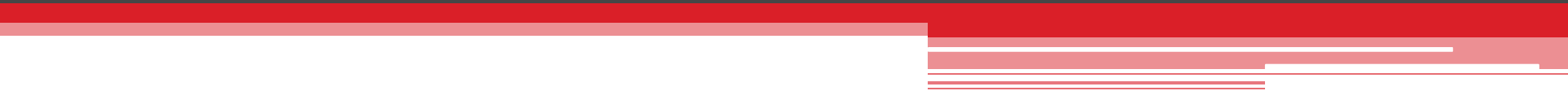


Introduction to Statistics



Dada

PNU-Math

Statistics is a scientific body of knowledge that deals with the *collection, presentation, analysis, and interpretation of data.*



Descriptive statistics refers to the methods that are employed in presenting and describing data. It involves gathering, organizing, presenting, and describing data.

It may answer questions such as:

1. How many students are interested to do Educational Research?
2. What are the highest and the lowest scores obtained by applicants in a managerial position?
3. What are the characteristics of the most successful sales representative according to customers?
4. Which group of employees produced more outputs?
5. What proportion of Math major students likes Statistics?

Inferential statistics refers to the methods of making inferences about a larger group called *population*, based on observations of a smaller group deemed representation of that population.

POPULATION is the total collection of all the elements of a well-defined group (e.g. people, events, objects, measurements) that one wishes to investigate.

SAMPLE is a subgroup obtained from a population where observations are made.

Examples:

1. Is there a significant difference between the proportion of students who are interested to take *Elementary Statistics* online and those who are not?
2. A company owner observes the correlation between the educational attainment and job performance rating of his employees.
3. A dermatologist tests the relative effectiveness of a new brand of medicine for whitening.
4. According to our recent poll, 43% of Americans brush their teeth incorrectly.

The *independent variable* is the variable that is systematically *manipulated* by the researcher usually to determine its effect on another variable called *dependent*.

Bogart wants to investigate the effect of kinds of food intake on mathematical ability. *Identify the independent variable and the dependent variable.*

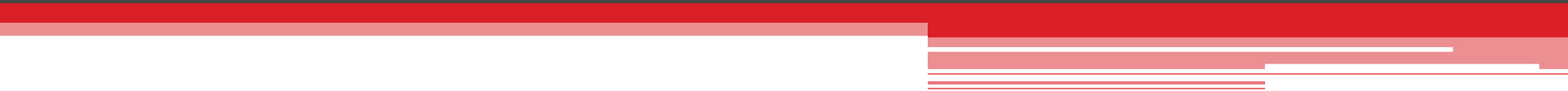
Quantitative Data are numbers representing counts or measurements.

Discrete data are values that are considered as whole units.

Continuous data are values that result from infinitely many possible values that can be associated with points on a continuous scale.

Qualitative Data are non numerical data.

Levels of Measurement



Dada

PNU-Math

Nominal

- consist of names or categories only
- no magnitude relationship
- data have no numerical significance and cannot be compared nor arranged according to a particular scheme
- no operations of addition, subtraction, multiplication, or division

Examples: nationality, location, marital status, gender (sexuality?)

Ordinal

- data can be ranked or ordered
- exact difference between values cannot be ascertained
- differences between the categories would not necessarily be the same
- with low level of the property of magnitude but does not have a property of equal intervals between adjacent units

Examples: size of t-shirt, movie ratings, socio-economic class

Interval

- possesses the properties of ordinal scale and has equal intervals between adjacent units
- have no '**true**' zero point

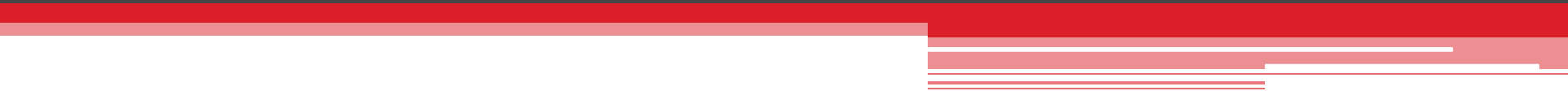
Examples: IQ scores, temperature

Ratio

- Ratio scale is an interval scale that has an absolute zero.
- When using a ratio scale, you can perform all mathematical operations.

