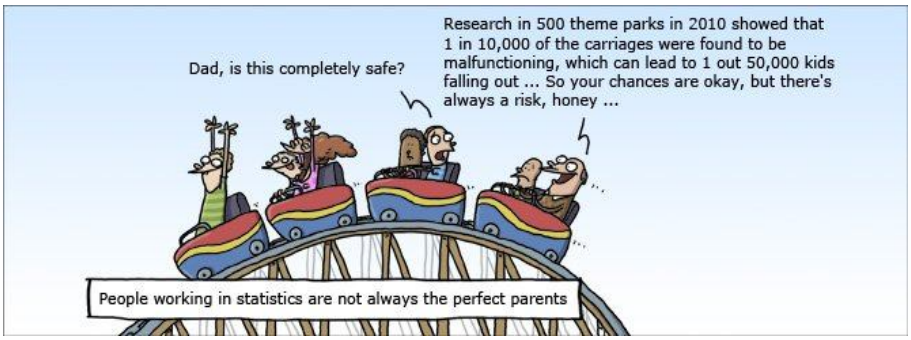
SCSI2134: PROBABILITY & STATISTICAL DATA ANALYSIS

CHAPTER 1

Introduction to Statistics

Introduction to Statistics

- Statistics?



Introduction to Statistics

Number of fatal road accidents up in 2016, more than 7,000 lives lost

KUALA LUMPUR: In 2016, total of 7,152 people died in road accidents in Malaysia, an alarming jump from 6,706 deaths in the year before, Transport Minister Datuk Seri Liow Tiong Lai announced today.

"Despite the number of campaigns and initiatives conducted by the government, the painful reality is that the accident rate keeps increasing every year.

"It's too late when tragic accidents like the bus crash in Pagoh, which claimed 14 lives, occur. There is no use accusing anyone and finger pointing then," he said at the launch of the Chinese New Year Road Safety campaign in Sunway Pyramid here.

He added that of that number, 62.7 per cent of the deaths involved motorcyclists in 2016.

"In 2016, 6,570 fatal road accidents involving were recorded while in 2015 there were 6,193 of such cases.

Liow said a total of 521,466 accidents were recorded in 2016, an increase from 489,606 in 2015.

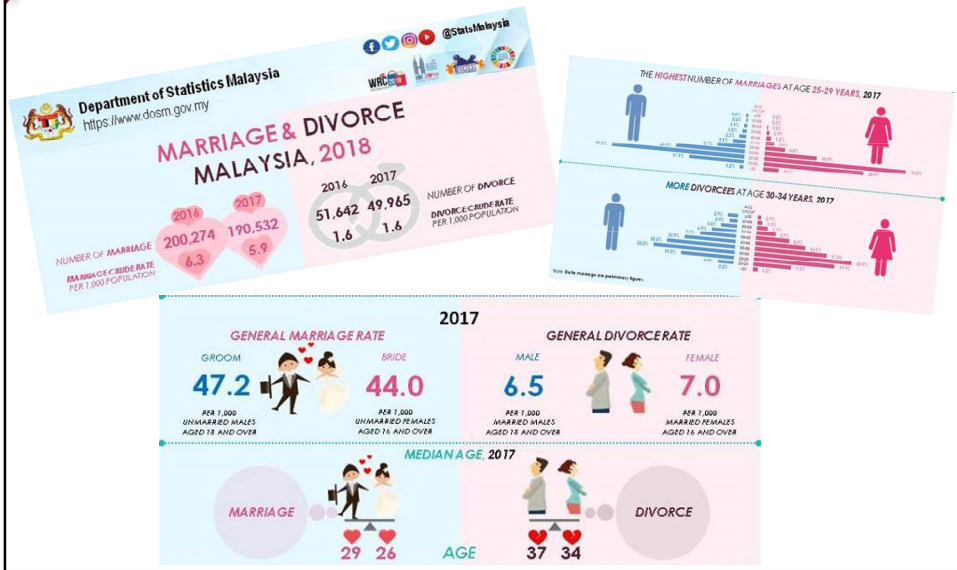
"A total of 80.6 per cent of the road accidents are caused by human error.

"I hope that those on the road this festive season do not speed to their destination (just) to arrive early and gamble their lives or that of their loved ones. It's okay to be late as long as you are safe," he advised.

The joint campaign involving the police and other transport authorities like the Road Transport Department will begin on Jan 21 and end on Feb 5.

He also said that there would be increased patrols by authorities at 102 road accident hotspots nationwide.

Introduction to Statistics



Introduction to Statistics

Introduction to Statistics

- 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.

