KNOW & UNDERSTAND THE KEY DEFINITION

BUSINESS STATISTICS vs DATA ANALYTICS vs DATA SCIENCE

- Business Statistics (BS) is a collection of statistical procedures and techniques that are used to convert data into meaningful information in a business environment

- Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.

- Data Science (DS) employs techniques and theories drawn from many fields within the broad areas in order to understand and analyze actual phenomena" with data.

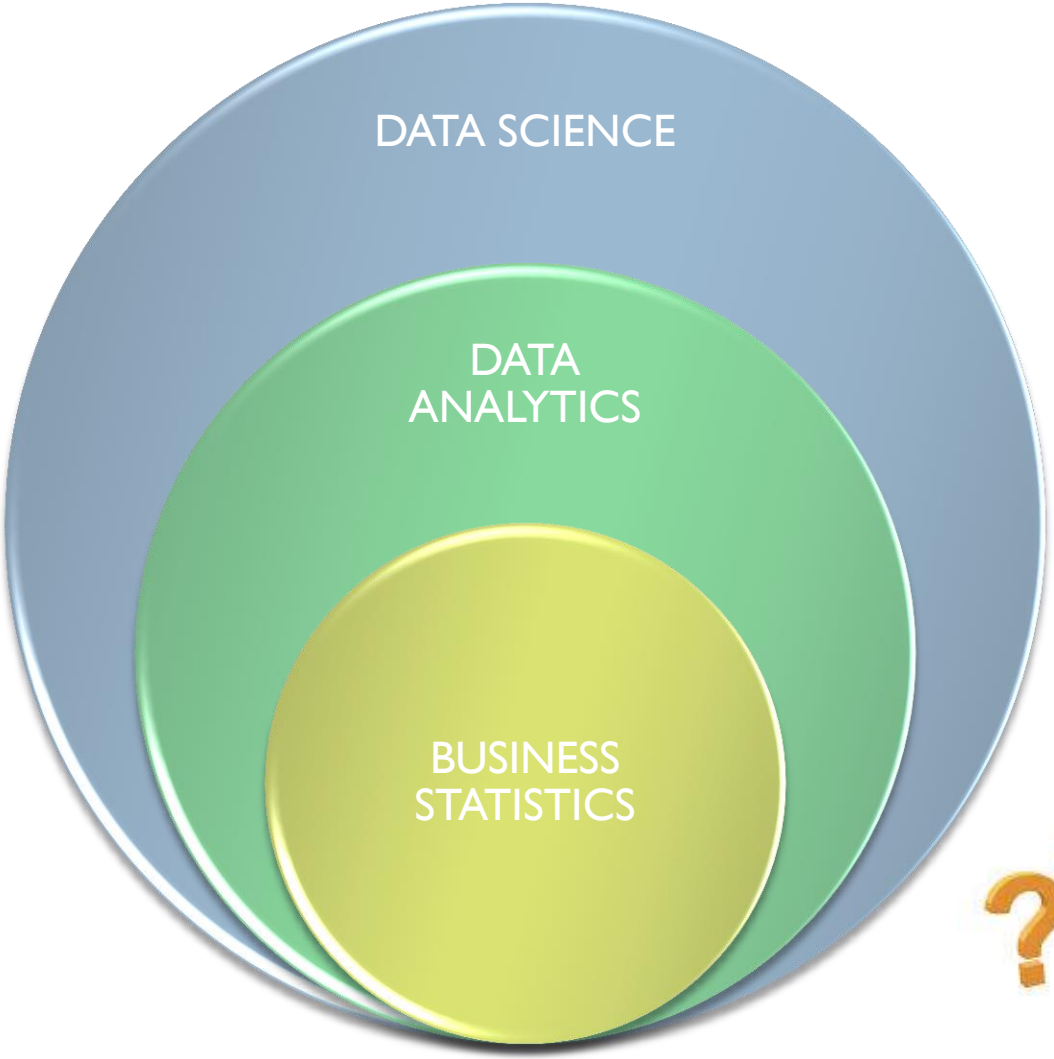
HOW THEY ARE CONNECTED TO EACH OTHER?

Data Science (DS)

Data Analytics (DA)

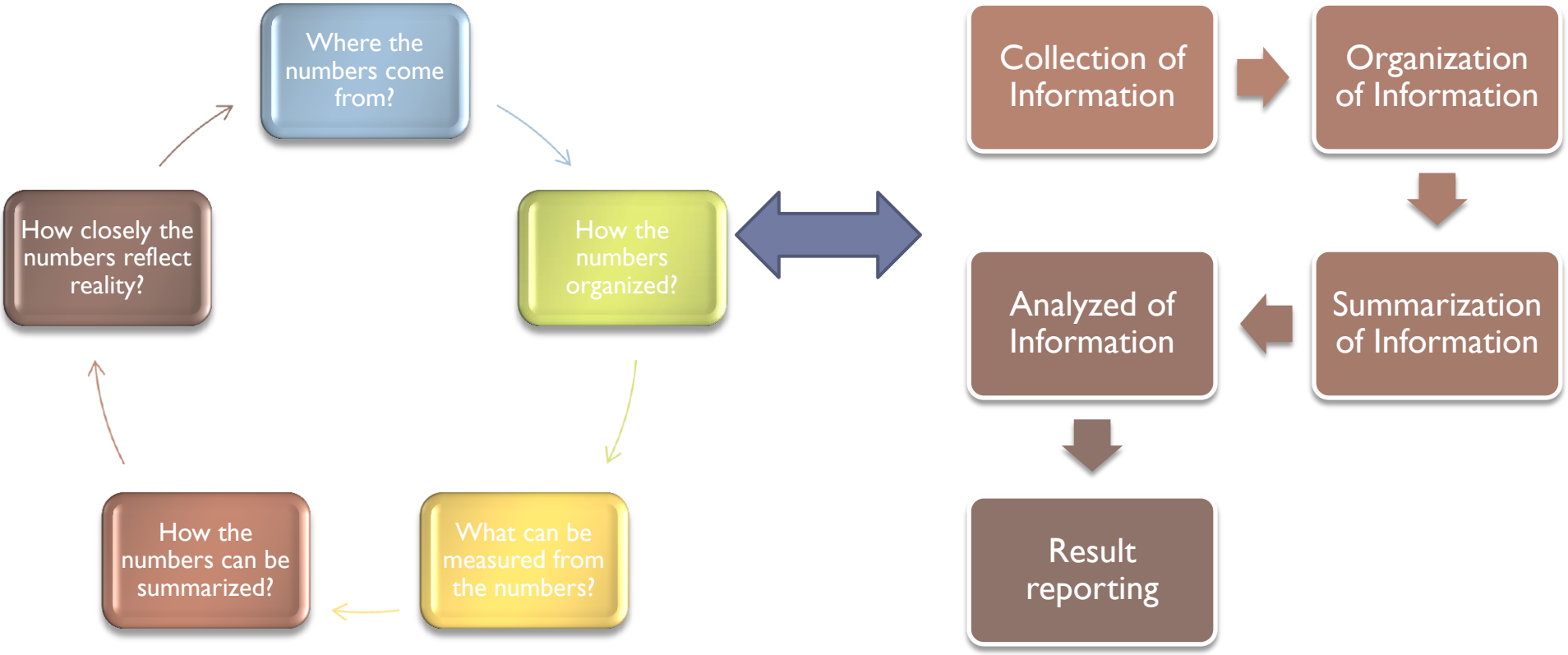
Business Statistics (BS)

BUSINESS STATISTICS vs DATA ANALYTICS vs DATA SCIENCE



STATISTICS: WHAT IT IS?

- Numbers: Only numbers? ➔ partially correct!!



STATISTICS: WHAT IT IS?

- Information in statistics refer to → Data
- Data:
  - describe characteristics of individual (in the study)
  - Characteristics:
    - Numerical: height
    - Non-Numerical: gender
  - varies (same weight? eat same amount of food everyday?)
    - variability in data may help to explain different result obtained
- Misused of data occurs when the data are incorrectly obtained (where the data comes from) or analyzed (does it is correctly/reliable)

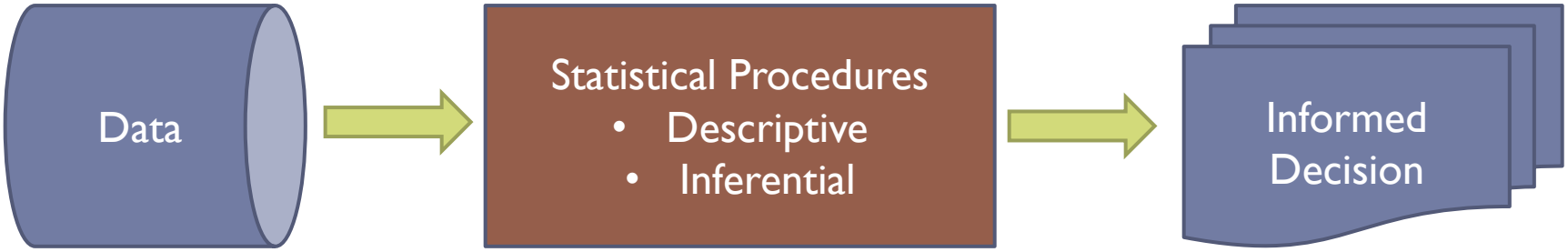
STATISTICS: WHAT IT IS?

- Mathematics versus Statistics?
  - Mathematics:
    - When the problem is solved correctly, the result can be reported as 100% certainty
  - Statistics:
    - When the problem is solved, the result do not have 100% certainty.
    - We might say, for example, we are 95% confidence that the average…..
- Example
  - Mathematic problem: Mary and Jane is asked to solve the value of x given $3x+5=11$
  - Statistic problem: Mary and Jane is asked to estimate the average commute time for workers in Dallas and Texas.

BUSINESS STATISTICS: WHY?

- Know about the business
  - Opening a business without assessing the need for it may affect its success
- Decision Making
  - allows business managers to analyze the data and arrive at meaningful conclusions
  - allows businesses to deal with the uncertainties of the business.
  - helps businesses to plan better and make predictions about the road ahead
- Competitors
  - allow a business to keep an eye on the competition.
  - plan the product(production) according to what the customer likes and wants

DESCRIPTIVE & INFERENTIAL



Descriptive Statistics:

- Consists of organizing and summarizing data.
- Describe about the data either through visual tools (chart or graph) and/or numerical measures.

Inferential Statistics:

- Uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.
- Making estimation & hypothesis testing to help in decision making process
- Include a level of a confidence in the result since sample cannot tells everything about a population.

PARAMETER VERSUS STATISTIC

PARAMETER

- Numerical summary of a population
- Example:

  Suppose 48.2% of all students in UTM KL campus own a car.

  - it is a numerical summary of a population. Why?

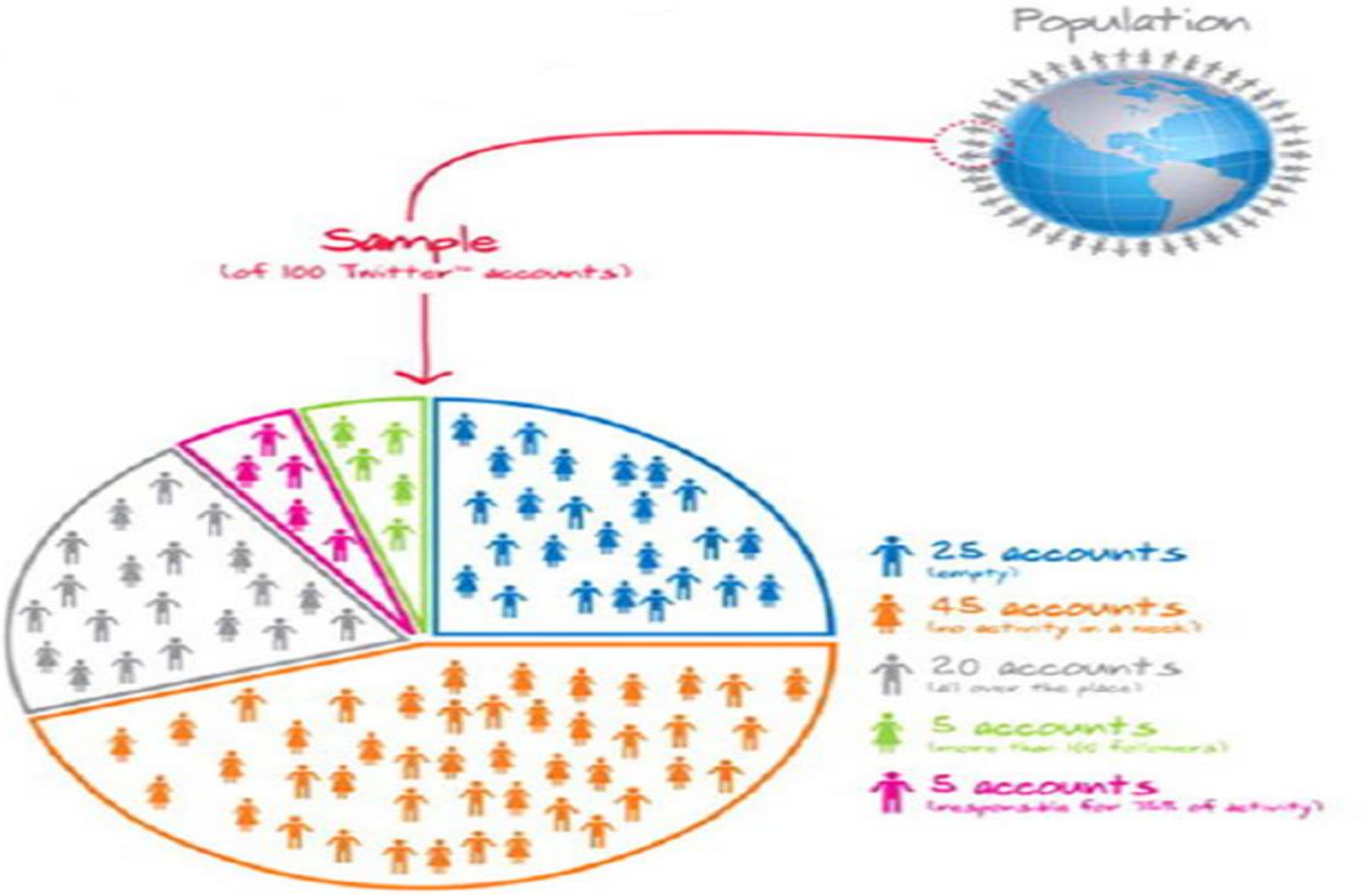
STATISTIC

- Numerical summary of a sample.
- Example:

  Suppose a sample of 100 students is obtained, and from this sample, we find that 46% own a car.