Examples: grams of fats, test score, number of chairs, reaction time

Sampling

A decorative graphic consisting of several horizontal stripes in red and white, located at the bottom of the slide.

SAMPLING is the process of selecting the sample.

- What is the size of the sample?
- What is the population?



Determining the Sample Size

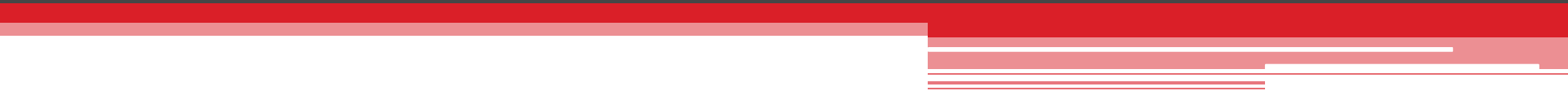
Slovin's Formula:
$$n = \frac{N}{1 + Ne^2}$$

n = sample size

N = population size

e = margin of error (usually $e = 0.05$)

Sampling Techniques



- (A) Probability Sampling
- (B) Non-Probability Sampling

A. Probability Sampling/ Random Sampling - each member of the population has equal chances of being chosen.

(1) *Simple Sampling*

(2) *Systematic Sampling*

(3) *Stratified Sampling*

(4) *Cluster Sampling*

Simple Sampling is a method of sampling in which at every selection from the population all remaining sampling units in the population have the same probability of being included in the sample.

- Lottery Method

Systematic Sampling

relies on arranging the target population according to some ordering **scheme** and then selecting elements at regular start and then proceeds with the selection of every ***Kth*** element from the onwards.

- (a) Every *Kth* Unit
- (b) Table of Random Numbers

Systematic Sampling

Example: Obtain a sample of 200 from a population of 2000.

STEP 1: Select a column of numbers and begin reading four-digit numbers. [note: Person 1 as 0001, Person 27 as 0027, Person 1999 as 1999]

STEP 2: Write down the first 200 numbers in the column that have a value of 2000 or less.

Table of Random Numbers

Table 6.1 *Part of a Table of Random Numbers*

011723	223456	222167	032762	062281	565451
912334	379156	233989	109238	934128	987678
086401	016265	411148	251287	602345	659080
059397	022334	080675	454555	011563	237873
666278	106590	879809	899030	909876	198905
051965	004571	036900	037700	500098	046660
063045	786326	098000	510379	024358	145678
560132	345678	356789	033460	050521	342021
727009	344870	889567	324588	400567	989657
000037	121191	258700	088909	015460	223350
667899	234345	076567	090076	345121	121348
042397	045645	030032	657112	675897	079326
987650	568799	070070	143188	198789	097451
091126	021557	102322	209312	909036	342045

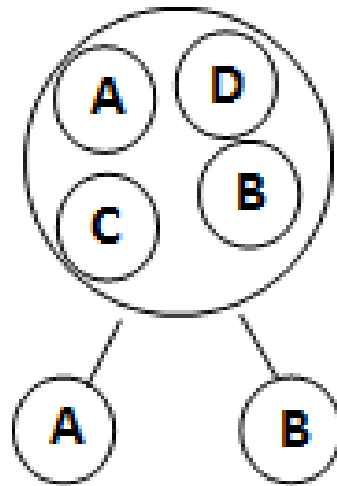
Simple Stratified Sampling

- population consists of definite **strata**, each of which is distinctly different, but the units within the stratum are as homogeneous as possible.
- Classification may be based on districts, socio-economic status, sex, work, etc. depending on the problem being studied.