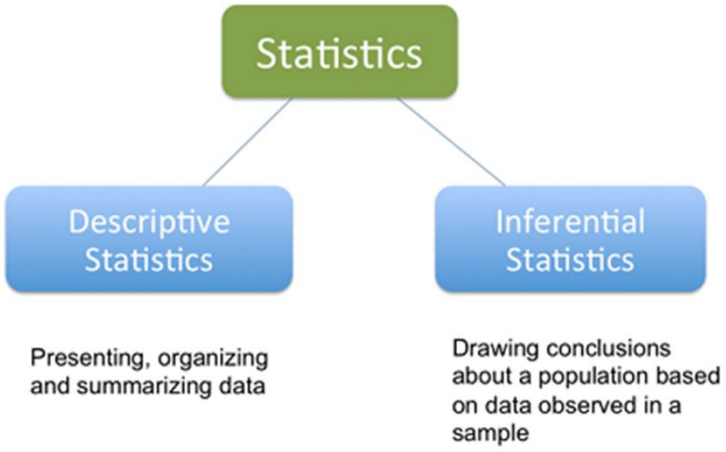
Introduction to Statistics

- 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.

Introduction to Statistics

- 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

Introduction to Statistics



Introduction to 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.

Descriptive Statistics

Graphical description

Example:
Graph
Bar chart
Pie chart

Descriptive Statistics

Tabular description

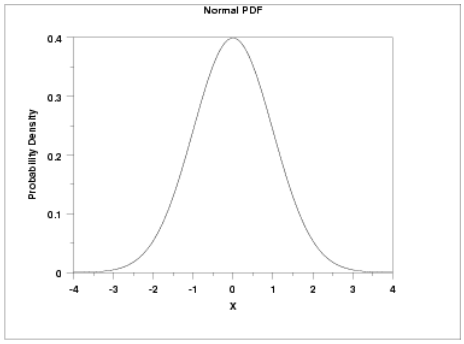
Example:
Frequency Table

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

Descriptive Statistics

Parametric description

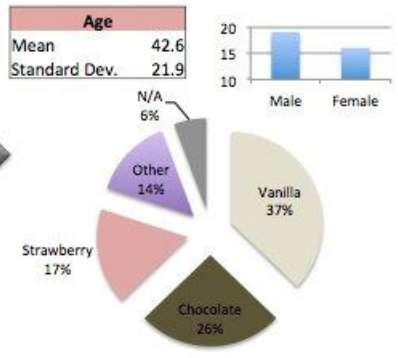
Mean	μ
Median	μ
Mode	μ
Range	Infinity in both directions.
Standard Deviation	σ
Skewness	0
Kurtosis	3



Example

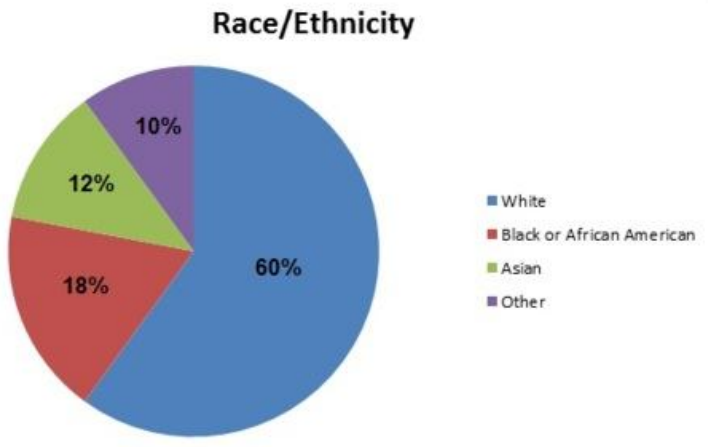
	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15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