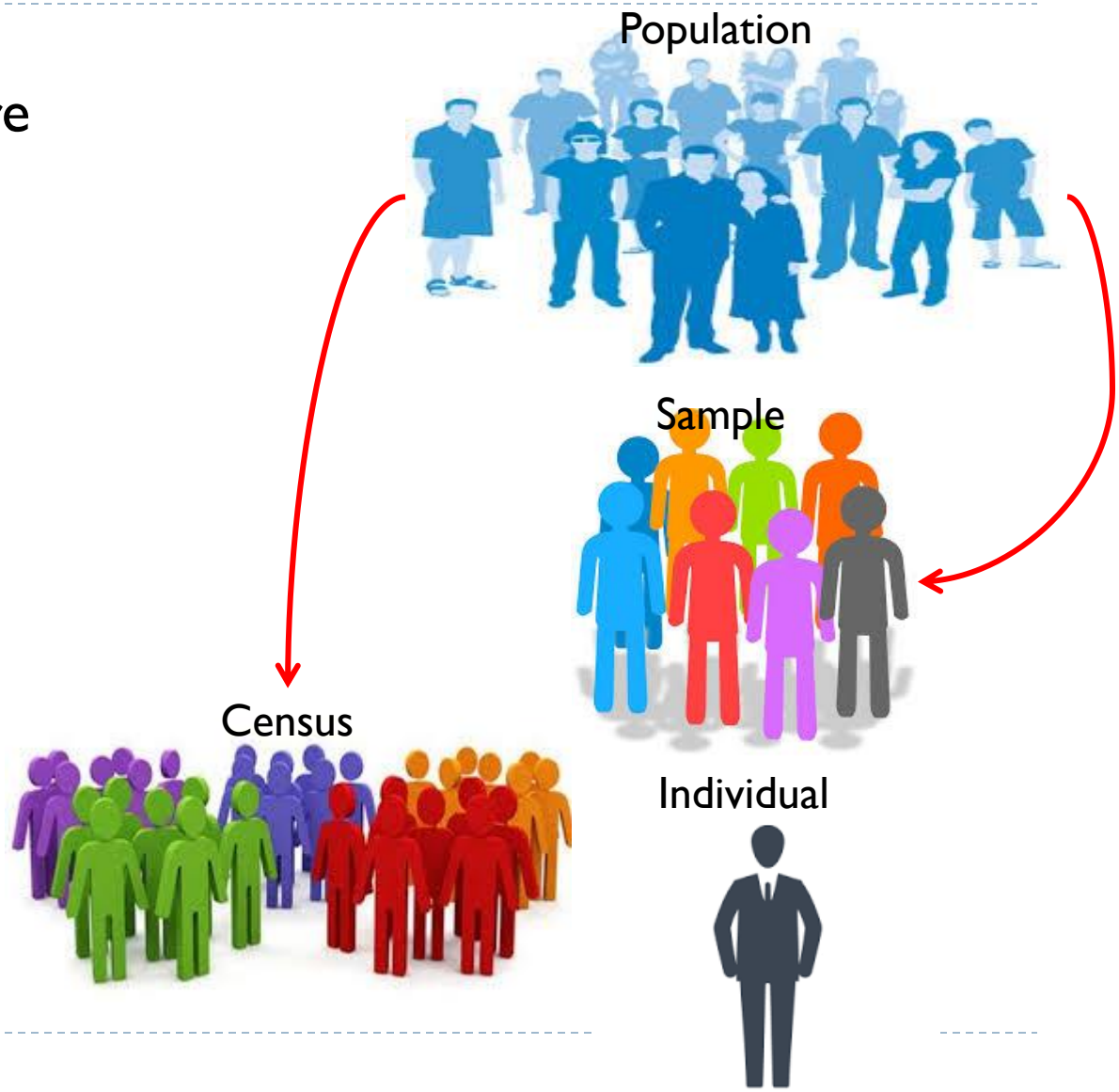
  - This value represent statistic. Why?

POPULATION & SAMPLING



POPULATION & SAMPLE

- Population – the entire group to be studied
- Individual – a person or object that is a member of the population being studied.
- Sample – a subset of the population that is being studied.
- Census – a list of all individuals in a population along with the characteristics of each individual



THE PROCESS OF STATISTIC

1. Identify the research objectives.
   - Determine the question (s) to answer
   - The question (s) must clearly identify the population that is to be studied.
2. Collect the data needed to answer the question (s) posed in (1).
   - Look for sample because population will be difficult and expensive.
3. Describe the data.
   - Descriptive allow to describe the data.
4. Perform Inference
   - Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results.

DISTINGUSHING BETWEEN DATA & VARIABLE

DATA VERSUS VARIABLE

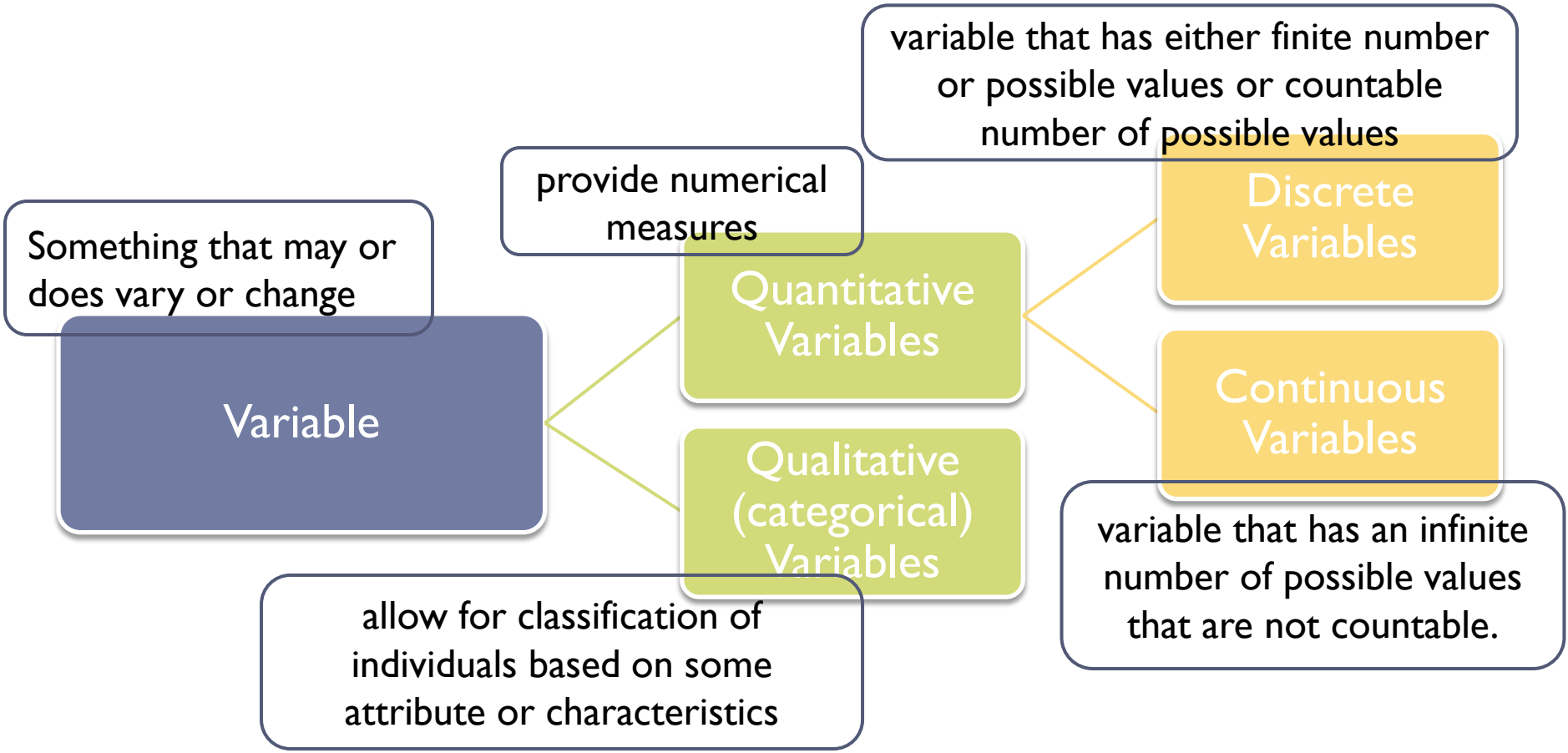
VARIABLE

- characteristics of the individuals within the population.
- Example: Recently my son and I planted tomato plant in our backyard. We collected information about the tomatoes harvested from the plant.
  - Individuals in the study : Tomato
  - Variables: Weight of the tomatoes

DATA

- The list of the observed value for a variables.
- Example: Gender is a variable, the observations male and female are data.

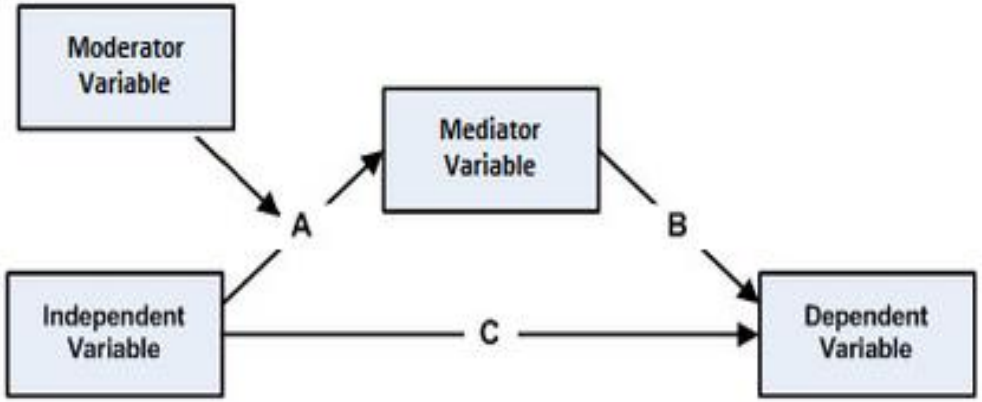
DATA VERSUS VARIABLE



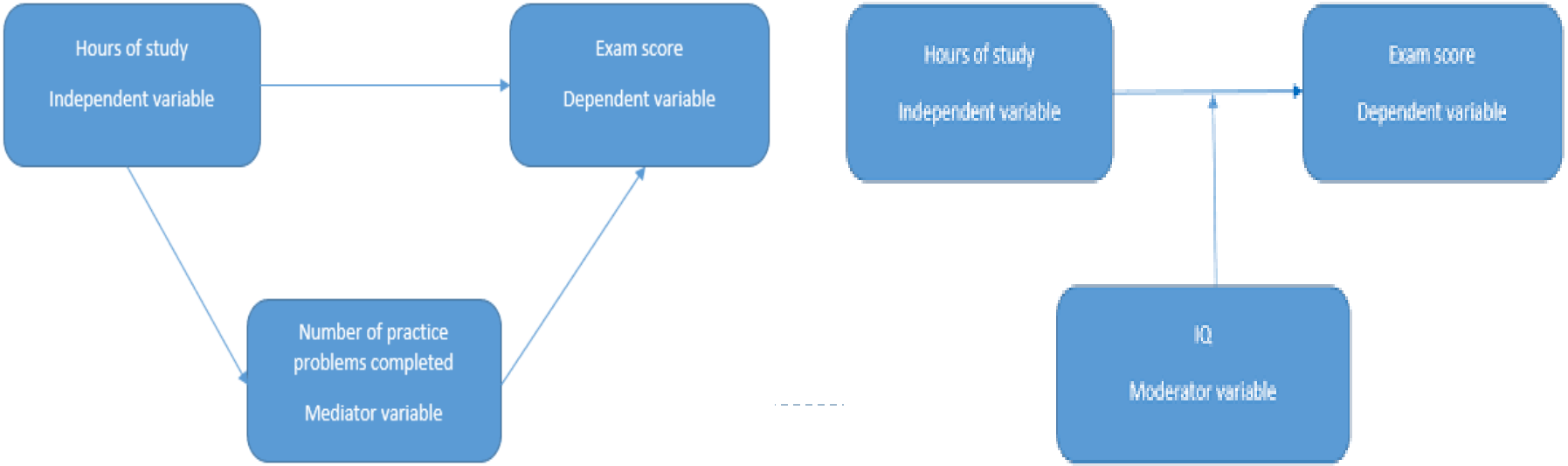
VARIABLES

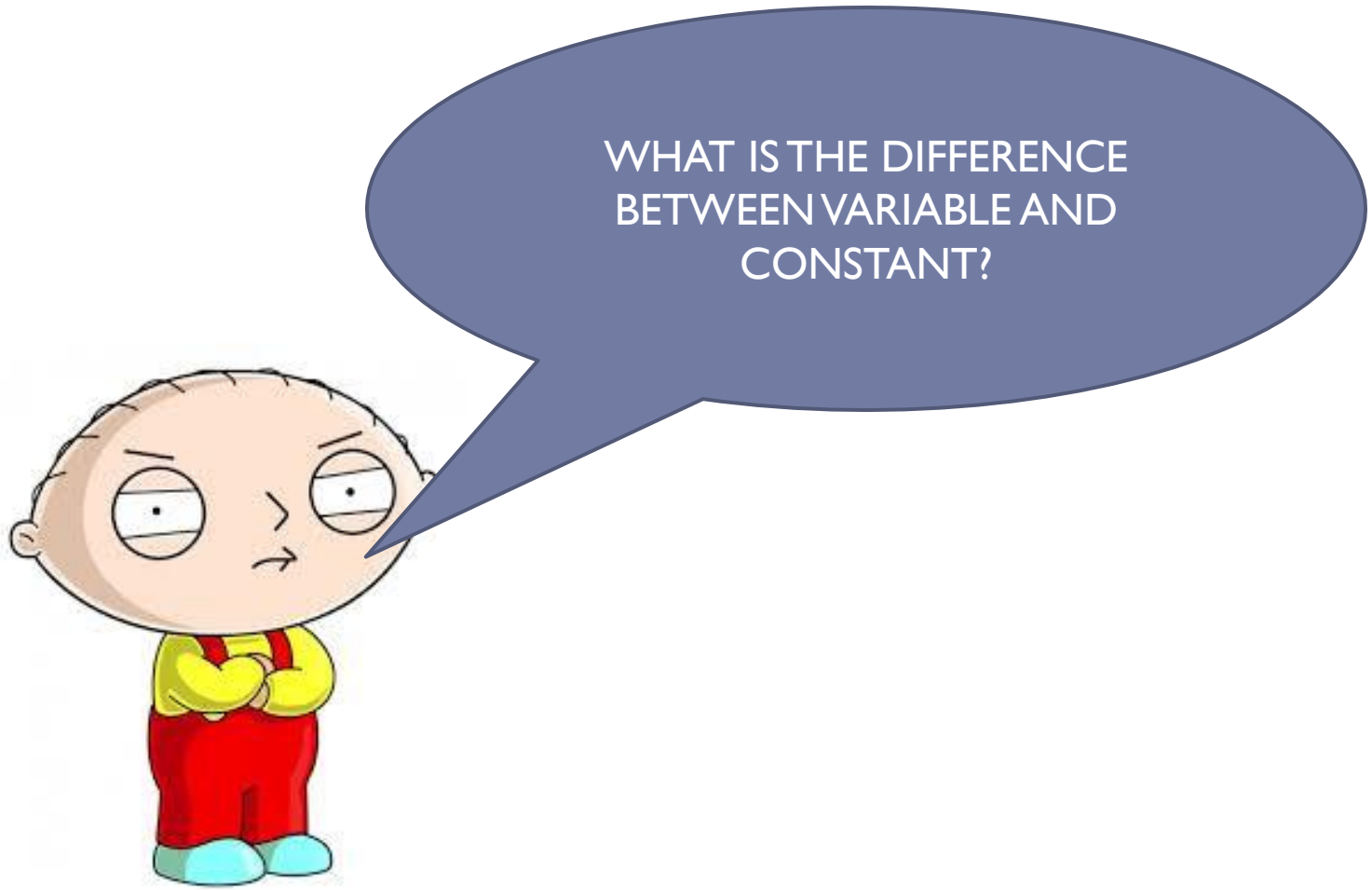
- Independent variable – represent inputs or causes
- Dependent variable – the output or outcome whose variation is being studied
- Moderator (intervening) variable – influenced by the independent variable, which in turn influences the dependent variable
- Mediator variable – influence the nature and strength of the relationship (dependent and independent). May reduce or increase cause-and-effect between variables

MODERATOR AND MEDIATOR VARIABLES



EXAMPLE



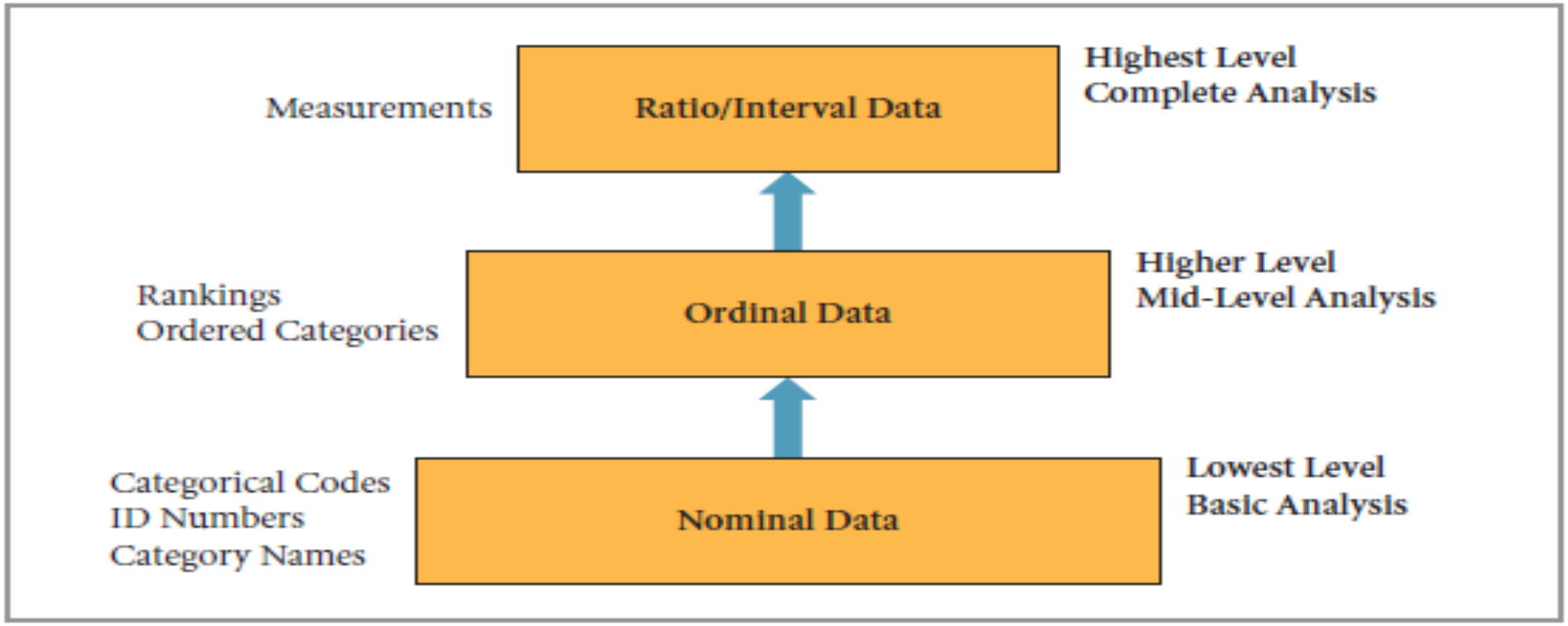

WHAT IS THE DIFFERENCE BETWEEN VARIABLE AND CONSTANT?