Cluster Sampling

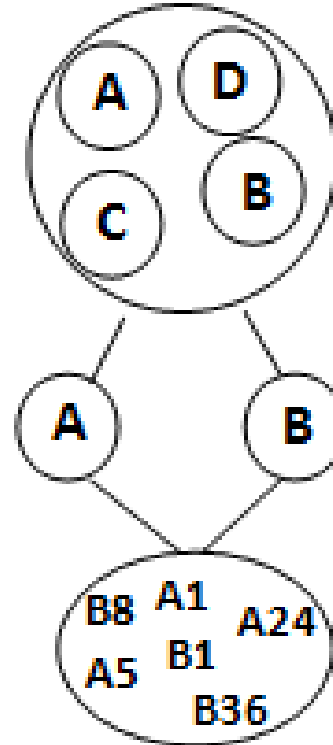
- A variation of the simple random sample that is particularly appropriate when *population of interest is infinite, when a list of the members of the population does not exist, or when geographic distribution of the individuals is widely scattered.*
 - single-stage cluster
 - two-stage cluster
 - multistage cluster

SS



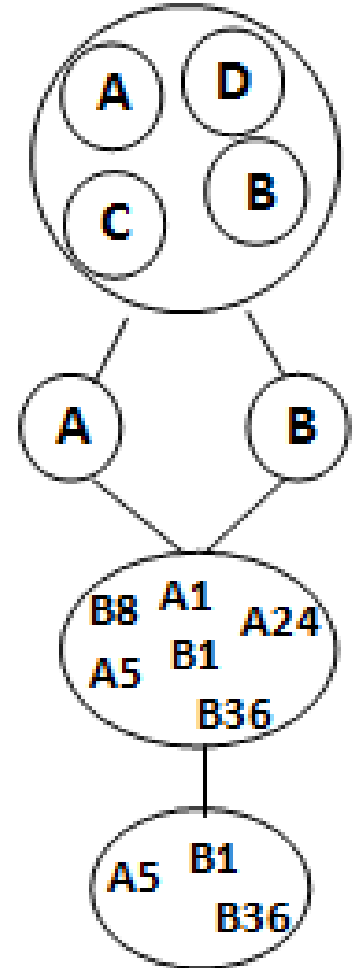
Sample: All the
A's and B's

TS



Sample: A1,
A5, A24, B1,
B8, & B36

MS



Sample: A5,
B1, & B36

*Tell the
procedure
of each type
of Cluster
Sampling.*

B. Non-Probability Sampling - there is no way of forecasting, estimating or guaranteeing that each element in the population will be represented in the sample

(1) *Purposive or Judgment Sampling*

(2) *Convenience Sampling*

(3) *Quota Sampling*

Purposive or Judgment Sampling

- This is based on certain criteria laid down by the researcher. People who satisfy the criteria are interviewed.

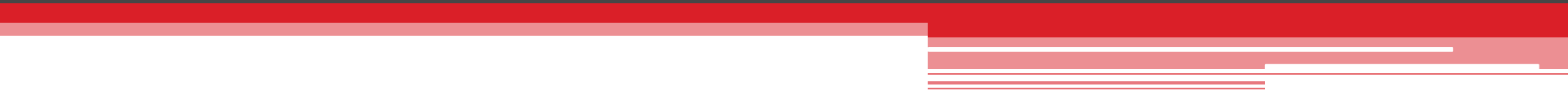
Convenience Sampling

- a sampling strategy based on the convenience of the researcher
- This strategy allows the use of any available group of research activities.
 - *Snowball Sampling*

Quota Sampling

- non-probability sampling that is equivalent to Stratified sampling
- sampling widely used in opinion polling and market research
- Interviewers are each given a quota of subjects of specified type to attempt to recruit.

Data Gathering



Dada

PNU-Math

(1) *Interview or Direct Method*

- the researcher obtains the information needed by **asking questions and inquiries** from the interviewee
- the researcher can **get more accurate answers or responses** since clarifications can be made if the interviewee or respondent does not understand the question
- costly and time-consuming

(2) *Questionnaire or Indirect Method*

- can save a lot of time and money in gathering data because questionnaires **can be given to a large number of respondents at the same time**
- cannot expect that all distributed questionnaires will be **retrieved** because some respondents simply ignore the questionnaires
- **clarification cannot be made** if the respondents do not understand the questions

Do You Have Math Anxiety?

A Self Test

Rate your answers from 1 to 5; add them up and check your score below.

(1) = Disagree, (5) = Agree.