Raw Data

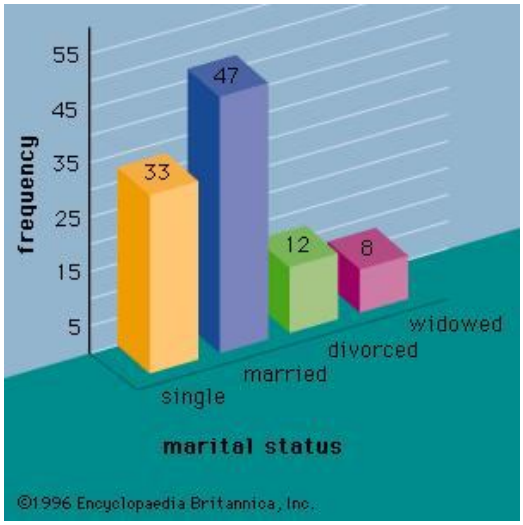


Descriptive Statistics

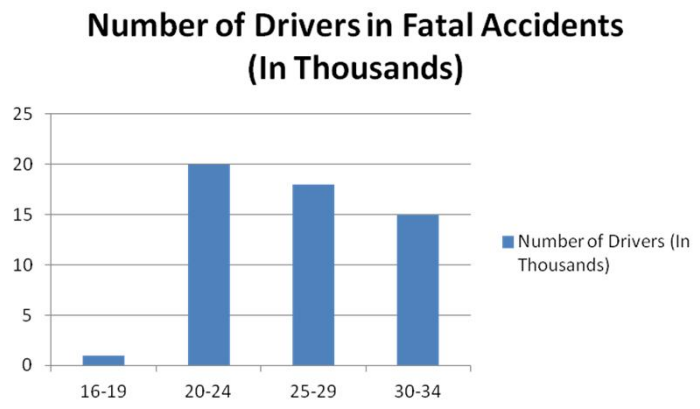
Example



Example



Example



Example

Average Monthly Temperature	
	Fahrenheit
January	26.4
February	28.0
March	35.8
April	46.6
May	56.3
June	65.8
July	71.2

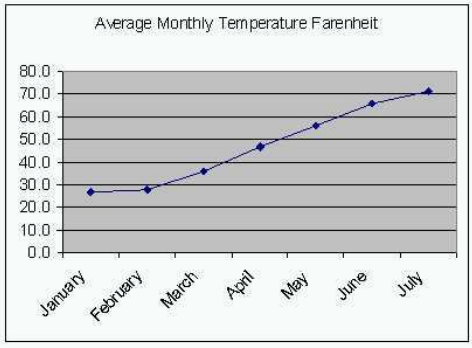
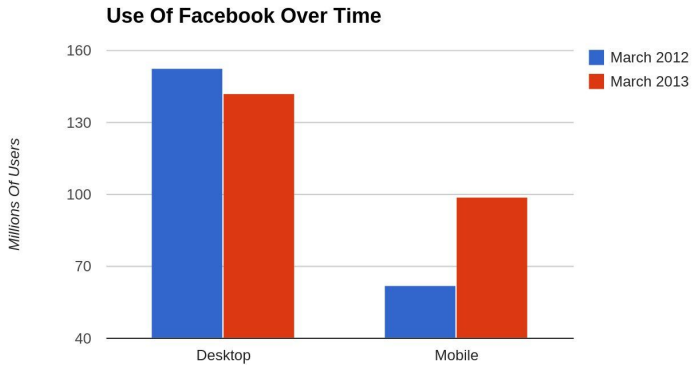


Figure 2

Example



Example

	A	B	C	D	E	F	G
5	*Sample 1*	*Sample 2*		*Sample 1*		*Sample 2*	
6	19	12					
7	41	27		Mean	30.46154	Mean	30.61538
8	29	18		Standard Error	4.673459	Standard Error	5.447345
9	18	23		Median	29	Median	27
10	8	72		Mode	29	Mode	27
11	29	27		Standard Deviation	16.8504	Standard Deviation	19.64068
12	11	27		Sample Variance	283.9359	Sample Variance	385.7564
13	59	53		Kurtosis	-1.15073	Kurtosis	0.062841
14	41	3		Skewness	0.265601	Skewness	0.78698
15	48	45		Range	51	Range	69
16	53	53		Minimum	8	Minimum	3
17	29	13		Maximum	59	Maximum	72
18	11	25		Sum	396	Sum	398
19				Count	13	Count	13

Example

Table 2 – Clinical and demographic characteristics of patients undergoing pharmacological prophylaxis.

Variable	n = 357			
	Minimum	Maximum	Average	Standard deviation
Age (years)	24	71	40.35	8.829
Weight (kg)	50	130	71.43	12.211
Height (cm)	148	184	162.75	5.980
BMI (kg/m^2)	17.93	52.07	26.9731	4.43736
Weight of the flap (g)	650	9200	1422.82	573.049
Surgery time (min)	55	240	135.30	41.267

BMI = body mass index; n = number of patients.

Example

TABLE 4
Descriptive statistics of students' attitudes toward statistics

	N	Mode	s.d.	Skewness	Kurtosis
Affect	234	4.6	1.32	-0.45	0.17
Cognitive Competence	232	5.33	1.16	-0.19	-0.62
Value	234	4.88	1.12	-0.62	0.25
Difficulty	233	3.4	1.13	-0.13	-0.65
Interest	231	7	1.26	-1.03	1.09
Effort	235	7	0.98	-2.32	7.85