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**

- Quantitative Data – data whose measurement scale are inherently numerical. Example: percentage, inches, accuracy…
- 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]
- Time-series Data – Set of consecutive data values observed at successive points in time.
- Cross-sectional Data – Set of data values observed at a fixed point in time.

DATA MEASUREMENT LEVEL (data hierarchy)



DATA MEASUREMENT LEVEL (data hierarchy)

- Nominal Data
  - Lowest form of data
  - Assigning code to categorize data
  - Example: 88. Single 11. Divorces 33. Married 55. Other
  - Nominal data can be compared only for equality. Cannot order nominal measurement

- Ordinal Data or Rank Data
  - one notch above nominal data
  - data elements can be rank-ordered on the basis of some relationship among them, with the assigned values indicating this order.
  - Allows decision makers to equate two or more observations or to rank-order the observations.
  - Example: Bank loan applicants are asked to indicate the category corresponding to their household incomes.

    1. Under RM10,000 2. RM10,000 - RM25,000

    3. Above RM25,000

DATA MEASUREMENT LEVEL (data hierarchy)

- Interval Data
  - The distance between two data items can be measured on some scale and the data have ordinal properties (<,>,=).
  - Example: Temperature scale (Celsius/ Fahrenheit)
  - Allow us to precisely measure the difference between any two values (possible with ordinal data because all we can say is that one value is larger than another)

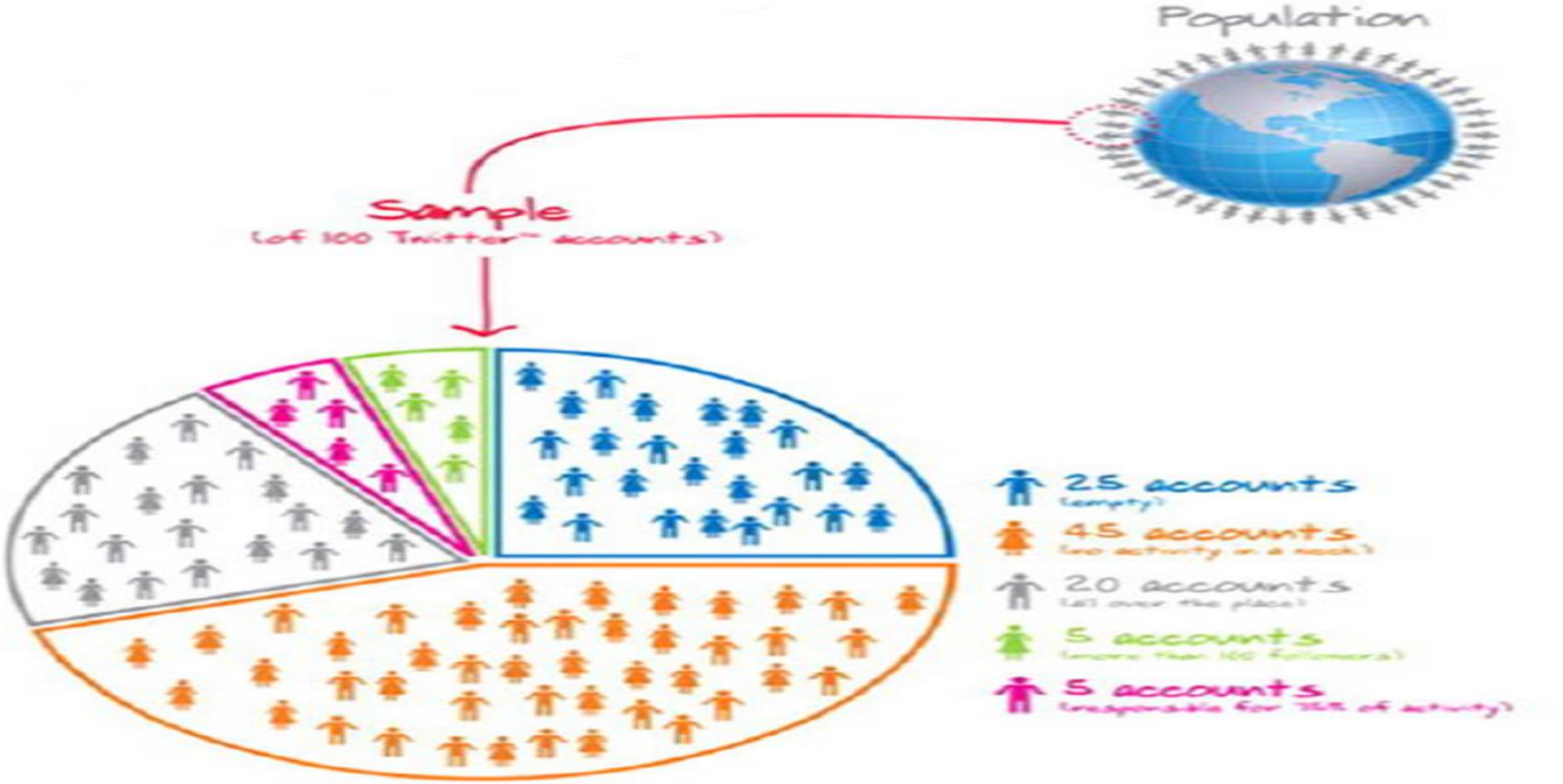
- Ratio Data
  - have all the characteristics of interval data but also have a true zero point (at which zero means "none").
  - Example: Packagers of frozen foods encounter ratio measures when they pack their products by weight. Weight = 0kg indicate the pack does not have weight.

TIPS: CATEGORIZING DATA

1. Identify each factor in the data set. (Look on the data title)
2. Determine whether the data are time-series or cross-sectional
3. Determine which factors are quantitative data and whish are qualitative data
4. Determine the level of data measurement for each factor

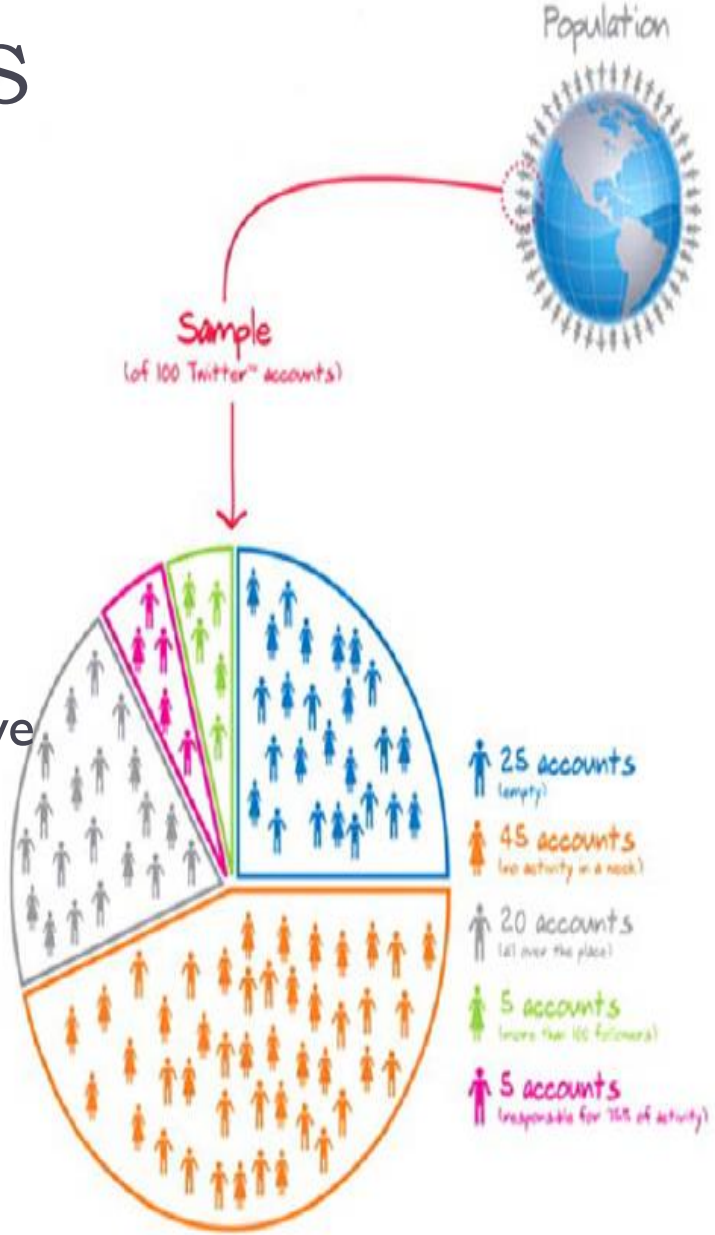
POPULATION & SAMPLING

- NON STATISTICAL SAMPLING
- STATISTICAL SAMPLING



SAMPLING TECHNIQUES

- Why sampling? – to gather information (data)
- Categories: (commonly use by decision makers)
  - Statistical sampling techniques
  - Non Statistical sampling techniques
- Non statistical sampling:
  - samples are selected based on the subjective judgement of the researcher, rather than random selection (i.e., probabilistic **methods**)
  - Approach:
    - Convenience sampling
    - Judgmental sampling
    - Ratio/Quota sampling
    - Snowball sampling



NON STATISTICAL SAMPLING

- Convenience Sampling
  - Select the items from the population based on accessibility and ease of selection
  - Example of business application:

    " Sun Citrus orchards owns and operates a large orange orchard and fruit picking plant in Australia. During harvest time in the orange grove, pickers load 20 kilos sacks of oranges, which are then transported to the packing plant. At the packing plant the orange are graded and boxed for shipping internationally or locally. Because of the volume of orange involved, it is impossible to assign a quality grade to each of the individual orange. Instead, as each sack moves up the conveyor into a packing plant, a quality manager selects an orange sacks every so often, grades the individual oranges in the sacks as to size, color and etc, and then assigns an overall quality grade to the entire shipment from which the sample was selected. The quality manager is wiling to assume that orange quality (size, color…) is evenly spread throughout the many sacks of oranges in the shipment. That is the oranges in the sacks selected are the same quality as those that were not inspected."

STATISTICAL SAMPLING

- Often known as – probability sampling
  - Allow every item in the population to have a known or calculable chance of being included in the sample.

- Measurement – use of the probability theory to evaluate sample results, including measurement of sampling risk

- Type of approaches:
  - Simple Random Sampling
  - Stratified Random Sampling
  - Systematic Sampling
  - Cluster Sampling

How it is implemented through business application ?

SIMPLE RANDOM SAMPLING

- the most basic sample survey design - each member of the subset has possible an equal probability of being chosen.
- Example of business application:

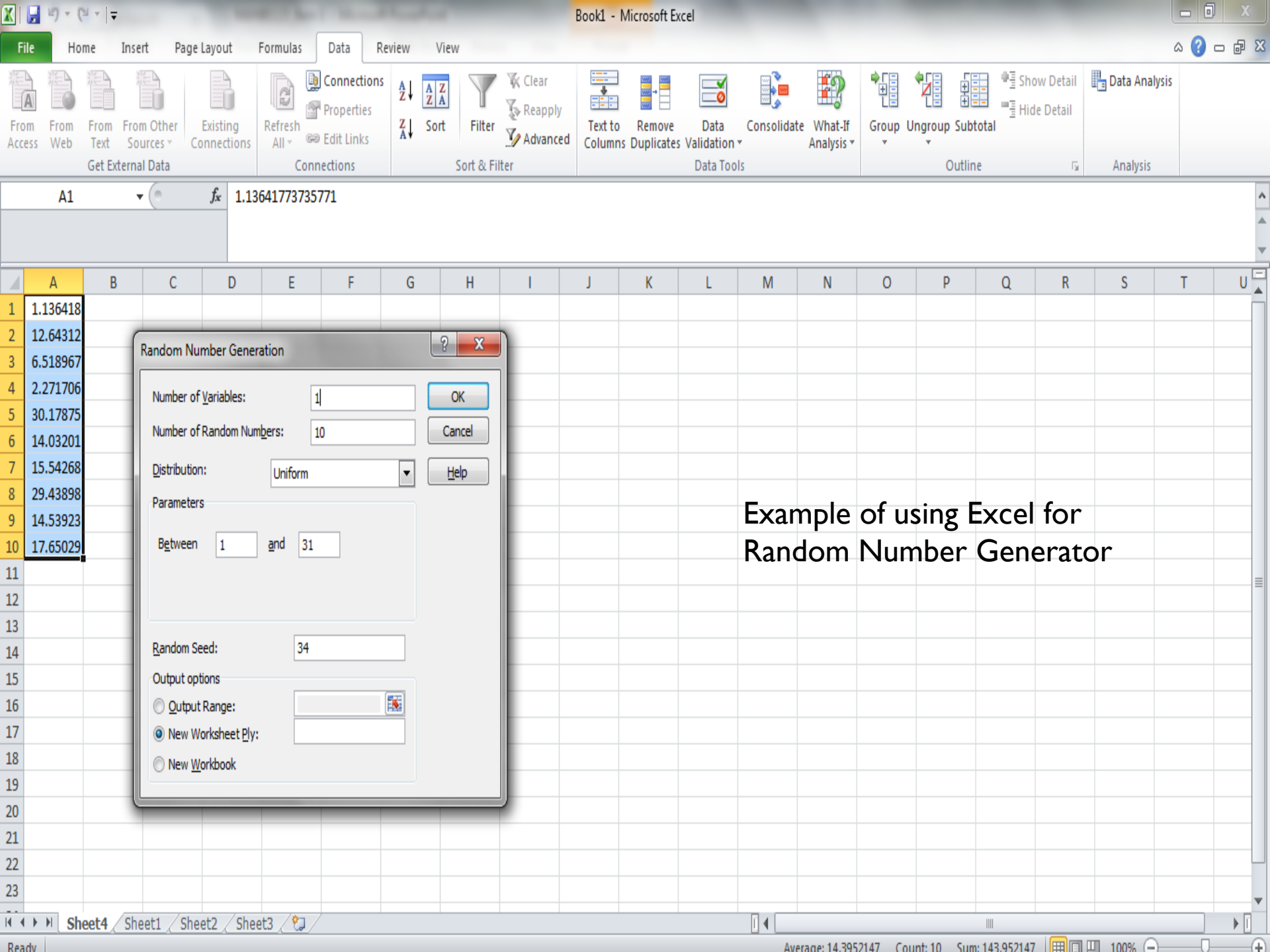
  " Social services – Director of Midwestern state's social services is considering changing the timing on food stamp distribution from once a month to once every two week. Before making any decision he wants to survey a sample of 100 citizens who are on food stamp from the 300 total food stamp recipients in that country."

  How?

  - assign recipients a number (001 to 300)
  - use the random generator number to determine which 100 recipients to include in the sample.

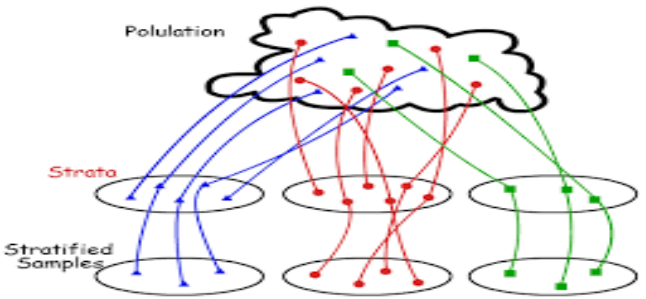
SIMPLE RANDOM SAMPLING

- Subtype of simple random sampling:
  - Sampling with replacement - after an element has been selected from the sampling frame, it is returned to the frame and is eligible to be selected again.
  - ** Sampling without replacement - after an element is selected from the sampling frame, it is removed from the population and is not returned to the sampling frame.