1. I cringe when I have to go to math class. 1 2 3 4 5
2. I am uneasy about going to the board in a math class. 1 2 3 4 5
3. I am afraid to ask questions in math class. 1 2 3 4 5
4. I am always worried about being called on in math class. 1 2 3 4 5
5. I understand math now, but I worry that it's going to get really difficult soon. 1 2 3 4 5
6. I tend to zone out in math class. 1 2 3 4 5
7. I fear math tests more than any other kind. 1 2 3 4 5
8. I don't know how to study for math tests. 1 2 3 4 5
9. It's clear to me in math class, but when I go home it's like I was never there. 1 2 3 4 5
10. I'm afraid I won't be able to keep up with the rest of the class. 1 2 3 4 5

Check Your Score:

40-50 Sure thing, you have math anxiety. Check my 10 hints on how to reduce math anxiety.

30-39 No doubt! You're still fearful about math.

20-29 On the fence!.

10-19 Wow! Loose as a goose!

Math anxiety is an emotional reaction to mathematics based on a past unpleasant experience which harms future learning. A good experience learning mathematics can overcome these past feelings and success and future achievement in math can be attained.

Copyright © 1997-2006

by Ellen Freedman, All Rights Reserved.

(3) *Registration Method*

- method of collecting data governed by laws
- very reliable method since this is enforced by law

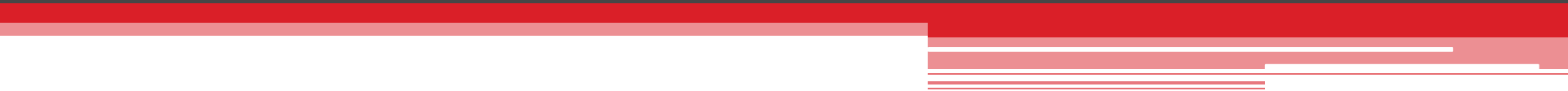
(4) *Experimental Method*

- is usually used to find out cause and effect relationships

(5) *Observation Method*

- is a way of gathering data by watching behavior, events, or noting physical characteristics in their natural setting

Presentation of Data



Dada

PNU-Math

Textual

Data are presented in paragraph or in sentences.

Official Population Count of NSO in the year 2010*

The final result of the latest Census of Population (POPCEN 2010) conducted by the National Statistics Office (NSO) placed the Philippine population at 92,337,852 / 92.34M persons as of May 1, 2010. The 2010 population is higher by 3,763,238 / 3.76M compared to 2007 population of 88, 574, 614 / 88.57M persons.

Among the 17 regions, CALABARZON (Region IVA) had the largest population with 12.61M, followed by the National Capital Region (NCR) with 11.86M and Central Luzon (Region III) with 10.14M. The population of these three regions together comprised more than one-third (37.47 percent) of the Philippine population.

Among the provinces, Cavite had the largest population with 3.09M. Bulacan had the second largest with 2.92M and Pangasinan had the third largest with 2.78M. On the other hand, the provinces with a population of less than 100,000 persons were Batanes (16,604), Camiguin (83,807), and Siquijor (91,066).

As projected by the NSO, this 2012 the Philippine population is about 97.6M and by 2014, the population may increase to 101.2M.

*Source: National Statistics Office, Manila.

April 04, 2012

Tabular

A **frequency distribution** is a table that indicates the number of times a score or a group of scores occur in a set of data.

Frequency Distribution

(1) Categorical Frequency Distribution

It is used for data that can be placed in specific categories, such as nominal or ordinal data.

(2) Frequency Distribution for Ungrouped Data

It is used when the range of the data is small.

(3) Frequency Distribution for Grouped Data

It is used when the range of the data is large.



Frequency Distributions for Ungrouped and Grouped Data

Frequency Distributions for *Ungrouped* Data

The scores obtained by ten students in a test in Math are 8, 12, 15, 9, 7, 10, 10, 6, 9, and 10. Prepare a frequency distribution of the scores.