s.d. = standard deviation
Higher value = more positive attitude

Example

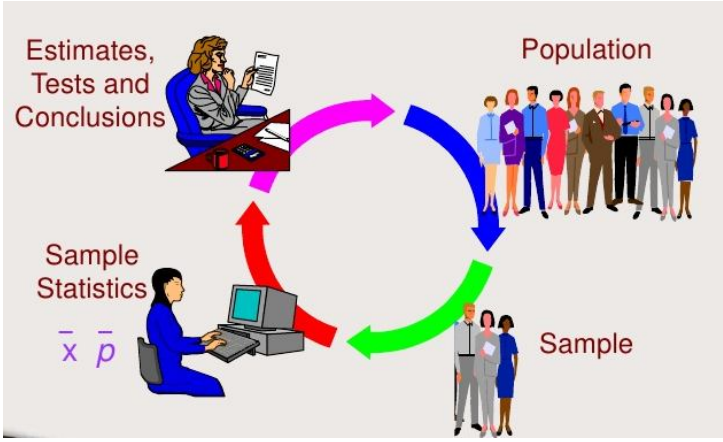
Variable	Mean	Std. Dev	Min	Max
Loans	0.204E + 07	0.234E + 07	0.000	0.178E + 08
Deposits	0.370E + 07	0.430E + 07	0.000	0.368E + 08
Physical capital	0.488E + 07	0.534E + 07	0.000	0.422E + 08

Source: Banking Supervision Department, Bank of Ghana, and the ARB Apex Bank.

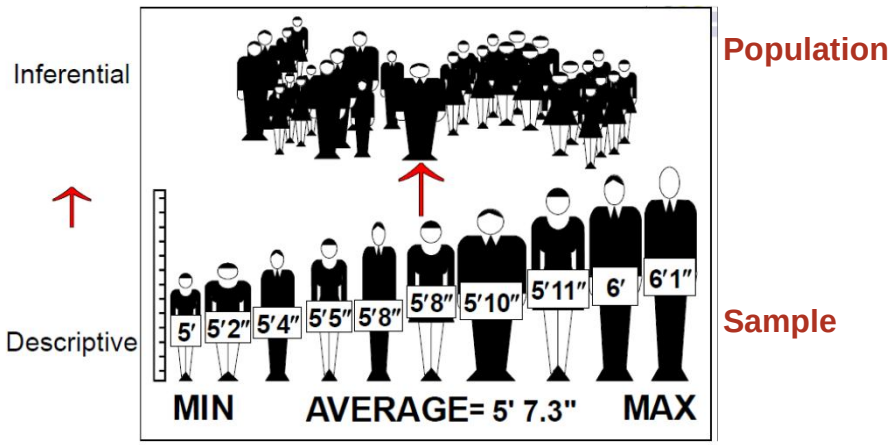
Introduction to Statistics

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

Inference Process



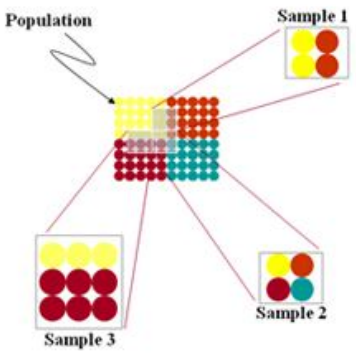
Introduction to Statistics



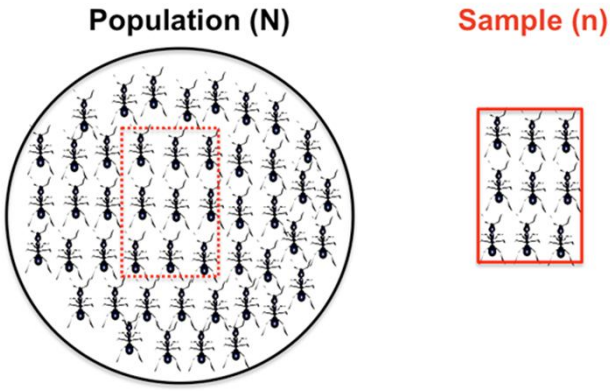
Population & Sample

Population & Sample

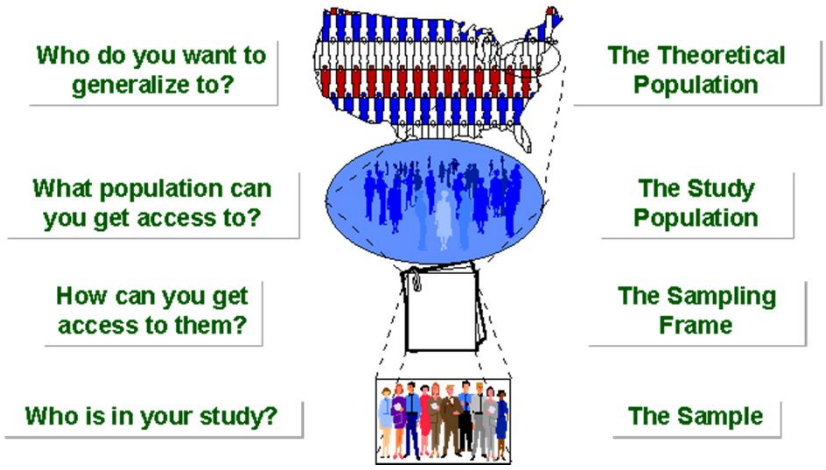
- The entire collection of individuals or object about which information is desired is called the **population** of interest.
- A **sample** is a subset of the population, selected for study in some prescribed manner