Book1 - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Get External Data
Connections
Sort & Filter
Data Tools
Outline
Analysis
A1
1.13641773735771
1.136418
12.64312
6.518967
2.271706
30.17875
14.03201
15.54268
29.43898
14.53923
17.65029
Random Number Generation
Number of Variables:
1
Number of Random Numbers:
10
Distribution:
Uniform
Parameters
Between
1
and
31
Random Seed:
34
Output options
Output Range:
New Worksheet Ply:
New Workbook
OK
Cancel
Help
Example of using Excel for
Random Number Generator
Sheet4
Sheet1
Sheet2
Sheet3
Ready

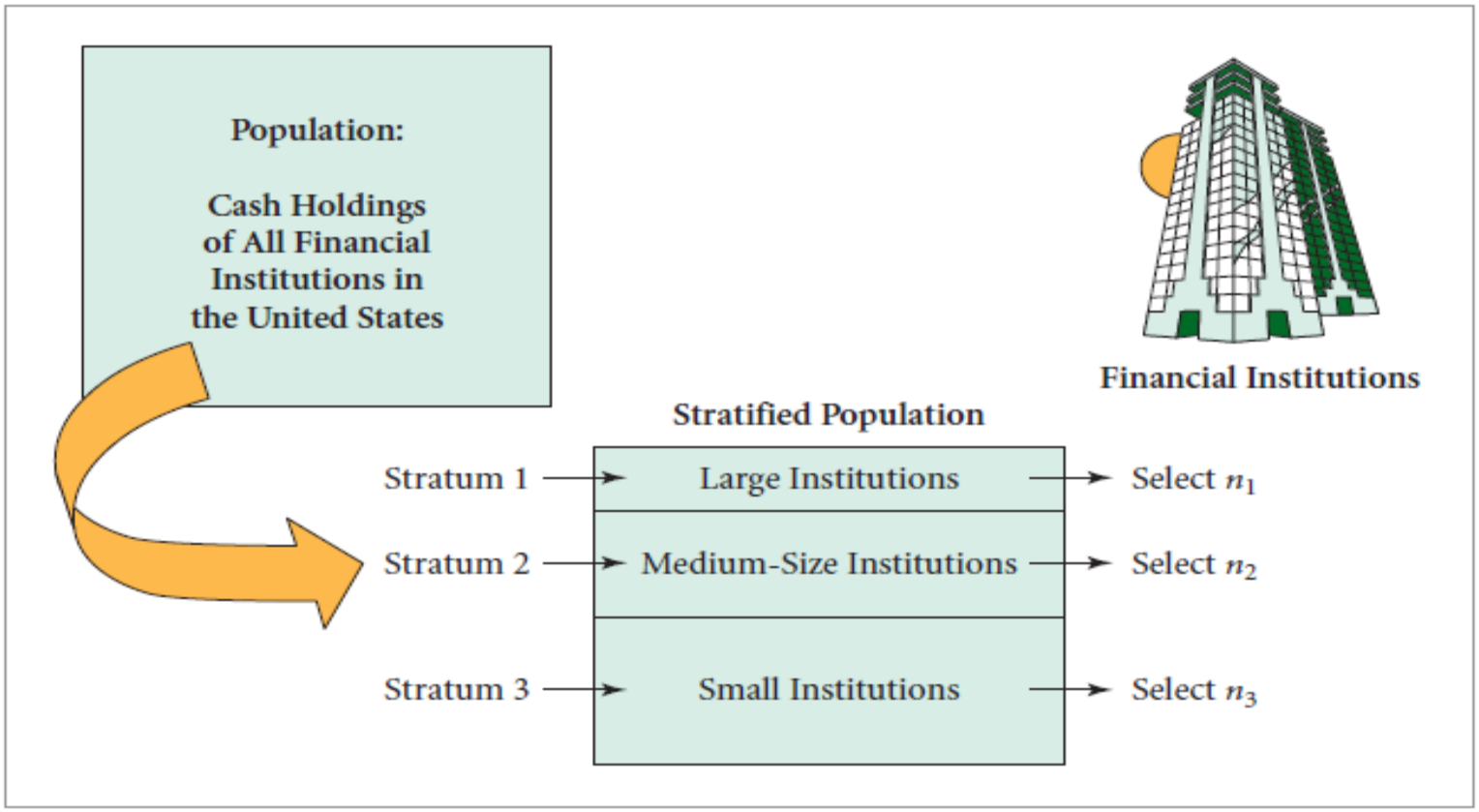
STRATIFIED RANDOM SAMPLING

- Is obtained by separating the population into no overlapping groups called strata (segments) and then obtaining a simple random sample from each stratum.
  - The individuals within each stratum should be homogeneous
- Provides more information about the population for less cost than simple random sampling.
  - because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next.



STRATIFIED RANDOM SAMPLING

Business Application Example: FINANCIAL INSTITUTION



SYSTEMATIC RANDOM SAMPLING

- Sampling procedure that involves selecting every *k*th item in the population after a randomly selected starting point between 1 and *k*.

  *k*th interval = population size, *N* / desired sample size, *n*

- Also known as Interval Random Sampling

Business application Example:

"National Association of Accountants (NAA) considered establishing a code of ethics. To determine the opinion of its 20,000 members, a questionnaire was sent to a sample of 500 members."

SYSTEMATIC RANDOM SAMPLING

How?

- Calculate kth interval: 20,000/500 = 40 → the questionnaire will be send to every 40th member from the list of members.
- generate a single random number in the range 1 to 40. Let say = 25, thus the 25th person in the alphabetic list would be selected.
- After that, every 40th member would be selected (25, 65, 105, 145, . . .) until there were 500 NAA members.

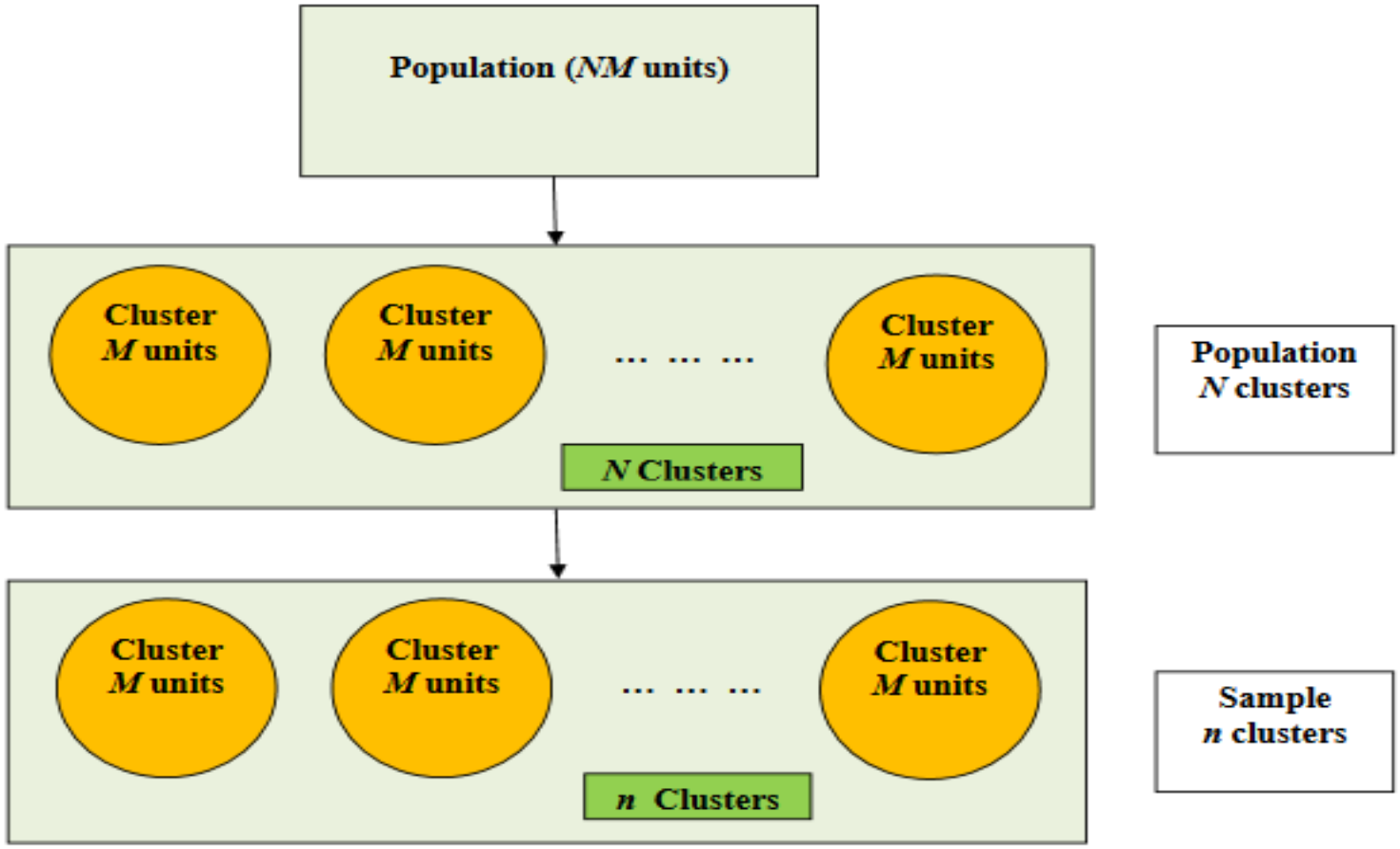
CLUSTER SAMPLING

- Sampling technique by which the population is divided into groups, or clusters, that are each intended to be mini-populations. A simple random sample of *m* clusters is selected. The items chosen from a cluster can be selected using any probability sampling technique.

- Subtype of cluster sampling:
  - Single cluster sampling - If all elements in each sampled cluster are sampled
  - Two-stage cluster sampling - Simple random subsample of elements is selected within each of these groups
  - Multi-stage cluster sampling - more than two steps are taken in selecting clusters from clusters

a random sample of cluster is drawn and then elements are randomly selected from the selected clusters

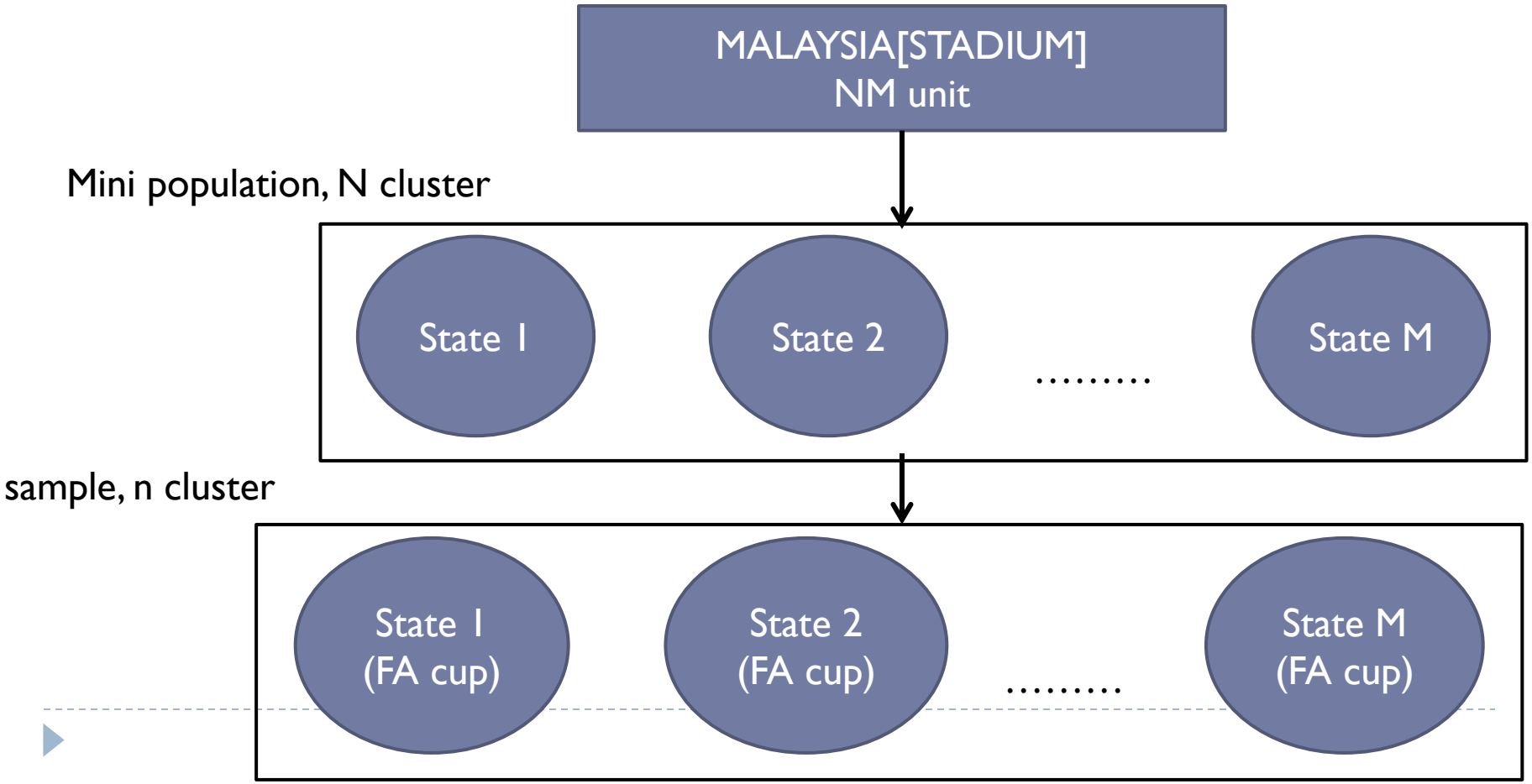
CLUSTER SAMPLING



CLUSTER SAMPLING

- Business Application Example:

"The potential for changing how season ticket pricing is structure." How?: Questionnaire to season ticket holder



CLUSTER SAMPLING

- Sub type of Cluster Sampling:
  - Single stage cluster - occurs when the researcher includes all the FA Cup season from all the randomly selected clusters as sample.
  - Two stage cluster - is obtained when the researcher only selects a number of FA Cup season from each cluster by using simple or systematic random sampling
  - Multi stage cluster – occurs when involved more than two steps in selecting clusters from clusters

- Season ticket holder strategy/technique:
  - What if (without cluster sampling):
    1. Simple Random sampling – season ticket holder of size, n from the whole population, N (Malaysia of season ticket holder) – cost high because require to sample all section
    2. Stratified or Systematic Sampling – require to visit each section

CLUSTER vs STRATIFIED SAMPLING

- Similarity:
  - Cluster sampling is similar to stratified sampling in that both involve separating the population into categories and then sampling within the categories
  - Both sampling procedures permit analysis of individual categories (strata or clusters) in addition to analysis of the total sample.

- Differences:
  - Stratified sampling: Once the categories (strata) are created, a random sample is drawn from each category (stratum). All the strata of the population is sampled.
  - Cluster sampling involve with 3 types of stages (one stage, two stage & multi-stage).

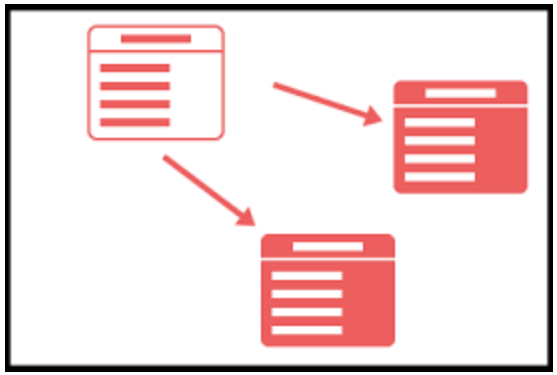
DESCRIPTIVE STATISTICS

- VISUAL WAY
- NUMERICAL WAY

DESCRIBING DATA: WHY?

- Business climate – companies collect massive amount of data → hope will be useful for making decision

- To get information/decision making from the data captured: Business communicate → customers, employees, suppliers and other constituents



- Better understanding the key concepts in an informative way

DESCRIBING YOUR DATA

VISUAL WAY:

TABLES

Data set: product categories per customer at the abx mart

Purpose of Study: Market Basket Analysis of item purchased at ABX Mart

Problem: no information that managers could use to determine the buying habits of their customers.