Score	Tally	Frequency
15	/	1
12	/	1
10	///	3
9	//	2
8	/	1
7		0
6	//	2

Frequency Distributions for *Grouped* Data

- (1)** *Decide on the number of classes:* between 5 and 20 or Sturge's Formula.

$$\text{Number of classes} = 1 + \frac{\log n}{\log 2}$$

- (2)** *Determine the range (R).*

- (3)** *Determine the class width (w).*

$$w = \frac{R}{\text{Number of Classes}}$$

Frequency Distributions for *Grouped* Data

- (4) *Determine the lowest class interval. The **lower limit of this class interval** is either the lowest score or any value slightly less than this lowest score. This value serves as the **starting point**. Add the width (w) to the starting point to obtain the second lower limit. Next, add the width to the second lower limit to get the third, and so on **until the class interval containing the highest score has been obtained**.*
- (5) *Tally the frequencies and compute for **class marks, class boundaries, cumulative frequency, and relative frequency**.*

Frequency Distributions for *Grouped* Data

A. In a survey of 30 patients who smoked, the following data were obtained. Each value represents the number of cigarettes the patient smoked per day.

10	8	6	14	10	20
22	13	17	19	13	11
11	9	18	14	10	10
13	12	15	15	17	15
5	11	16	11	14	16

B. Guess how many F's are there in the paragraph, which will be shown in just 30 seconds!

- Make a frequency distribution table of the 40 guesses.
- Compute also the *class marks*, *class boundaries*, *cumulative frequency*, and *relative frequency*.

Frequency Distributions for *Grouped* Data

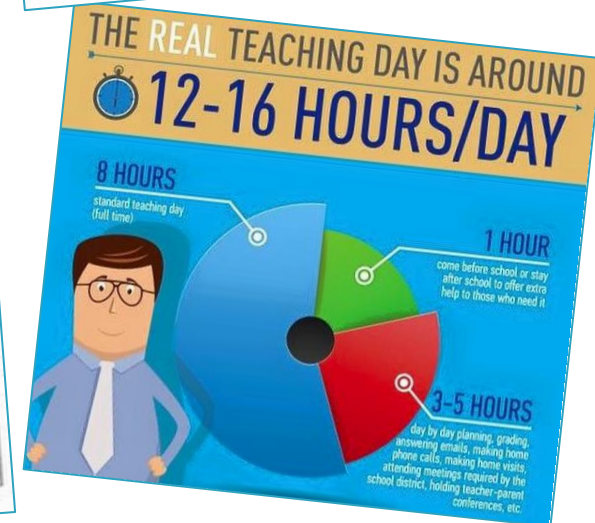
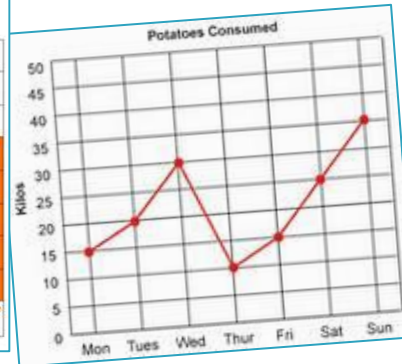
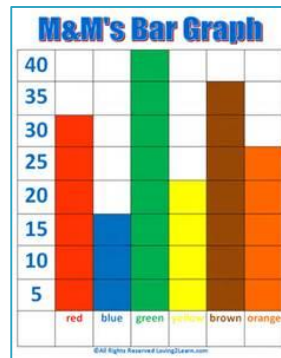
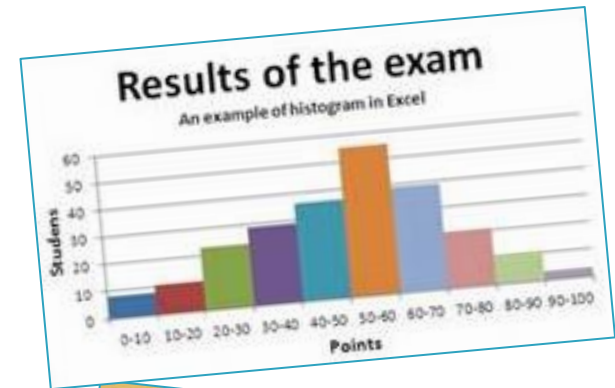
B. The necessity of training farm hands for first class farms in the fatherly handling of farm livestock is foremost in the minds of effective farm owners. Since the forefathers of the farm owners trained the farm hands for first class farms in the fatherly handling of farm livestock, the farm owners feel they should carry on with the former family tradition of training farmhands of first class farms in the effective fatherly handling of farm livestock, however futile, because of their belief it forms the basis of effective management efforts.