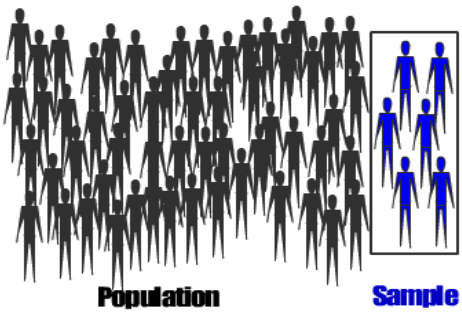
Population & Sample



Population & Sample

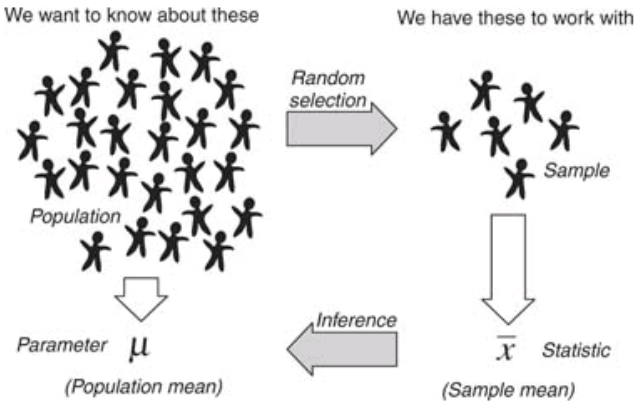


Population & Sample

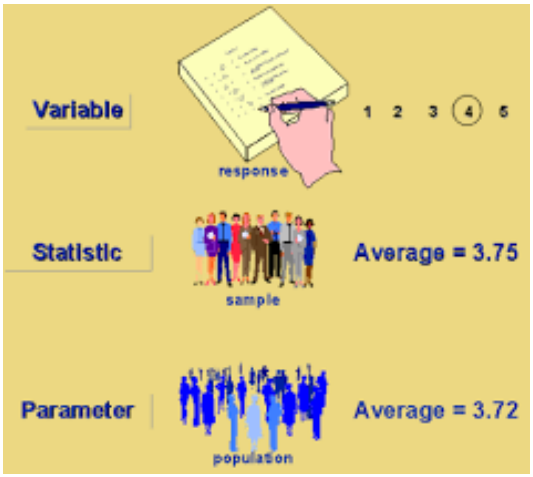


Population	Sample
quantity (count) = N	quantity (count) = n
mean = μ	mean = $\bar{x}$
variance = σ^2	variance = s^2
standard deviation = σ	standard deviation = s

