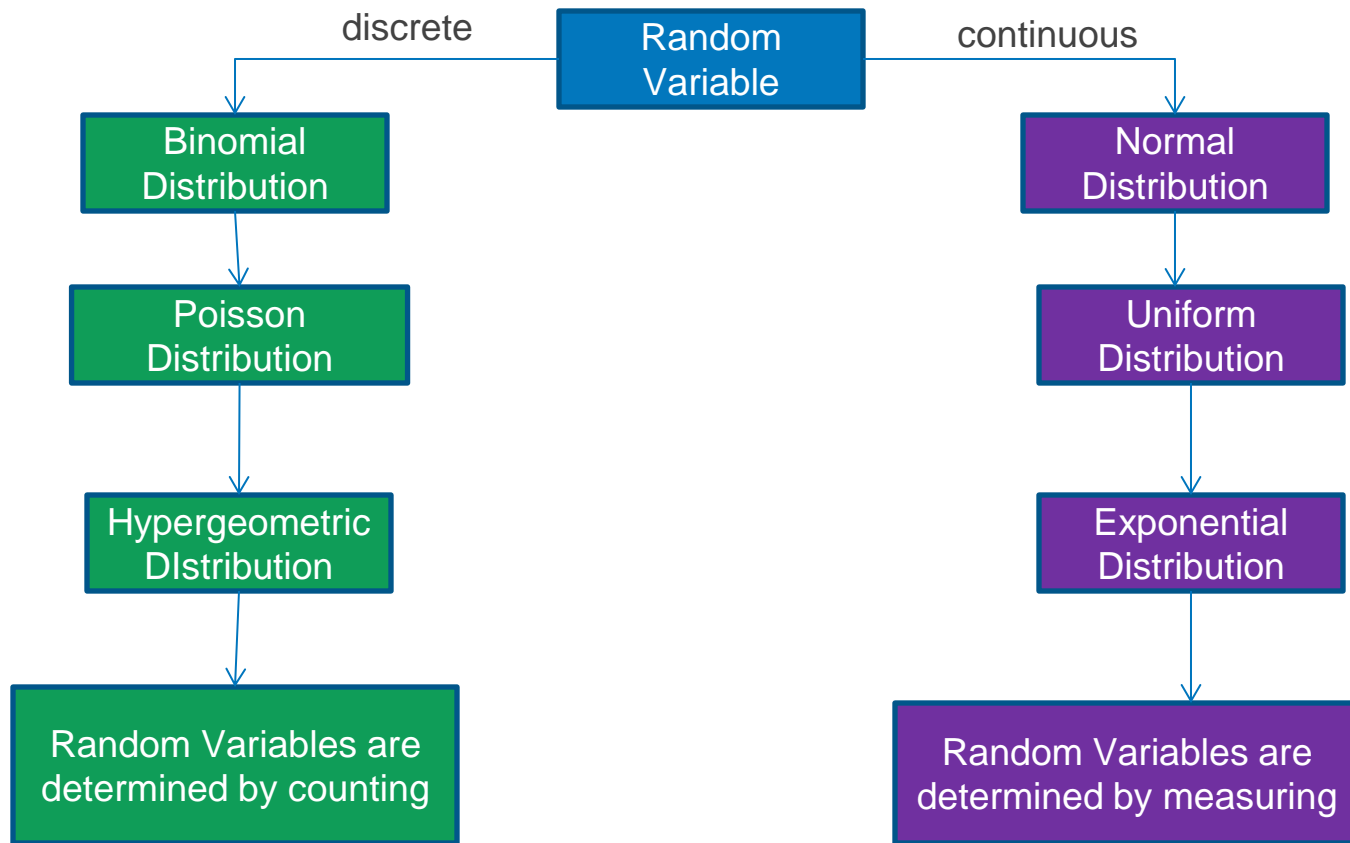
**Example:** Airlines overbook flights to make sure that the seats on the plane are as full as possible because they know there is a certain probability that customers will not show for their flight.

2. Probability is the way decision makers express their uncertainty about outcomes and events.

**Example:** personnel manager has a chance to promote 3 people from 10 equally qualified candidates. Suppose none of 6 women are selected by the manager. Is this evidence of gender bias or would we expect to see this type of result?



# PROBABILITY DISTRIBUTION





# APPLICATION OF LINEAR ALGEBRA

# WHAT IS LINEAR ALGEBRA?

**Algebra** is the branch of mathematics that deals with **unknown values** being represented in the form of variables.

**Linear algebra** is an area that primarily deals with **representation of data** conforming to certain notations and practices.

- **linear algebra data** is represented in the form of **linear equations**.
- These linear equations are in turn represented in the form of **matrices and vectors**.
- **Vectors**: can be looked at as a single dimensional matrix

# THE APPLICATION OF LINEAR ALGEBRA IN MACHINE LEARNING

Machine Learning:

- Data intensive → Machine Learning algorithms aren't very effective unless they're trained and are operating on large data sets, **ranging from hundreds of training examples to millions of testing data.**

With linear algebra:

- representing large sets of data in the form of matrices help us visualize the data better.
- **Matrices help us to look at all the data as a single entity and also let us process them as and how we look at them.**

# Thanks!

Machine Learning for Data Science Interest Group  
Advanced Informatics School  
Universiti Teknologi Malaysia

@utmmls

[ais.utm.my/mls](http://ais.utm.my/mls)

[huda@utm.my](mailto:huda@utm.my)

February 2017



**MLDS**