How to solve?: Frequency and Relative Frequency Distribution

4	2	5	8	8	10	1	4	8	3	4	1	1	3	4
1	4	4	5	4	4	4	9	5	4	4	10	7	11	4
10	2	6	7	10	5	4	6	4	6	2	3	2	4	5
5	4	11	1	4	1	9	2	4	6	6	7	6	2	3
6	5	3	4	5	6	5	3	10	6	5	7	7	4	3
8	2	2	6	5	11	9	9	5	5	6	5	3	1	7
6	6	5	3	8	4	3	3	4	4	4	7	6	4	9
1	6	5	5	4	4	7	5	6	6	9	5	6	10	4
7	5	8	4	4	7	4	6	6	4	4	2	10	4	5
4	11	8	7	9	5	6	4	2	8	4	2	6	6	6
6	4	6	5	7	1	6	9	1	5	9	10	5	5	10
5	4	7	5	7	6	9	5	3	2	1	5	5	5	5
5	9	5	3	2	5	7	2	4	6	4	4	4	4	4
6	5	8	5	5	5	5	5	2	5	5	6	4	6	5
5	7	10	2	2	6	8	3	1	3	5	6	3	3	6
5	4	5	3	3	7	9	4	4	5	10	6	10	5	9
4	3	8	7	1	8	4	3	1	3	6	7	5	5	5
4	7	4	11	6	6	3	7	9	4	4	2	9	7	5
1	6	6	8	3	8	4	4	1	9	3	9	3	4	2
9	5	5	7	10	5	3	4	7	7	6	2	2	4	4
4	7	3	5	4	9	2	3	4	3	2	1	6	4	6
1	8	1	4	3	5	5	10	4	4	4	6	9	2	7
9	4	5	3	6	5	5	3	4	6	5	7	3	6	8
3	6	1	5	7	7	5	4	6	6	6	3	6	9	5
4	5	10	1	5	5	7	8	9	1	6	5	6	6	4
10	6	5	5	5	1	6	5	6	4	7	9	10	2	6
4	4	6	11	9	5	4	4	3	5	4	6	2	6	7
3	5	6	7	4	5	4	6	9	4	3	3	6	9	4
3	7	5	6	11	4	4	8	4	2	8	2	4	2	3
6	5	1	10	5	9	5	4	5	1	4	9	5	4	4

FREQUENCY AND RELATIVE FREQUENCY DISTRIBUTION

- Frequency distribution - A summary of a set of data that displays the number of observations in each of the distribution's distinct categories or classes.

Number of Product Catagories	Frequency
1	25
2	29
3	42
4	92
5	83
6	71
7	35
8	19
9	29
10	18
11	7
	Total = 450

Table: Market Basket Analysis of item purchased at ABX Mart

FREQUENCY AND RELATIVE FREQUENCY DISTRIBUTION

- Relative Frequency distribution - The proportion of total observations that are in a given category. Useful when the number in total distribution is different.

Relative Frequency

$$\text{Relative frequency} = \frac{f_i}{n}$$

where:

f_i = Frequency of the ith value of the discrete variable

$n = \sum_{i=1}^{k} f_i$ = Total number of observations

k = The number of different values for the discrete variable

The relative frequencies can be converted to percentages by multiplying by 100

Table: Frequency distribution of years of college

Dallas		Knoxville	
Years of College	Frequency	Years of College	Frequency
0	35	0	187
1	21	1	62
2	24	2	34
3	22	3	19
4	31	4	14
5	13	5	7
6	6	6	3
7	5	7	4
8	3	8	0
	Total = 160		Total = 330

FREQUENCY AND RELATIVE FREQUENCY DISTRIBUTION

Purpose: To compare the distribution for years of college for Dallas and Knoxville. How do the two cities' distributions compare

Years of College	Dallas		Knoxville	
	Frequency	Relative Frequency	Frequency	Relative Frequency
0	35	35/160 = 0.219	187	187/330 = 0.567
1	21	21/160 = 0.131	62	62/330 = 0.188
2	24	24/160 = 0.150	34	34/330 = 0.103
3	22	22/160 = 0.138	19	19/330 = 0.058
4	31	31/160 = 0.194	14	14/330 = 0.042
5	13	13/160 = 0.081	7	7/330 = 0.021
6	6	6/160 = 0.038	3	3/330 = 0.009
7	5	5/160 = 0.031	4	4/330 = 0.012
8	3	3/160 = 0.019	0	0/330 = 0.000
Total	160		330	

Table: Relative Frequency distribution of years of college

Summary: Knoxville has relatively more people without any college (56.7%) or with one year of college (18.8%) than Dallas (21.9% and 13.1%). At all other levels of education, Dallas has relatively more people than Knoxville.

FREQUENCY AND RELATIVE FREQUENCY DISTRIBUTION : LETS PRACTICE

PROBLEM: ATHLETIC SHOE SURVEY

" The status symbol for many students has been the brand and style of athletic shoes they wear. Companies like Nike and Adidas compete for the top position in the sport shoe market. A survey was recently conducted in which 100 college students were asked a number of questions, including how many pairs of Nike shoes they currently own."

What can you conclude from the analysis of the survey?

Data Set: Sportshoes

GROUPED DATA FREQUENCY DISTRIBUTION

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

GROUPED DATA FREQUENCY DISTRIBUTION

- HOW TO SOLVE?

1. Determine the desired number of classes or groups. (Rule of thumb is to use 5 to 20 classes. Or the $2^k \geq n$ rule can also be used.)
2. Determine the minimum class width using:
   $$w = \frac{largest\ value - smallest\ value}{number\ of\ classes}$$
   Round the class width up to a more convenient value.
3. Define the class boundaries.(Ideally, the classes should have equal widths and should all contain at least one observation)
4. Determine the class frequency for each class.

GROUPED DATA FREQUENCY DISTRIBUTION

- 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.

DESCRIBING THE DATA- VISUAL WAY

To what extent did you increase your skills in OOP?				
	A lot	**Some**	**A little**	**Not at all**
Students (N=30)	14	9	5	2

Frequency Distribution

To what extent did you increase your skills in putting together a household budget?				
	A lot	**Some**	**A little**	**Not at all**
Women (N=30)	46%	30%	17%	7%

Percentage Distribution

STATISTICAL PACKAGE: CROSS TABULATION

Cross-tabulation/ Contingency Table is one of the most useful analytical tools

- account for more than 90% of all research analyses.
- used to analyse categorical (nominal measurement scale) data.

A cross-tabulation is a two (or more) dimensional table that records the number (frequency) of respondents that have the specific characteristics described in the cells of the table.

CROSS TABULATIONS

Program Type	Area of Inquiry	Outcome
Web site	Employment law	Satisfied
I & R Line	Family law	Not satisf
Law clinic	Immigration	Pending
Web site	Immigration	Satisfied
I & R Line	Immigration	Satisfied
I & R Line	Family law	Not satisf
Web site	Employment law	Not satisf
Law clinic	Other	Satisfied
I & R Line	Other	Not satisf
I & R Line	Other	Satisfied
Law clinic	Employment law	Satisfied
Web site	Family law	Satisfied
Law clinic	Family law	Satisfied
Web site	Immigration	Not satisf
Law clinic	Immigration	Not satisf
I & R Line	Family law	Satisfied
I & R Line	Immigration	Not satisf
I & R Line	Employment law	Not satisf
Law clinic	Other	Pending

Count of Outcome	Outcome			
Program Type	Not satisfied	Pending	Satisfied	Grand Total
I & R Line	7		5	12
Law clinic	1	3	7	11
Web site	6		5	11
Grand Total	14	3	17	34

Count of Outcome	Outcome			
Program Type	Not satisfied	Pending	Satisfied	Grand Total
I & R Line	58%	0%	42%	100%
Law clinic	9%	27%	64%	100%
Web site	55%	0%	45%	100%
Grand Total	41.18%	8.82%	50.00%	100.00%

DESCRIBING YOUR DATA

VISUAL WAY –GRAPH, CHARTS

(HISTOGRAM, BAR CHART, PIE CHART, STEM AND LEAF DIAGRAM, LINE CHART AND SCATTER DIAGRAM)

HISTOGRAMS & OGIVE

- Information gained from Histogram:
  - provides a visual indication of where the approximate center of the data is.
  - can gain an understanding of the degree of spread (or variation) in the data.
    - The more the data cluster around the center, the smaller the variation in the data.
    - If the data are spread out from the center, the data exhibit greater variation.
  - Can observe the shape of the distribution. Is it flat? Or weighted to one side or the other? Or is it balanced around the center, or is it bell shape?

- Ogive = graph that represent relative cumulative frequency.

HISTOGRAMS & OGIVE

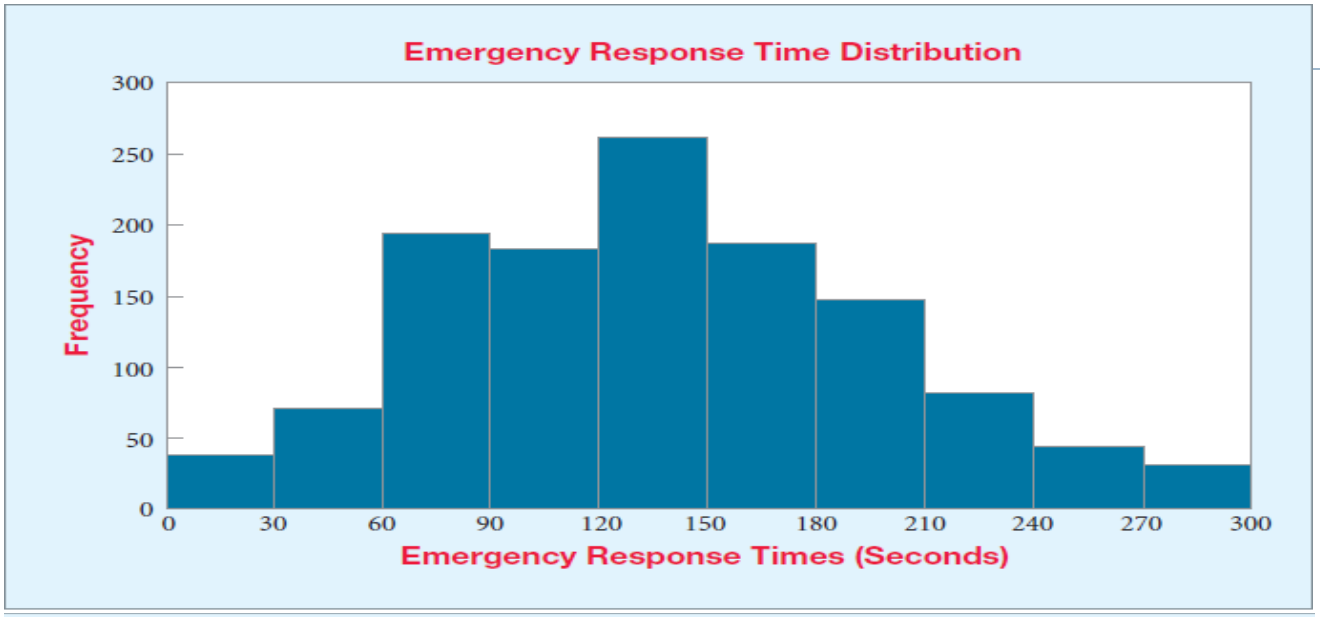
PROBLEM: EMERGENCY RESPONSE

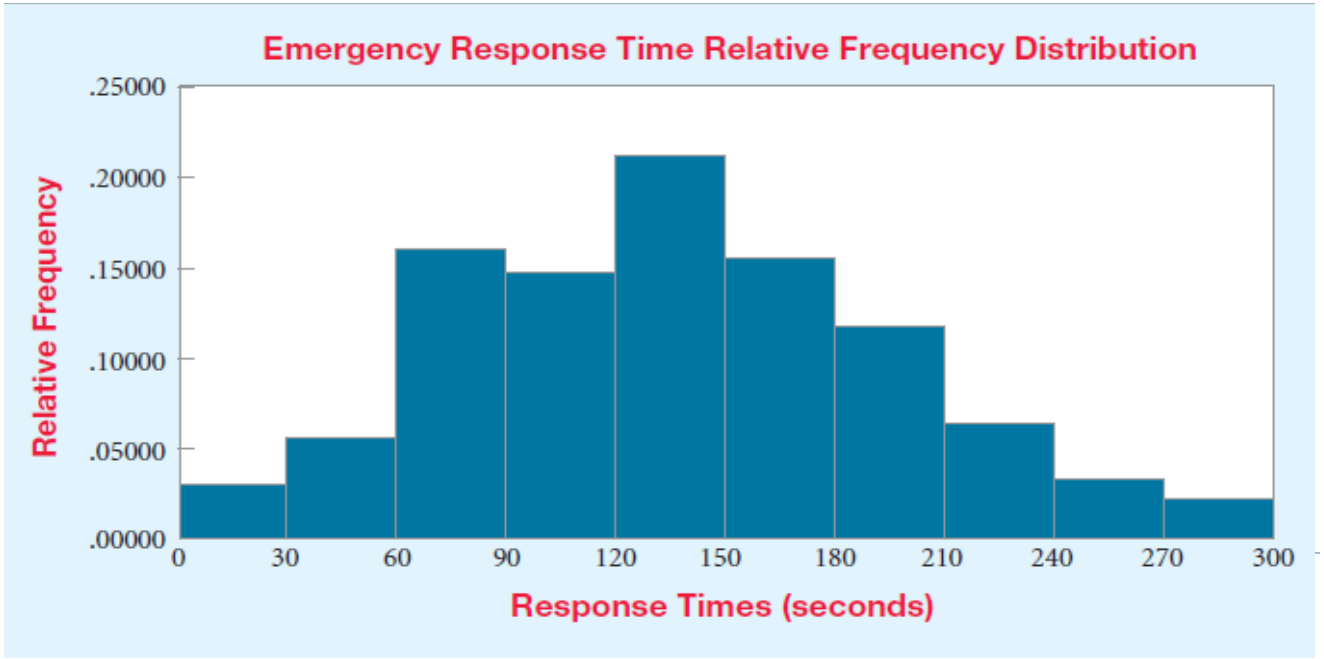
"The emergency responses team in Hospital XYZ is interested in analyzing the time needed for response teams to reach their destinations in emergency situations after leaving their stations. The team has acquired the response times for 1,220 calls last month."

What can you summarize from the graph?