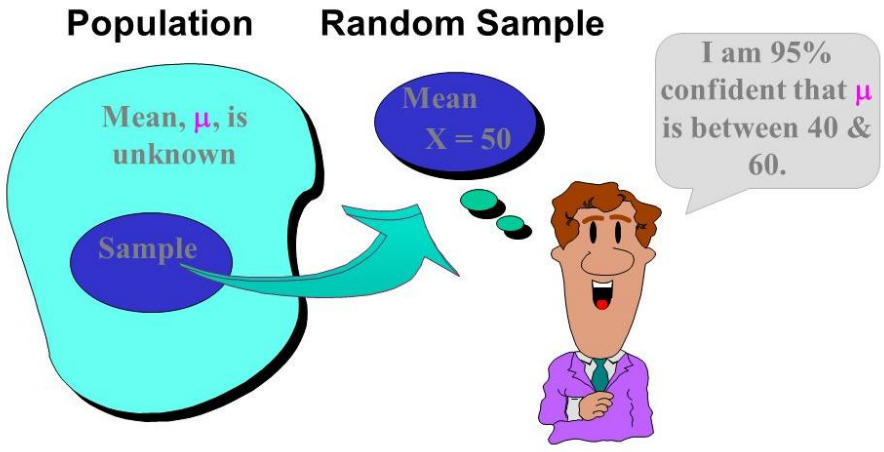
Population & Sample



Example

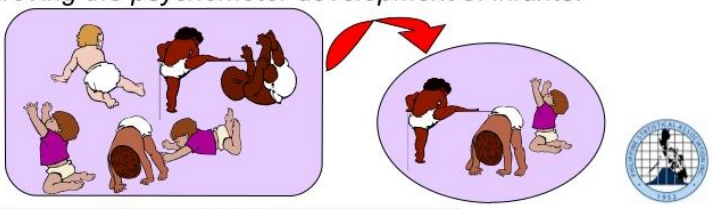


Example

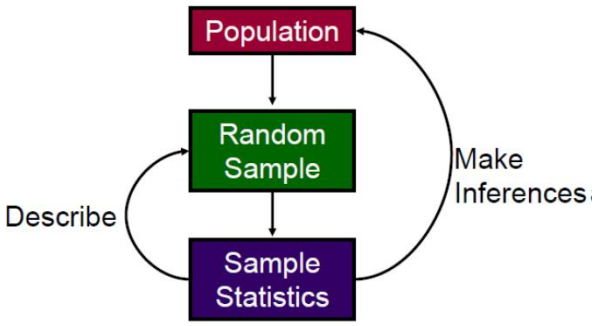


Example

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.



Data Analysis Process



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.

Data Analysis Process

- 6 steps:
  - Understanding the nature of the problem
  - Deciding what to measure and how to measure it
  - Data collection
  - Data summarization and preliminary analysis
  - Formal data analysis
  - Interpretation of results

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
   - 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.

Data Analysis Process

– but in other cases the choice of information is not as straightforward, e.g. 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.

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.

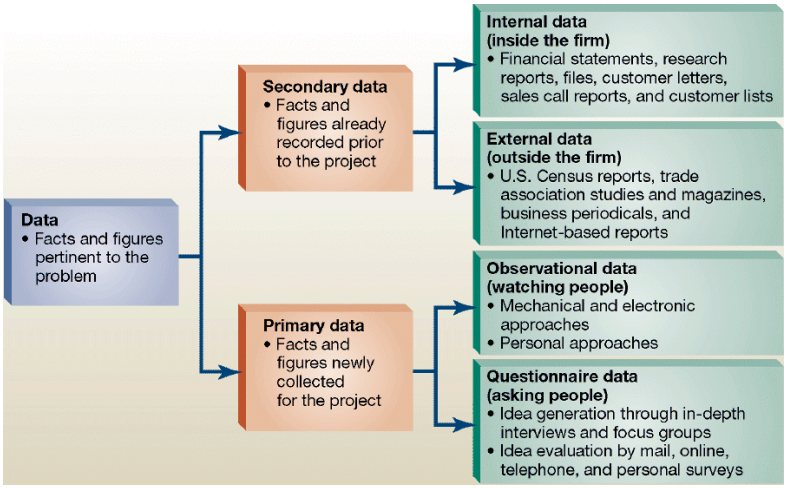
6. Interpretation of results
   - What conclusions can be drawn from the analysis?
   - How do the result of the analysis inform us about the stated research problem or question?
   - How can our results guide future research?

Primary & Secondary Data

- **Secondary data** is data which has been collected by individuals or agencies for purposes other than those of our particular research study.
  - For 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 data**, by contrast, are collected by the investigator conducting the research.

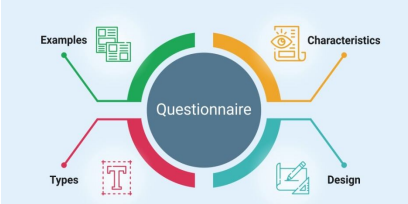
Primary & Secondary Data

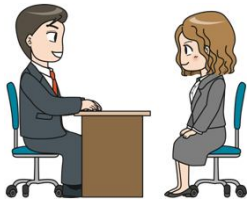


Data Sources

Primary data: Experiment

Primary data: Survey (Questionnaire, Interview)





Data Sources

Secondary data: existing databases, record review

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 $300

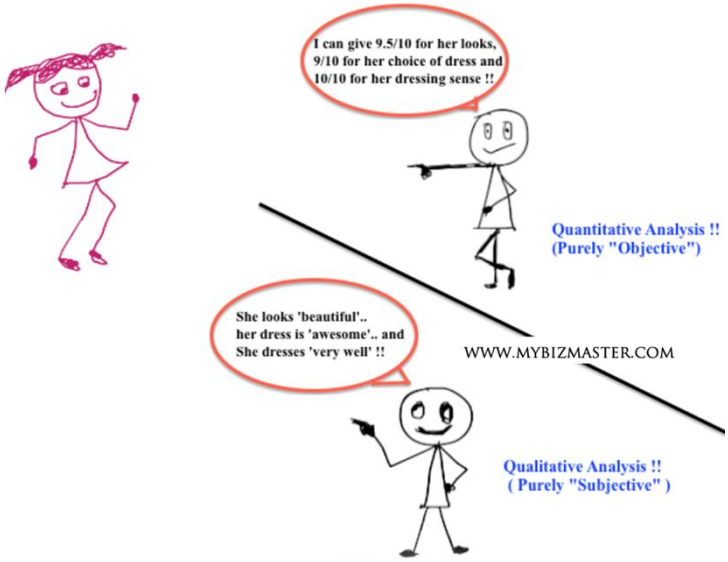
Qualitative & Quantitative Data

Example 2:

- ***Latte***

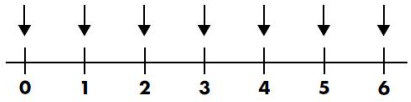
Qualitative data:	Quantitative data:
• robust aroma • frothy appearance • strong taste • burgundy cup	• 12 ounces of latte • serving temperature 150º F. • serving cup 7 inches in height • cost $4.95

Qualitative & Quantitative Data

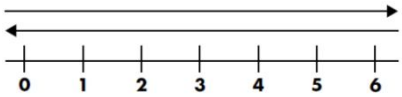


Discrete & Continuous Data

Discrete data can only take on certain individual values.



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



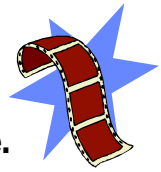
Discrete & Continuous Data

Example 1

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

Example 2

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



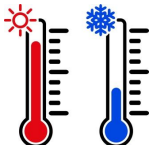
Example 3

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

Discrete & Continuous Data

Example 4

Temperature is a **continuous variable.**



Example 5

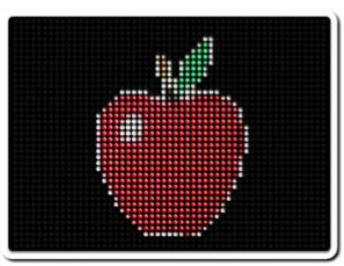
Number of people in a gathering is a **discrete variable.**

Example 6

Time taken to run a race is a **continuous variable.**

Discrete & Continuous Data

Discrete vs. Continuous

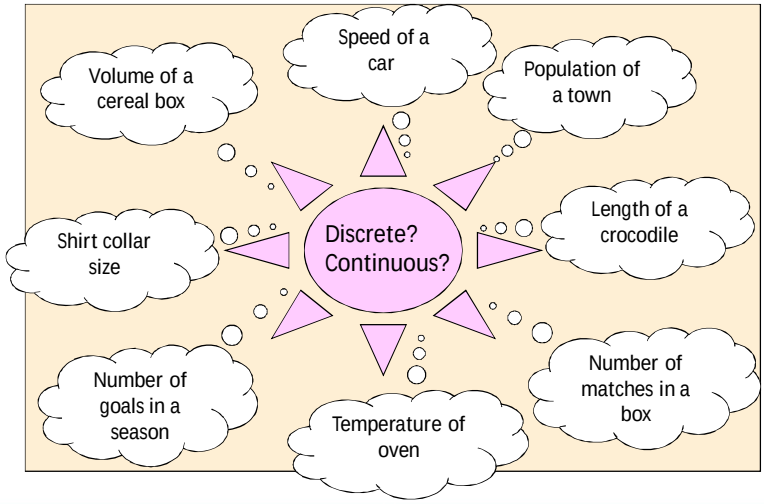


LITE BRITE
discrete

ETCH-A-SKETCH
continuous

Discrete & Continuous Data

Group the following as either discrete or continuous data.



Discrete & Continuous Data

Discrete

- Population of a town
- Number of matches in a box
- Shirt collar size
- Number of goals in a season

Continuous

- Volume of a cereal box
- Top speed of a car
- Length of a crocodile
- Temperature of oven

Discrete & Continuous Data

- How about money? Is it a discrete or continuous data?

Discrete!!
But why?

Levels of Measurement

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

Levels of Measurement

- Each level of measurement is characterized by its properties.
  - Nominal 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.

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

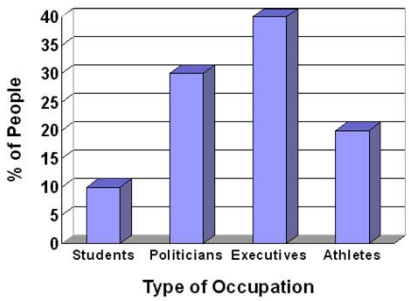
- Numbers assigned to categories (as identification codes) have no numeric value (we cannot add, subtract, divide or multiply nominal data) and any ordering of categories is arbitrary.
- This is the most primitive form of measurement. The presence vs. absence of something is a form of nominal measurement ("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