Graphical is a diagram that shows relationships between variables.

(1) Pie Graph

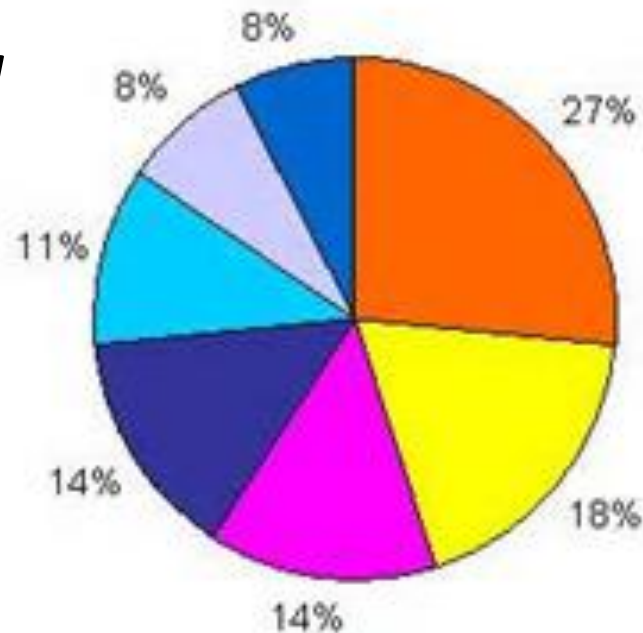
(2) Bar Graph & Histogram

(3) Frequency Polygon



PIE GRAPH

- circle graph showing the proportion of each class through either the **relative or percentage frequency**
- shows division of a whole into its parts
- more used to display ***categorical data***

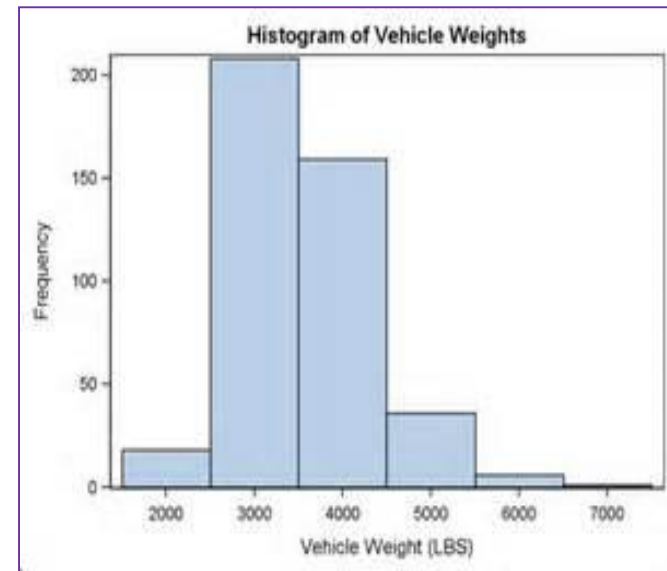
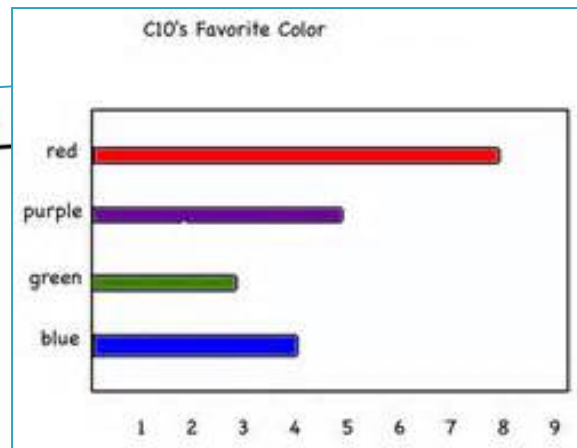
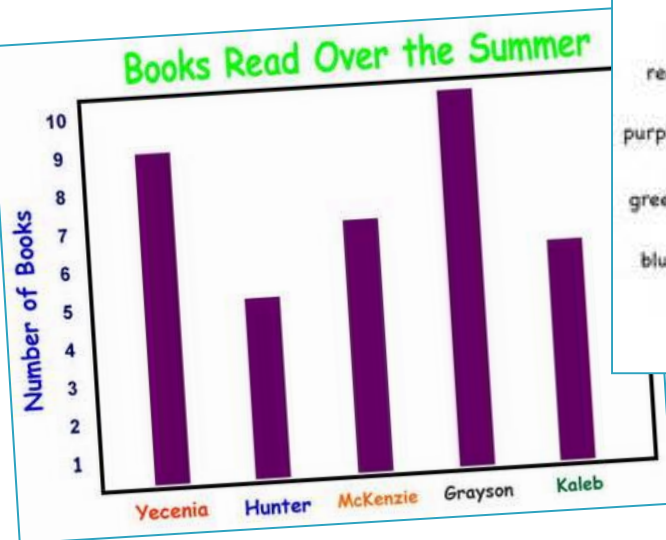