Response Time	Frequency	Relative Frequency	Cumulative Relative Frequency
0 and under 30	36	36/1220 = 0.0295	0.0295
30 and under 60	68	68/1220 = 0.0557	0.0852
60 and under 90	195	195/1220 = 0.1598	0.2451
90 and under 120	180	180/1220 = 0.1475	0.3926
120 and under 150	260	260/1220 = 0.2131	0.6057
150 and under 180	182	182/1220 = 0.1492	0.7549
180 and under 210	145	145/1220 = 0.1189	0.8738
210 and under 240	80	80/1220 = 0.0656	0.9393
240 and under 270	43	43/1220 = 0.0352	0.9746
270 and under 300	31	31/1220 = 0.0254	1.0000
	1,220	1.0000	

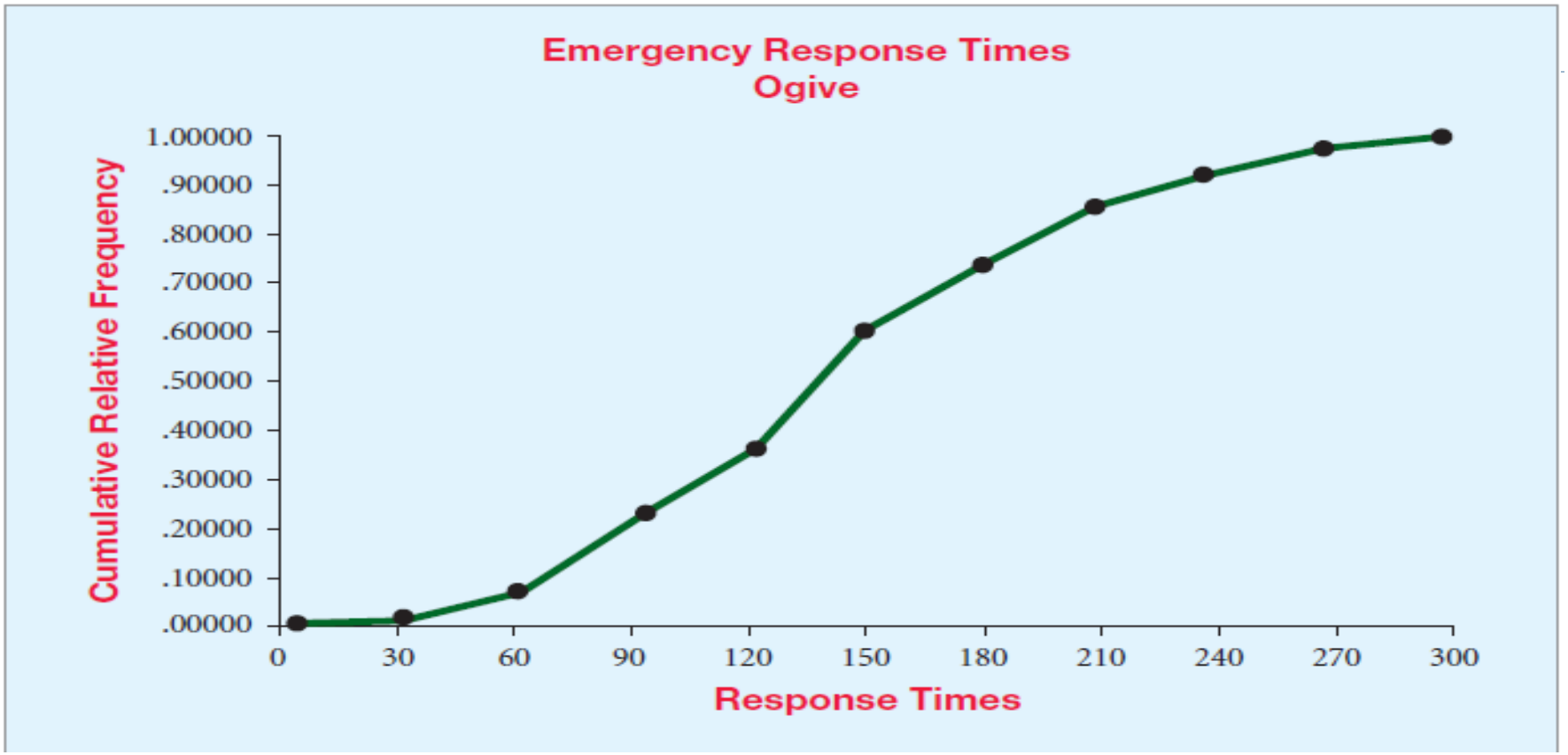
HISTOGRAMS & OGIVE





Both have same shape either with Frequency or relative frequency

HISTOGRAMS & OGIVE



Why Ogive?

- To best used when you want to display the total at any given time.
- The relative slopes from point to point will indicate greater or lesser increases; for example, a steeper slope means a greater increase than a more gradual slope.

BAR CHART

- Previously, tools to describe discrete and continuous data in raw form. What if data is a categorical type? ➔ Bar chart

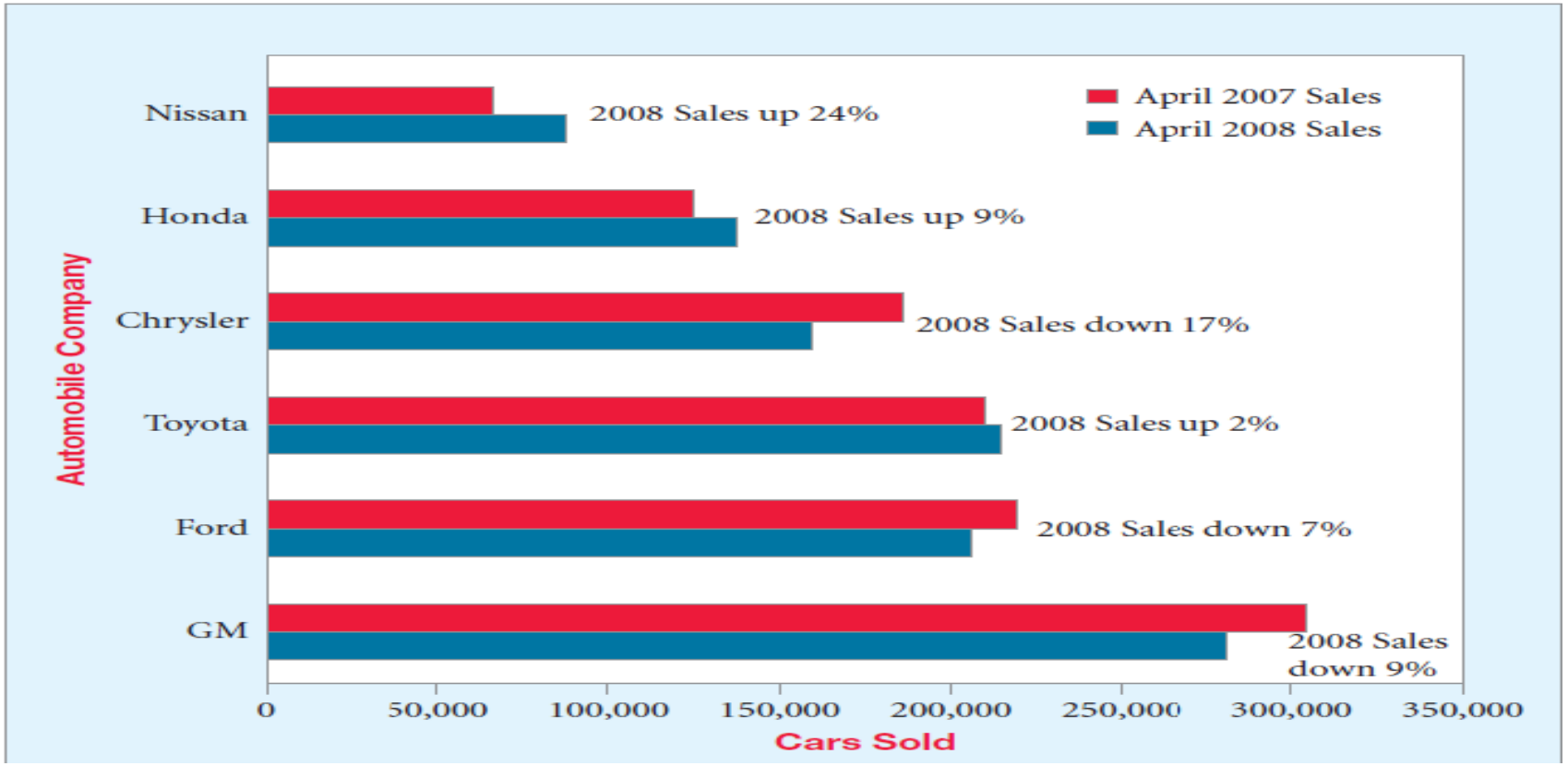
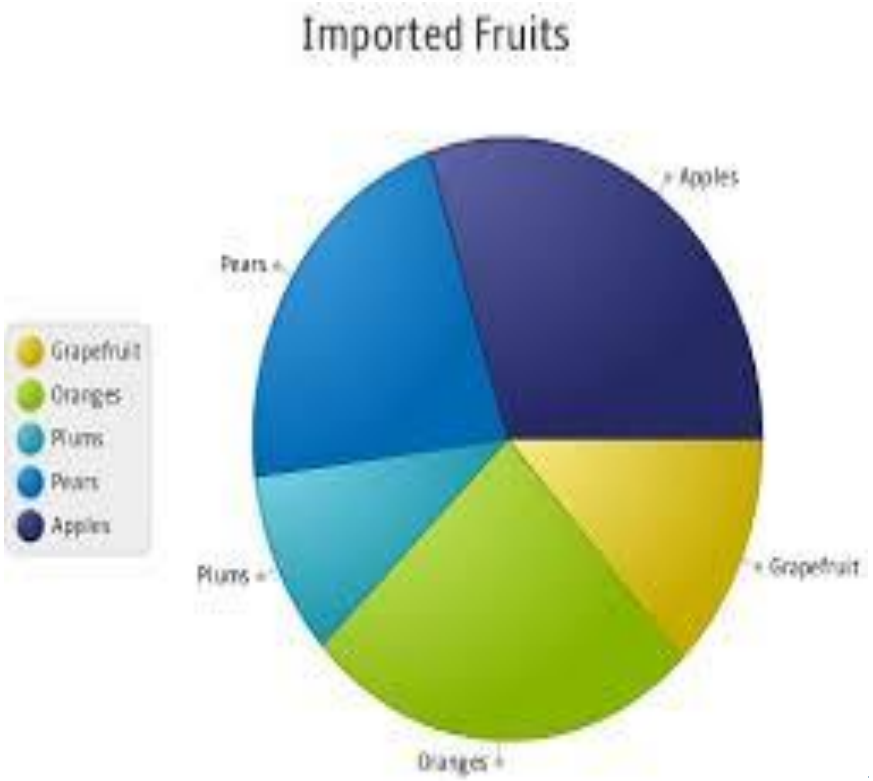
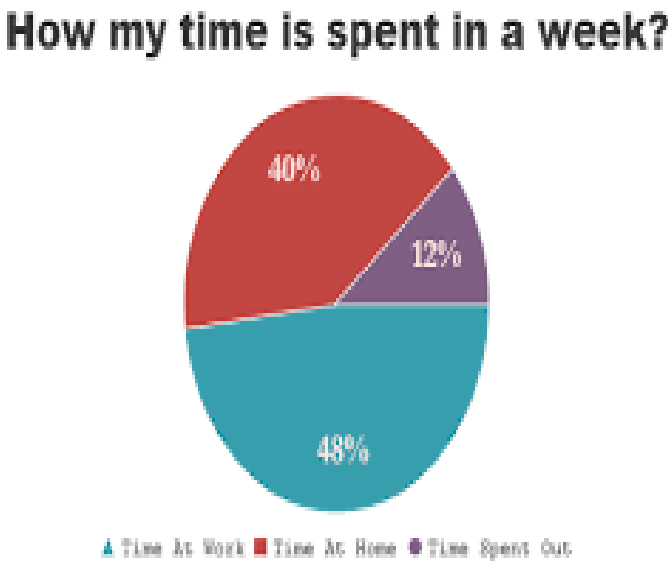


Figure: Bar Chart comparing April 2007 and January 2008 Cars Sold

PIE CHART

- Graph which in the shape of circle and the circle is divided into "slices" which corresponding to classes or categories.





STEM AND LEAF DIAGRAM

- Used for an exploratory analysis of quantitative data and it shows individual data value.

PROBLEM: WALK-IN HEALTH CLINIC

"The administrator for the ABC Walk-in Health Clinic is interested in performing an analysis of the number of patients who enter the clinic daily. One method for analyzing the data for a sample of 200 days is the stem and leaf diagram. The following data represent the number of patients on each of the 200 days."

What can you summarize from the graph?

113	112	63	127	110	129	142	115	192	94
165	121	105	140	85	93	105	140	93	126
183	118	67	104	162	110	76	109	91	132
88	96	132	80	144	112	57	139	123	124
172	149	198	114	88	111	133	117	138	134
53	147	108	109	153	89	159	99	130	93
161	118	115	117	128	98	125	184	134	132
117	127	166	72	122	109	124	92	82	69
110	128	151	67	142	177	135	121	143	89
160	115	138	79	104	76	89	110	44	140
117	103	59	109	145	117	162	108	141	139
148	175	107	117	87	87	150	152	80	168
88	127	131	85	143	101	137	111	128	147
110	81	111	149	154	90	150	117	101	116
153	176	112	147	87	177	190	66	62	154
143	122	176	153	97	106	86	62	146	98
134	135	127	118	109	143	146	152	140	95
102	137	158	69	122	135	136	129	91	136
135	86	131	154	132	59	136	85	142	137
155	190	120	154	102	109	97	157	144	149

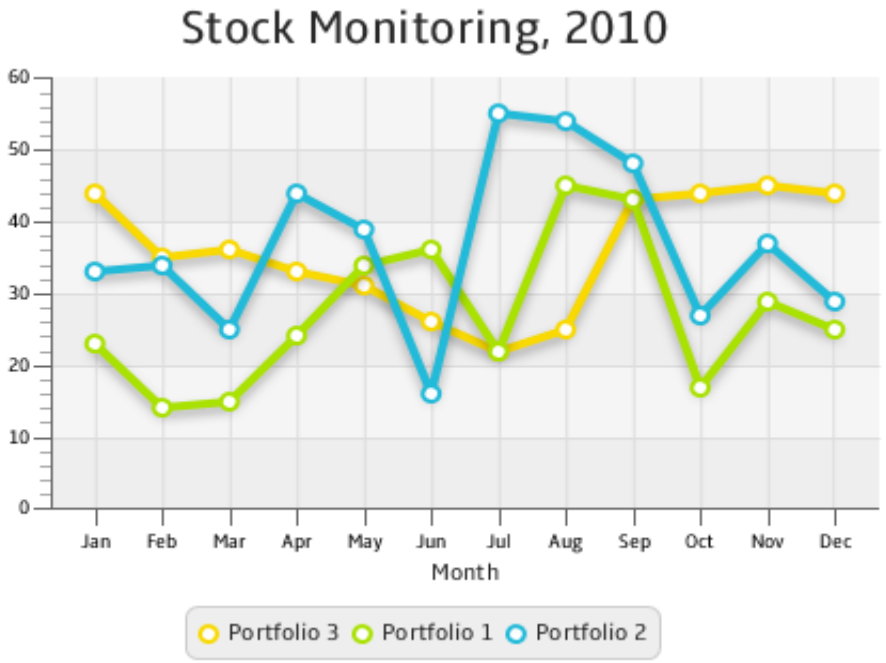
4	4
5	3 7 9 9
6	2 2 3 6 7 7 9 9
7	2 6 6 9
8	0 0 1 2 5 5 5 6 6 7 7 7 8 8 8 9 9 9
9	0 1 1 2 3 3 3 4 5 6 7 7 8 8 9
10	1 1 2 2 3 4 4 5 5 6 7 8 8 9 9 9 9 9 9
11	0 0 0 0 0 1 1 1 2 2 2 3 4 5 5 5 6 7 7 7 7 7 7 7 8 8 8
12	0 1 1 2 2 2 3 4 4 5 6 7 7 7 7 8 8 8 9 9
13	0 1 1 2 2 2 2 3 4 4 4 5 5 5 5 6 6 6 7 7 7 8 8 9 9
14	0 0 0 0 1 2 2 2 3 3 3 3 4 4 5 6 6 7 7 7 8 9 9 9
15	0 0 1 2 2 3 3 3 4 4 4 4 5 7 8 9
16	0 1 2 2 5 6 8
17	2 5 6 6 7 7
18	3 4
19	0 0 2 8

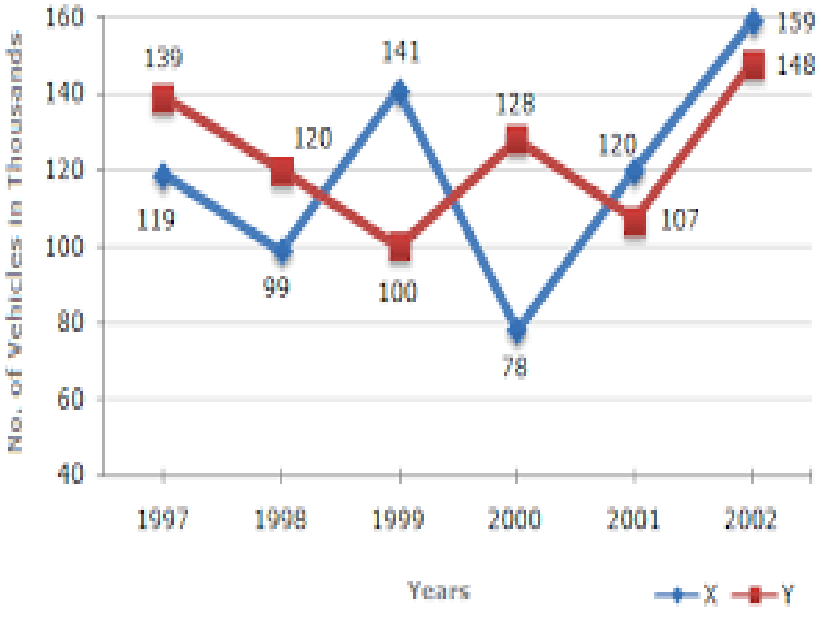
Figure: Stem and Leaf Diagram for Walk-in Clinic Patients

Summary: The stem and leaf diagram shows that most days have between 80 and 160 patients, with the most frequent value in the 110- to 120-patient range.