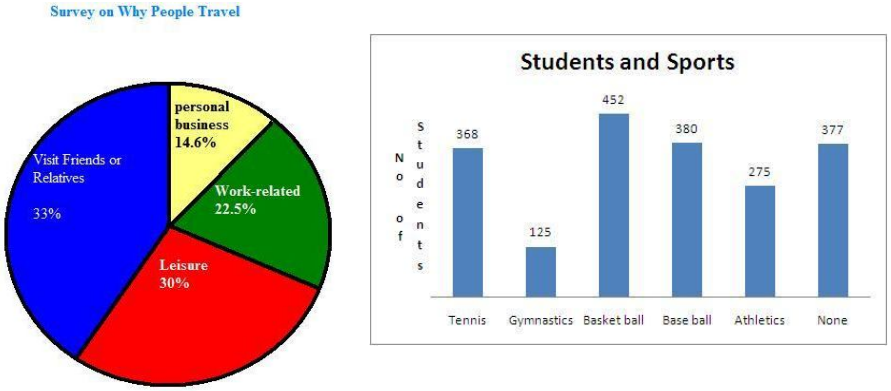
Example

Nominal Data

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



Levels of Measurement - Nominal Scales



Levels of Measurement - Nominal Scales

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

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

- Similarly if Jozan is ranked 1st for funniness, and Nabil is ranked 4th we have no way of knowing how much funnier Jozan is than Nabil.
- We cannot assume that they are four times funnier.
- They may be more or less than four times funnier.
- But we do know that they are more funny than Nabil, and more funny than the comedians ranked 2nd and 3rd places as well.
- We know about order but we have no information about the size of the interval between points.

Levels of Measurement -Ordinal Scales

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

Levels of Measurement - Ordinal Scales

Example

An ordinal data example

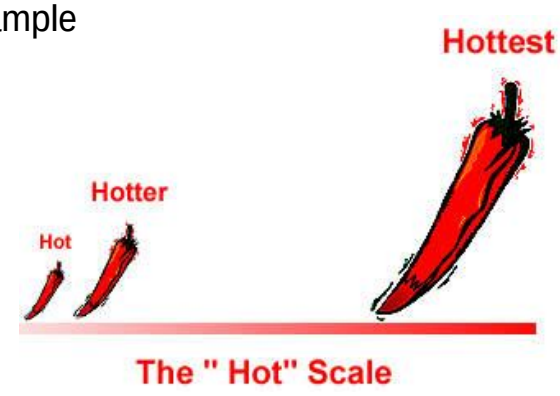
How often do you eat cheese for breakfast?

		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 - Ordinal Scales

Example



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.
- 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.

Levels of Measurement - Interval Scales

- However, ratios are meaningless on an interval scale because an interval scale has no true zero.
- Temperature scales are an example of this, so are decibel scales.
- Zero degrees Fahrenheit does not mean the total absence of temperature.
- 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.
- For ratio data both differences and ratios are interpretable.
- Ratio data have a natural zero.

Levels of Measurement - Ratio Scales

- Examples of ratio scale data are 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

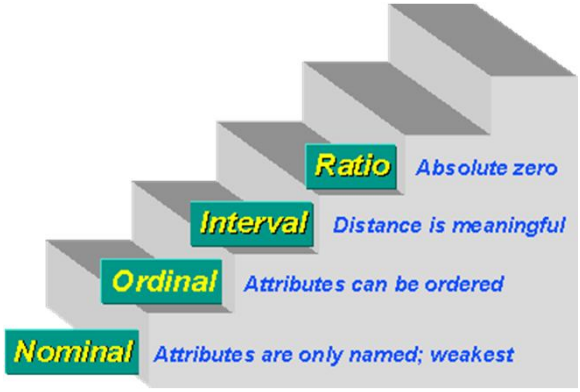
Level of Measurement	Properties	Examples	Descriptive statistics	Graphs
Nominal / Categorical	Discrete Arbitrary (no order)	Dichotomous • Yes / No • Gender Types / Categories • colour • shape	Frequencies Percentage Mode	Bar Pie
Ordinal / Rank	Ordered categories Ranks	Ranking of favourites Academic grades	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)

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

Example

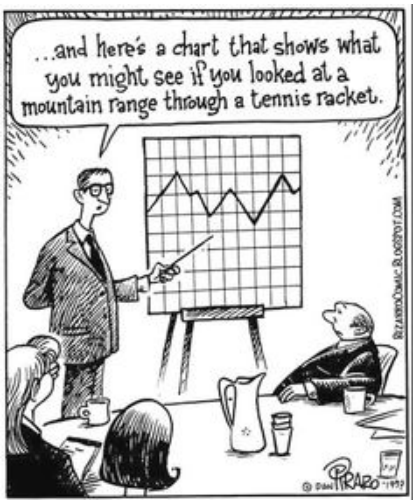
Scales to classify different measurements

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

Exercise

Identify the following as nominal level, ordinal level, interval level, or ratio level data.

1. Flavours of frozen yogurt
2. Amount of money in savings 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


...and here's a chart that shows what you might see if you looked at a mountain range through a tennis racket.