PIE GRAPH

Exercise: The table shows the highest educational attainments of the faculty of a certain university. **Construct a pie graph.**

Educational Qualifications	Frequency
BS/BSE	40
MS/MA	65
PhD/ED	50

BAR GRAPH and HISTOGRAM



BAR GRAPH

Exercise: Construct a bar graph using the data below showing the number of students enrolled in various undergraduate majors in a college of arts and sciences.

Majorships	Number of Enrollees
Psychology	650
Mathematics	600
Biology	240
PE	400
Chemistry	90
English	550

HISTOGRAM

Exercise: For 108 randomly selected college applicants, the following distribution for entrance exam scores was obtained. **Construct a Histogram.**

Lowest score = 90; Highest score = 134;
Frequencies from lowest to highest class intervals are respectively: 11, 9, 18, 15, 26, 13, 11, 5

FREQUENCY POLYGON

- graph that displays the data by using lines that connect points plotted for the **frequencies at the midpoints of the classes**
- a graphical device for understanding the shape of distribution

Exercise: Construct a frequency polygon of the data used in histogram.

Measures of Central Tendency

A decorative graphic consisting of a solid red horizontal bar that spans the width of the slide. Below this bar, on the right side, there are several thin, parallel white lines of varying lengths, creating a stepped or layered effect.

MEAN

is simply the sum of the data divided by the number of data.

$$\bar{X} = \frac{\sum x}{n}$$

Properties of MEAN

- (1) It is the most reliable.
- (2) It lends itself to higher statistical treatment.
- (3) It is the most sensitive.
- (4) It is used only if the data are interval or ratio.

Example 1: Compute Rory's

(a) 1st Long Test Average;

(b) 2nd Long Test Average; and,

(c) Grade Point Average (GPA) in the four Math subjects.

Subject	1st Long Test Score	2nd Long Test Score	Grade	Unit
Elem Statistics	65	65	79	3
College Algebra	63	99	88	3
Trigonometry	62	62	76	5
Plane Geometry	68	68	81	5

Example 2: Likert Scale

The Rosenberg Self-Esteem Scale

Circle one response for each of the following ten items.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. I feel that I am a person of worth, at least on an equal basis with others.	1	2	3	4	5
2. I feel that I have a number of good qualities.	1	2	3	4	5
3. All in all, I am inclined to feel that I am a failure.*	1	2	3	4	5
4. I am able to do things as well as most other people.	1	2	3	4	5
5. I feel I do not have much to be proud of.*	1	2	3	4	5
6. I take a positive attitude toward myself.	1	2	3	4	5
7. On the whole, I am satisfied with myself.	1	2	3	4	5
8. I wish I could have more respect for myself.*	1	2	3	4	5
9. I certainly feel useless at times.*	1	2	3	4	5
10. At times I think I am no good at all.*	1	2	3	4	5

Mean Interpretation

1.00 – 1.79	Strongly Disagree	3.40 – 4.19	Agree
1.80 – 2.59	Disagree	4.20 – 5.00	Strongly Agree
2.60 – 3.39	Neutral		



MEAN of Grouped Data

- **Midpoint Method** $\bar{X} = \frac{\sum Mf}{n}$

- **Class Deviation Method**

$$\bar{X} = AM + w \left(\frac{\sum fd}{n} \right)$$

Example 3: Ages of the passengers in the Titan ship were tallied in the table below. What is the average age of the passengers in the ship?