LINE CHART

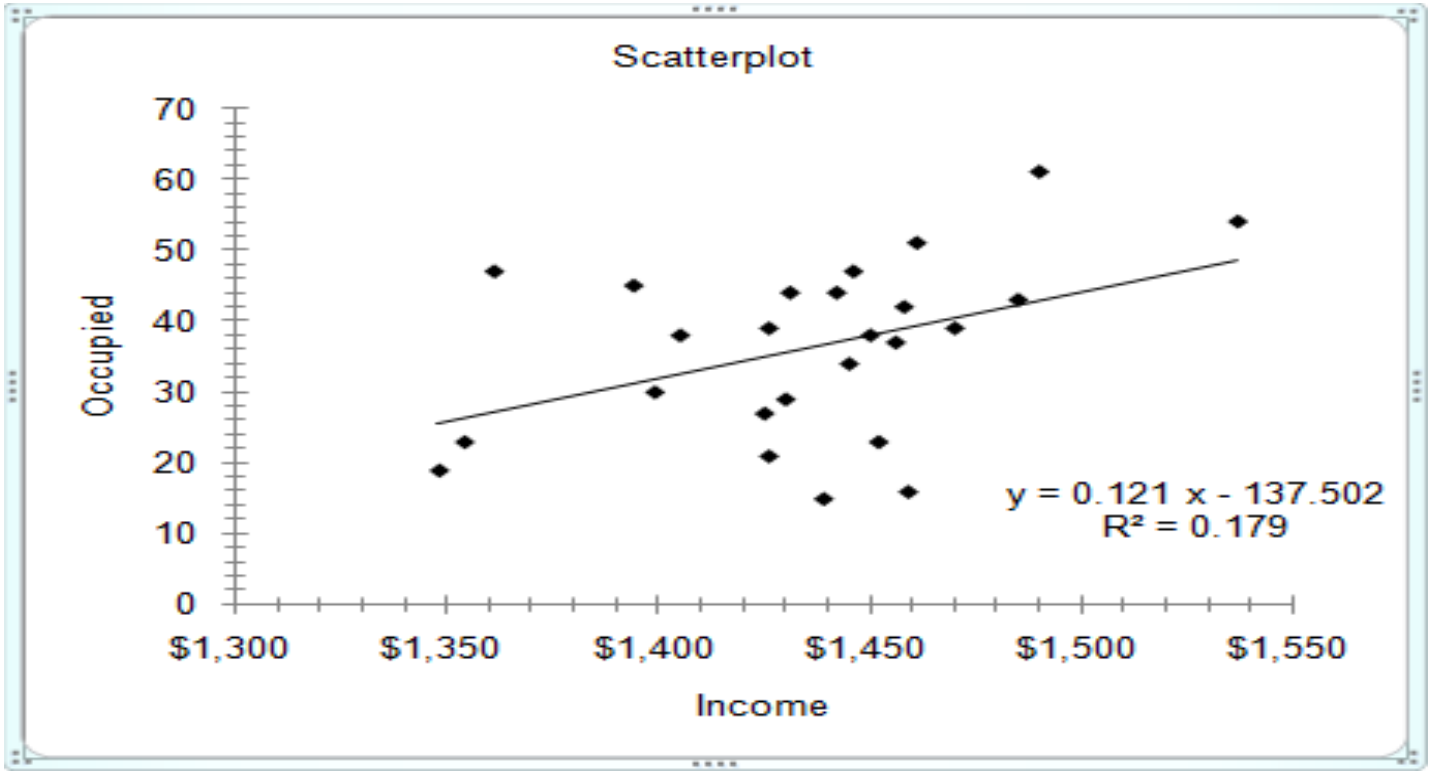
- Two dimensional chart showing time on the horizontal axis and the variable of interest on the vertical axis
- Tools which best illustrate time-series data that are measured over time (example: daily, monthly, quarterly and etc)





SCATTERPLOT/SCATTER DIAGRAM

Definition: graph of two variables with dot to represent each observation



"COMMENT ON THE SCATTERPLOT"

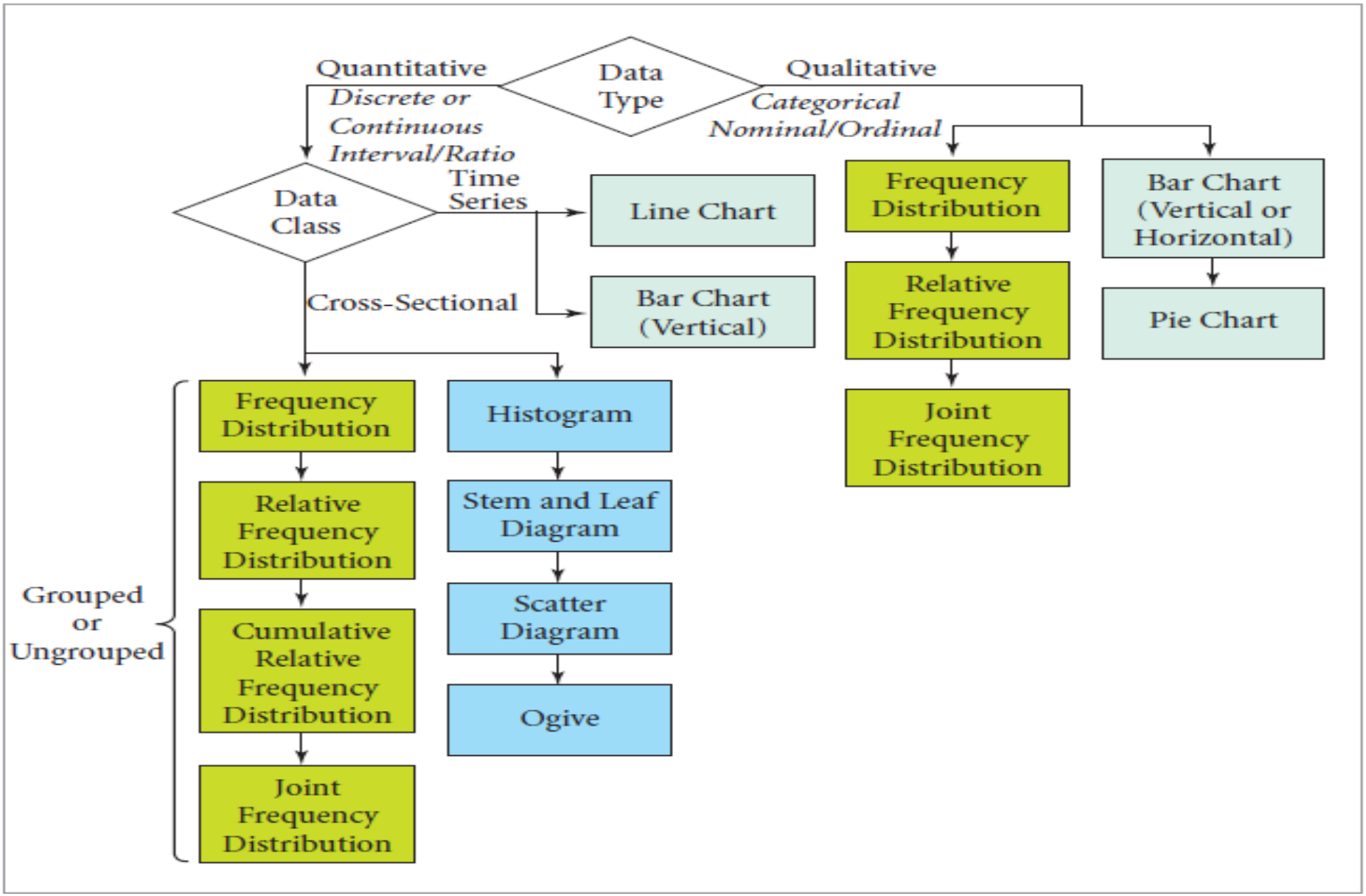
Shape: Does relationship look linear?

Outliers: Are there any unusual points?

Direction: Is linear relationship positive or negative?

Strength: Is the line strong?

SUMMARY – DESCRIBING DATA (VISUAL WAY)



DESCRIBING YOUR DATA

NUMERICAL WAY

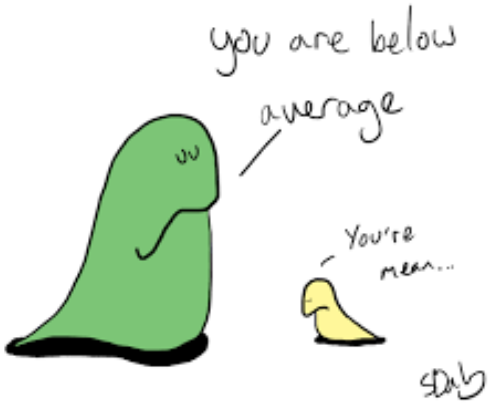
- √ Measures of Center and Location
- Measures of Variation

MEASURES OF CENTER AND LOCATION

- Previously (Histogram): visual indication of where data are centered and how much spread there is in the data around the center.

- Need a measurement to compute center and location about the data.
  - Measurements: Mean, Median, Mode & Weighted Mean
  - Display and measure the data range: Box and Whisker Plot

1. MEAN



Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where:

μ = Population mean (mu)
N = Population size
x_i = ith individual value of variable x

ALERT!!!
The Mean measure can be affected by extreme values (for example: income/salary data)

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where:

$\bar{x}$ = Sample mean (pronounced "x-bar")
n = Sample size

2. MEDIAN

Arrange the data point from smallest to largest. Then compute median using equation 3.3:

Median Index

$$i = \frac{1}{2}n \qquad (3.3)$$

where:

i = The index of the point in the data set corresponding to the median value

n = Sample size

If i is not an integer, round its value up to the next highest integer. This next highest integer then is the position of the median in the data array.

If i is an integer, the median is the average of the values in position i and position $i + 1$.

Example 1:

Personnel manager has hired 10 new employees. The age of the employees is as follows:

23 25 25 34 35 45 46 47 52 54

Thus, median index = $\frac{1}{2}n = \frac{1}{2}(10) = 5$, therefore the $M_d = \frac{35+45}{2} = 40$

Example 2:

Customers at a restaurant are asked to rate the service they receives on a scale of 1 to 100. A total of 15 customers were asked to provide the ratings. The data are presented as follows:

60 68 75 77 80 80 80 85 88 90 95 95 95 95 99

Thus, median index = $\frac{1}{2}n = \frac{1}{2}(15) = 7.5$ (round up to 8), therefore the M_d = 85

3. MODE

Mode is the value in a data set that occurs most frequent.

ALERT!!!

Occasionally used as a measure of central location. HOWEVER, useful in describing the central location value for sizes (clothes, shoes and etc)

Example:

The owners of Smoky Mountain Pizza are planning to expand their restaurant to include an open-air patio. Before finalizing the design, the managers want to know what the most frequently occurring group size is so they can organize the seating arrangements to best meet demand. A sample of 20 groups was selected at random. These data are:

$\{x_i\} = \{people\} = \{2, 4, 1, 2, 3, 2, 4, 2, 3, 6, 8, 4, 2, 1, 7, 4, 2, 4, 4, 3\}$

Solution: Organize the data into frequency distribution

x_i	Frequency
1	2
2	6
3	3
4	6
5	0
6	1
7	1
8	1
	Total = 20

Summary: The frequent occurring group size of people is 2 and 4 people

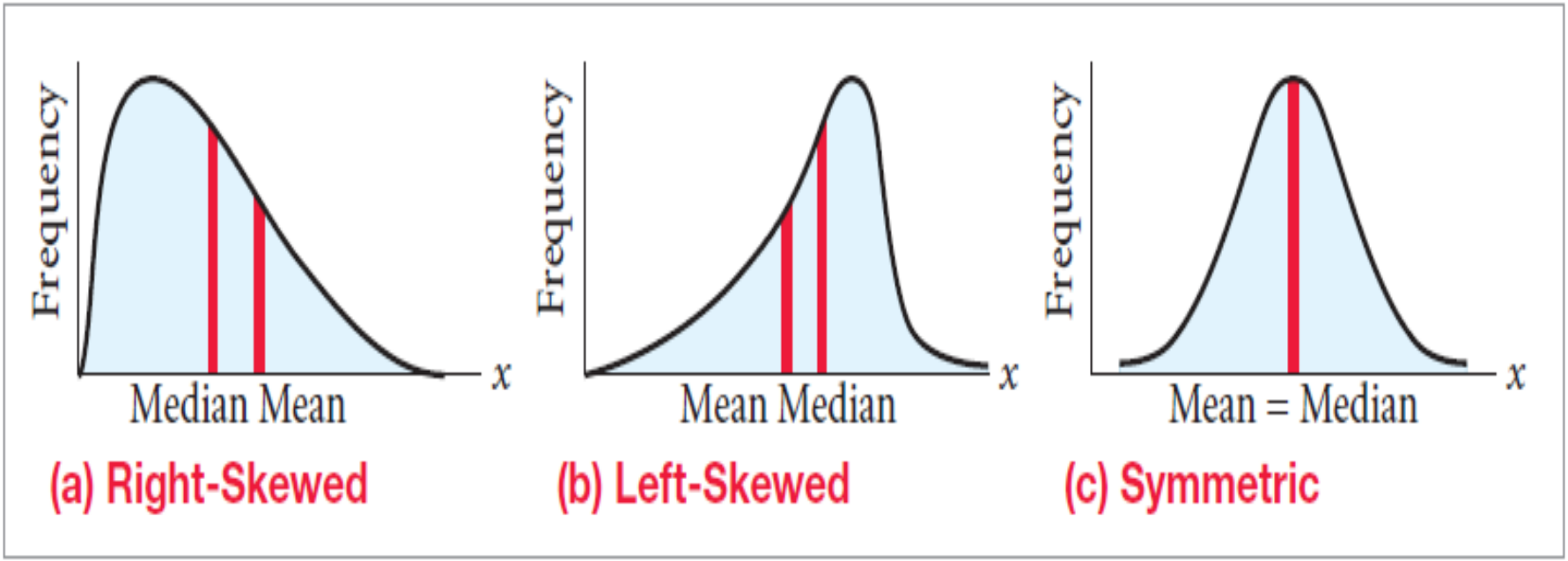
SKEWED AND SYMMETRIC DISTRIBUTION

- Data in population or sample can be either:
  - Symmetric (Normal – Bell Shaped) Distribution – data sets whose values are evenly spread around the center.
  - Skewed Distribution – data sets that are not symmetric.

- Skewness statistics – implies the direction of skewness.
  - The higher the absolute value, the more the data are skewed.
  - When the data is highly skewed, median is a useful measure of the center

- Kurtosis statistics – implies **how tall and sharp the central peak is**, relative to a standard bell curve.

SKEWED AND SYMMETRIC DISTRIBUTION



DESCRIPTIVE STATISTICS: LETS PRACTICE

PROBLEM: College and Universities Tuition Fees

"The cost of tuition is an important factor that most students and their families consider when deciding where to attend college. The data file **Colleges and Universities** contains data for a sample of 718 colleges and universities in the United States. The cost of out-of-state tuition is one of the variables in the data file. Suppose you are the guidance counselor who will be advising students about the college choices wishes to conduct a descriptive analysis. What can you conclude from the descriptive analysis?"

OTHER MEASURES OF LOCATION

- Other measure of location are:
  1. Weighted Mean (population & sample)
  2. Percentiles
  3. Quartiles
  4. Box-Whisker Plots

1. Weighted Mean - 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 ith data value
x_i = The ith data value

Example for Weighted Mean Business Application:

"Recently, the law firm of Myers & Associates was involved in litigating a discrimination suit concerning ski instructors at a ski resort in Colorado. One ski instructor from Germany had sued the operator of the ski resort, claiming he had not received equitable pay compared with the other ski instructors from Norway and the United States. In preparing a defense, the Myers attorneys planned to compute the mean annual income for all seven Norwegian ski instructors at the resort. However, because these instructors worked different numbers of days during the ski season, a weighted mean needed to be computed." What can you conclude from the weighted mean calculated?

x_i = Income:	$7,600	$3,900	$5,300	$4,000	$7,200	$2,300	$5,100
w_i = Days:	50	30	40	25	60	15	50

OTHER MEASURES OF LOCATION

2. 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

PERCENTILE APPLICATION

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, the80th 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

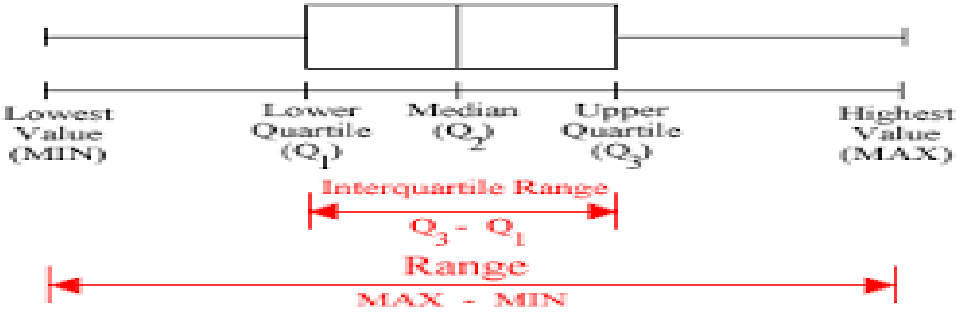
3. Quartiles

- Those value that divide the data set into four equal-sized groups.
- The 1st quartile (Q_1) = 25th percentile, 2nd quartile (Q_2) = median, 3rd quartile (Q_3) = 75th percentile and 4th quartile = 100th percentile

4. Box and Whisker Plot

- Descriptive tool that many decision maker use.
- A graph that composed of 2 parts: box and whiskers.
- Box has the width that ranges from Q_1 to Q_3 and vertical line through the box is placed at the median. Whisker represent the lower limit and the upper limit.
- Used to identify outliers.

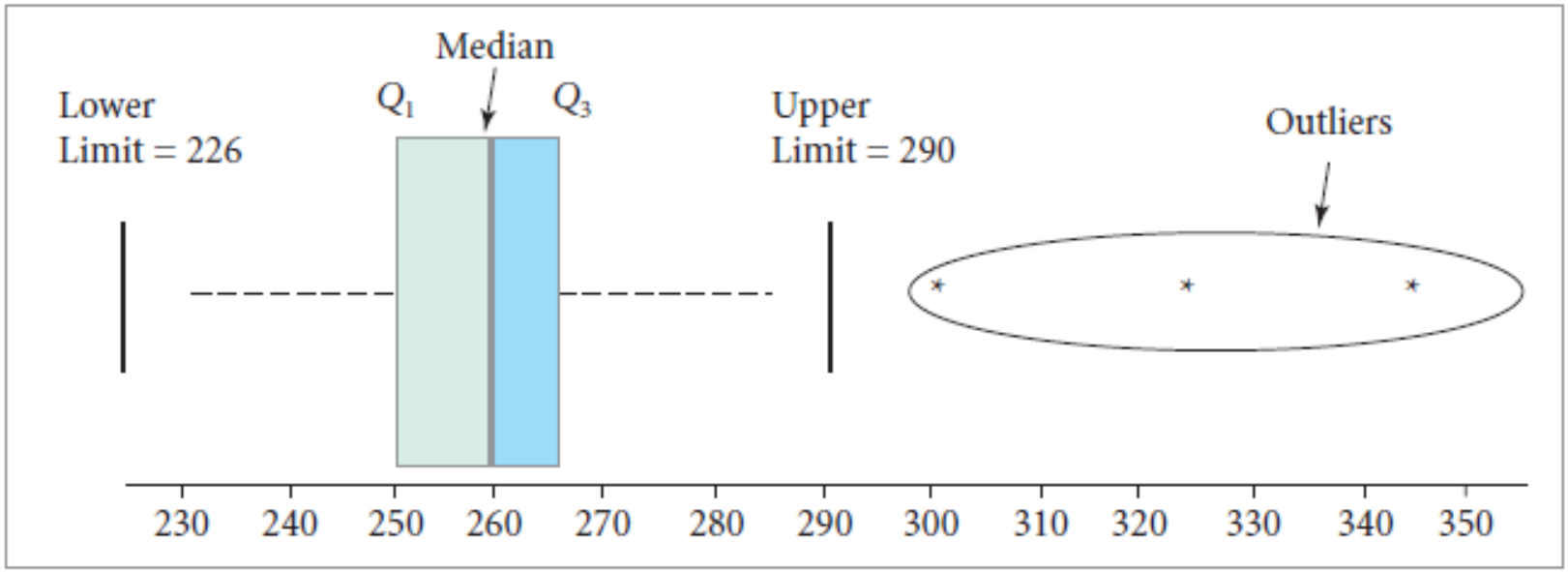
OTHER MEASURES OF LOCATION



Problem:

A demand analyst for XYZ Petroleum company has recently performed a study at one of the company's stores in which he asked customers to set their trip odometer to zero when they filled up. Then, when the customers returned for their next fill-up, he recorded the miles that had been driven. He now plans to make a presentation to the board of directors and wishes to construct a box and whisker plot as part of the presentation as a way to describe the data and identify any outliers. The sorted sample data showing the miles between fill-ups is as follows:

231	236	241	242	242	243	243	243	248
248	249	250	251	251	252	252	254	255
255	256	256	257	259	260	260	260	260
262	262	264	265	265	265	266	268	268
270	276	277	277	280	286	300	324	345


Median
Lower
Limit = 226
Q_1
Q_3
Upper
Limit = 290
Outliers
230
240
250
260
270
280
290
300
310
320
330
340
350

DATA-LEVEL ISSUES

- You should be aware of the level of data before computing the numerical measures.
  - Example 1: common mistakes to compute means on nominal-level data.
    - Problem example: An electronic manufacturer recently surveyed a sample of customers to determine whether they preferred black, white or colored stereo cases. The data were coded as:

      1 = black

      2 = white

      3 = colored

      The responses are: response = {1,1,3,2,1,2,2,2,3,1,1,1,3,2,2,1,2}

      If mean is calculated, then sample mean, $\bar{x} = 1.765$

      It can be seen that it would be between black and white. Thus, it is meaningless.

      Don't Let it Happen!!!

DATA-LEVEL ISSUES

- Example 2: common mistakes to compute means on ordinal-level data.
  - Problem example: In market research, 5 or 7 point scale is often used to measure customer's attitude about products or TV commercials. For example, we might set up the scale:

    1 = Strongly Agree

    2 = Agree

    3 = Neutral

    4 = Disagree

    5. = Strongly Disagree

    10 customer responses are: response = {2,2,1,3,3,1,5,2,1,3}

    If mean is calculated, then sample mean rating, $\bar{x} = 2.3$

    It can be seen that it would be Agree. Thus, we assume the distance of attitude (1,2,3,4 and 5) is same (respond to the survey have same definition). Thus, it is meaningless.

    Don't Let it Happen!!!

DATA-LEVEL ISSUES

Descriptive Measure	Computation Method	Data Level	Advantages/ Disadvantages
Mean	Sum of values divided by the number of values	Ratio Interval	• Numerical center of the data • Sum of deviations from the mean is zero • Sensitive to extreme values
Median	Middle value for data that have been sorted	Ratio Interval Ordinal	• Not sensitive to extreme values • Computed only from the center values • Does not use information from all the data
Mode	Value(s) that occur most frequently in the data	Ratio Interval Ordinal Nominal	• May not reflect the center • May not exist • Might have multiple modes

DESCRIBING YOUR DATA

NUMERICAL WAY

- Measures of Center and Location
- √ Measures of Variation

MEASURES OF VARIATION (SPREAD)

- Variation = A set of data exhibits variation if all the data is not in the same value.

Example:

Plant A	Plant B
15 units	23 units
25 units	26 units
35 units	25 units
20 units	24 units
30 units	27 units

Plant A	Plant B
$\bar{x} = 25$ units	$\bar{x} = 25$ units
$M_d = 25$ units	$M_d = 25$ 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.

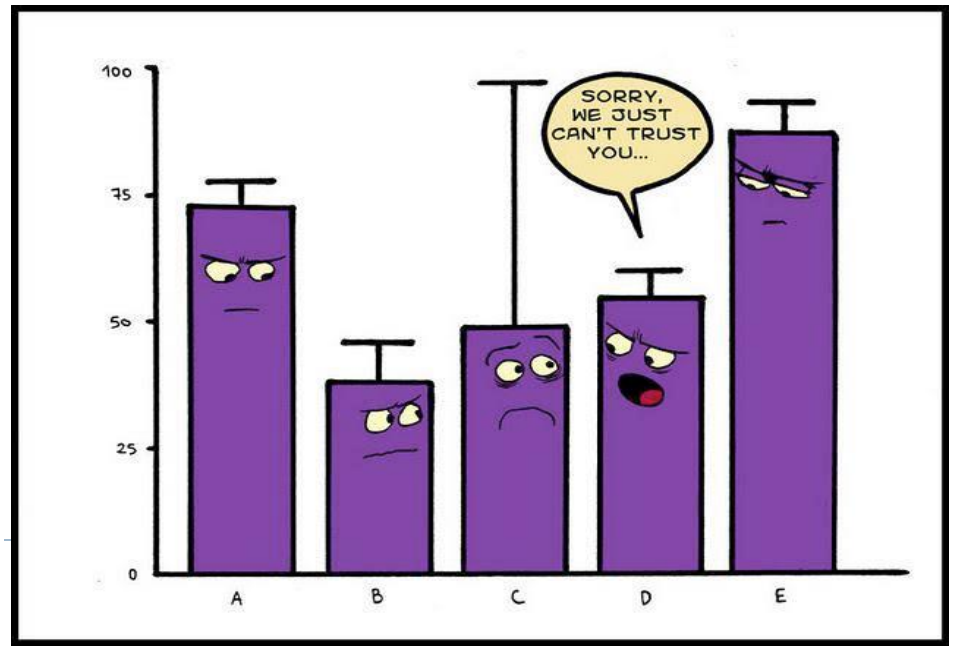
HOWEVER!!!: There is a HUGE variation in terms of day by day of the production output. Plant B is very stable, producing almost the same amount every day. Plant A varies considerably, with some high-output days and some low-output days.

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

MEASURES OF VARIATION (SPREAD)

- Variation is either a natural part of a process (or inherent to a product) or can be attributed to a special cause that is not considered random.

- Several different measures of variation that are used in business decision making:
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation



1. RANGE

- The range is a measure of variation that is computed by finding the difference between the maximum and minimum values in a data set.
- Range = Maximum value – Minimum value
- But, range sensitive to extreme values.

2. INTERQUARTILE

- A measure of variation that tends to overcome the range's susceptibility to extreme values
- Interquartile = Third Quartile(Q_3)– First Quartile(Q_1)
- But, the measure did not uses all the available data in its computation

3 & 4. VARIANCE AND STANDARD DEVIATION

- The measure (variance and standard deviation) incorporate all the values in the data set.

- These two measures are closely related.
  - The standard deviation is the square root of the variance.

- Standard deviation – is the original unit, thus the standard deviation are usually used to measure variation in a population and sample.
  - Why? – because dealing with original units are easier compare with the square of the unit (variance)

COEEFICIENT OF VARIATION

- If 2 different sample having same mean ➔ can use standard deviation to measure variation

- BUT, if 2 distribution did not have same mean, it does not make sense to measure based on standard deviation.
  - need other measurement to measure the variation ➔ COEEFICIENT OF VARIATION

- Coefficient of variation (CV):
  - The ratio of the standard deviation to the mean.
  - It is expressed as a percentage
  - the distribution with the largest *CV* is said to have the greatest relative spread.

STANDARDIZED DATA VALUES

- When do we need this measure?
  - when the mean scores and measures of spread (variation) are different between the 2 sample distribution.

- Standardized data values are sometimes refer as z scores.

Standardized Sample Data

$$z = \frac{x - \bar{x}}{s}$$

where:

x = Original data value
$\bar{x}$ = Sample mean
s = Sample standard deviation
z = The standard score

STANDARDIZED DATA VALUES APPLICATION

Problem:

Consider a company that uses placement exams as part of its hiring process. The company currently will accept scores from either of two tests: ABC Hiring and XYZ-Screen. The problem is that the ABC Hiring test has an average score of 2,000 and a standard deviation of 200, whereas the XYZ Screen test has an average score of 80 with a standard deviation of 12. (These means and standard deviations were developed from a large number of people who have taken the two tests.) Suppose the company is considering two applicants, Ahmad and Sarah. Ahmad took the ABC Hiring test and scored 2,344, whereas Sarah took the XYZ-Screen and scored 95.How can the company compare applicants when the average scores and measures of spread are so different for the two tests?

Solution:

Ahmad: ABC Hiring test mean, $\mu = 2000$ and standard deviation, $\sigma = 200$. Ahmad score = 2,344.

$$z = \frac{2344 - 2000}{200} = 1.72$$

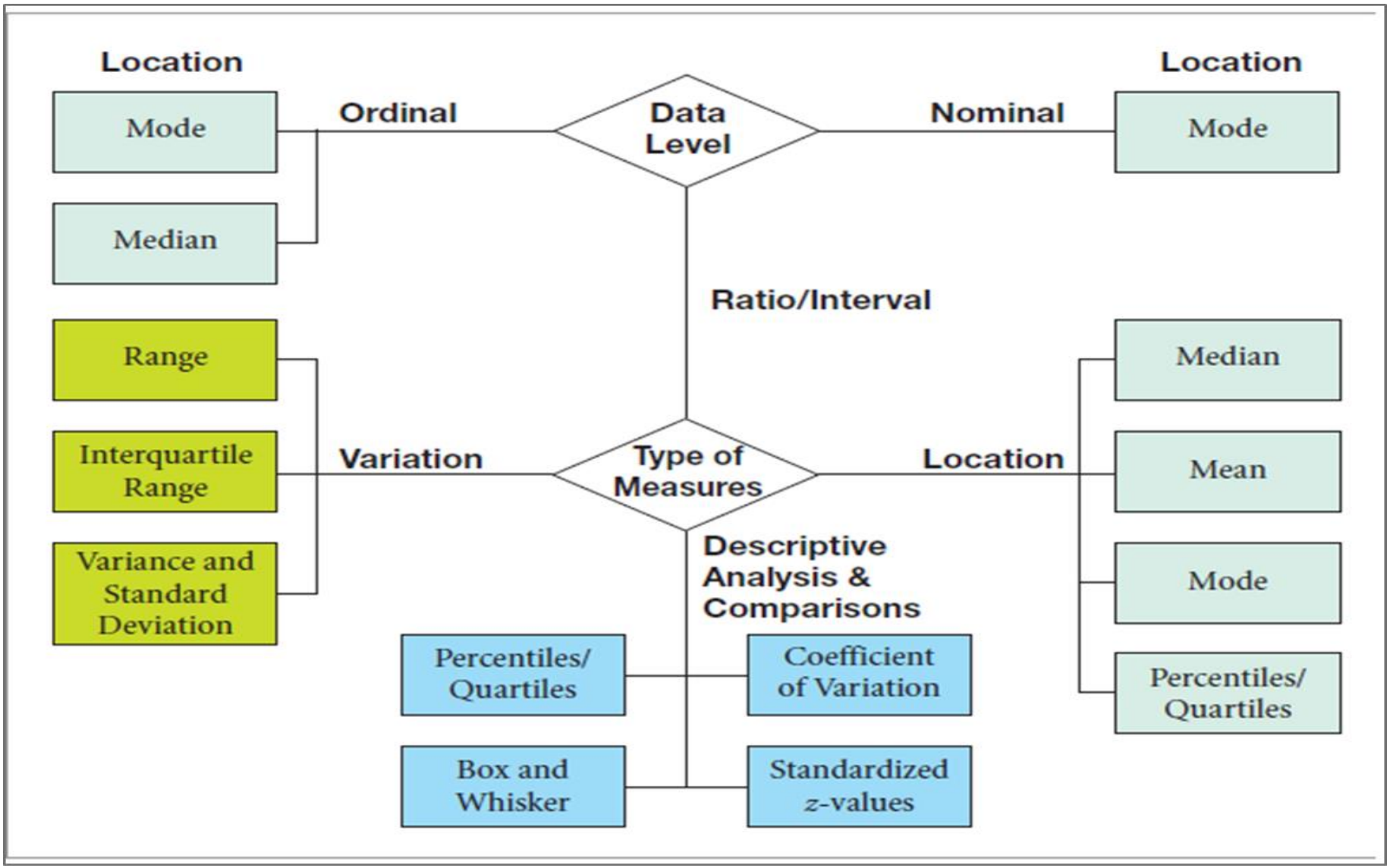
Sarah: XYZ Screen test mean, $\mu = 80$ and standard deviation, $\sigma = 12$. Ahmad score = 95.

$$z = \frac{95 - 80}{12} = 1.25$$

Summary:

Therefore, even though the two tests used different scales, standardizing the data allows us to conclude Ahmad scored relatively better on his test than Sarah did on her test.

SUMMARY: NUMERICAL STATISTICAL MEASURES



THE END OF ITEM 1