<i>Class Limits</i>	<i>f</i>
45-49	2
40-44	0
35-39	12
30-34	13
25-29	10
20-24	5
15-19	4
10-14	4

Individual Seatwork



Compute for D_5 and the mean of the grouped distribution using Midpoint Method. Then verify the mean value using Class deviation Method.

Score	f
98 – 104	1
91 – 97	7
84 – 90	15
77 – 83	26
70 – 76	31
63 – 69	22
56 – 62	20
49 – 55	11
42 – 48	10
35 – 41	4
28 – 34	3

MEDIAN is the point that divides the (arranged) distribution into two equal parts such that an equal number of scores fall above and below that point.

Properties of Median

1. It is NOT affected by extreme or deviant values.
2. It is used when the data are ratio or interval (in ascending/descending order).

Median of Ungrouped Data

$$\tilde{x} = \left[\frac{n+1}{2} \right]^{th} ; \text{ if } n \text{ is odd}$$

$$\tilde{x} = \frac{m_1 + m_2}{2} ; \text{ if } n \text{ is even, and } m_1 \text{ and } m_2 \text{ are the two middle values}$$

Examples:

(1)	15	18	10	14	17	22	13
-----	----	----	----	----	----	----	----

(2)	15	18	10	14	17	22	13	16
-----	----	----	----	----	----	----	----	----

(3)	15	18	14	16	22	13	16
-----	----	----	----	----	----	----	----

(4)	15	18	14	10	16	22	13	16
-----	----	----	----	----	----	----	----	----

(5)	15	18	14	10	16	22	14	16
-----	----	----	----	----	----	----	----	----

(6)	15	18	16	10	16	22	14	11	16
-----	----	----	----	----	----	----	----	----	----

Median of Grouped Data

$$\tilde{x} = LL + w \left[\frac{\frac{N}{2} - cf_b}{f} \right]$$

Example:

The Age Distribution of Patients in the ICU of
XYZ Hospital

Class Interval	f
56 – 63	8
48 – 55	10
40 – 47	12
32 – 39	6
24 – 31	3
16 – 23	1

MODE is the value that occurs with the highest frequency (and more than once).

Properties of Mode

- (1) It is used when you want to find the value which occurs most often.
- (2) It requires no calculation and it can be determined even for nominal data.
- (3) It may not exist in some set of data, or there may be more than one mode in other sets of data.

Example:

A sample of the records of a motor vehicle agency shows that 18 drivers in a certain group received

3 2 0 0 2 3 3 1 0 1 0 3 4 0 3 2 3 0

traffic tickets during the last three years. Find the mode/s and describe the distribution.

Mode of Grouped Data

(1) Delta Method

$$\hat{X} = LL + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] w$$

(2) Crude Method

(3) Refined Method

$$\hat{X} = 3\tilde{X} - 2\bar{X}$$

Example:

Find the mode of the distribution of scores on the right using

- (1) Delta Method,
- (2) Crude Method, and
- (3) Refined Method.

Score	f
98-104	1
91-97	7
84-90	15
77-83	25
70-76	32
63-69	22
56-62	20
49-55	11
42-48	10
35-41	4
28-34	3

SEATWORK



SEATWORK:

Compute for the **(1) $\tilde{X}$**
and **(2) $\hat{X}$** using

(2.1) Delta Method

(2.2) Crude Method

(2.3) Refined Method

(Note: use Class
Deviation Method
to get the mean).

Score	f
75.5 - 81.5	9
68.5 - 74.5	12
61.5 - 67.5	11
54.5 - 60.5	9
47.5 - 53.5	17
40.5 - 46.5	20
33.5 - 39.5	12
26.5 - 32.5	10