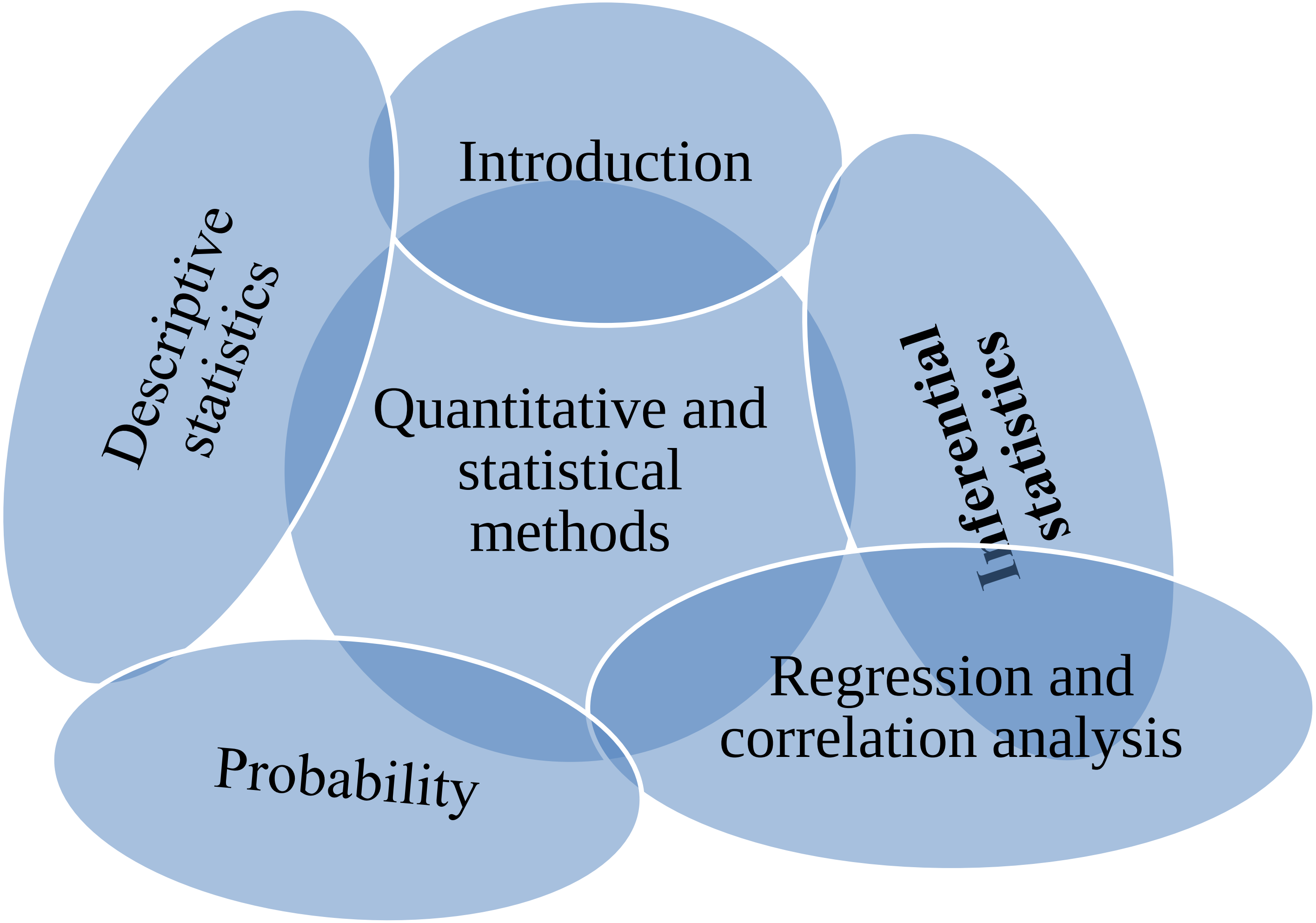
Introduction to Business Statistics

MGMT – 1071
MANKUL COLLEGE DESSIE COMPUS

HUSSIEN A. (MSC)

Plan of the course


Introduction
Descriptive statistics
Quantitative and statistical methods
Inferential statistics
Probability
Regression and correlation analysis

UNIT ONE

Introduction to Business Statistics

What Is Statistics?

1. **Collecting Data**

   e.g., Survey

2. **Presenting Data**

   e.g., Charts & Tables

3. **Characterizing Data**

   e.g., Average

Data Analysis

Why?

Decision-Making

What Is Statistics?

- **Statistics** is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information.
- Bowley has defined statistics as:

i) statistics is the science of counting,

ii) Statistics may rightly be called the science of averages,

iii) statistics is the science of measurement of social organism regarded as a whole in all its manifestations

Characteristics of Statistics

i) Statistics are the aggregates of facts. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

ii) Statistics are affected by a number of factors. For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers ...

iii) Statistics must be reasonably accurate. Wrong figures, if analysed, will lead to erroneous conclusions.

Characteristics of Statistics

iv) Statistics must be collected in a systematic manner. If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

v) Collected in a systematic manner for a pre-determined purpose

vi) Lastly, Statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

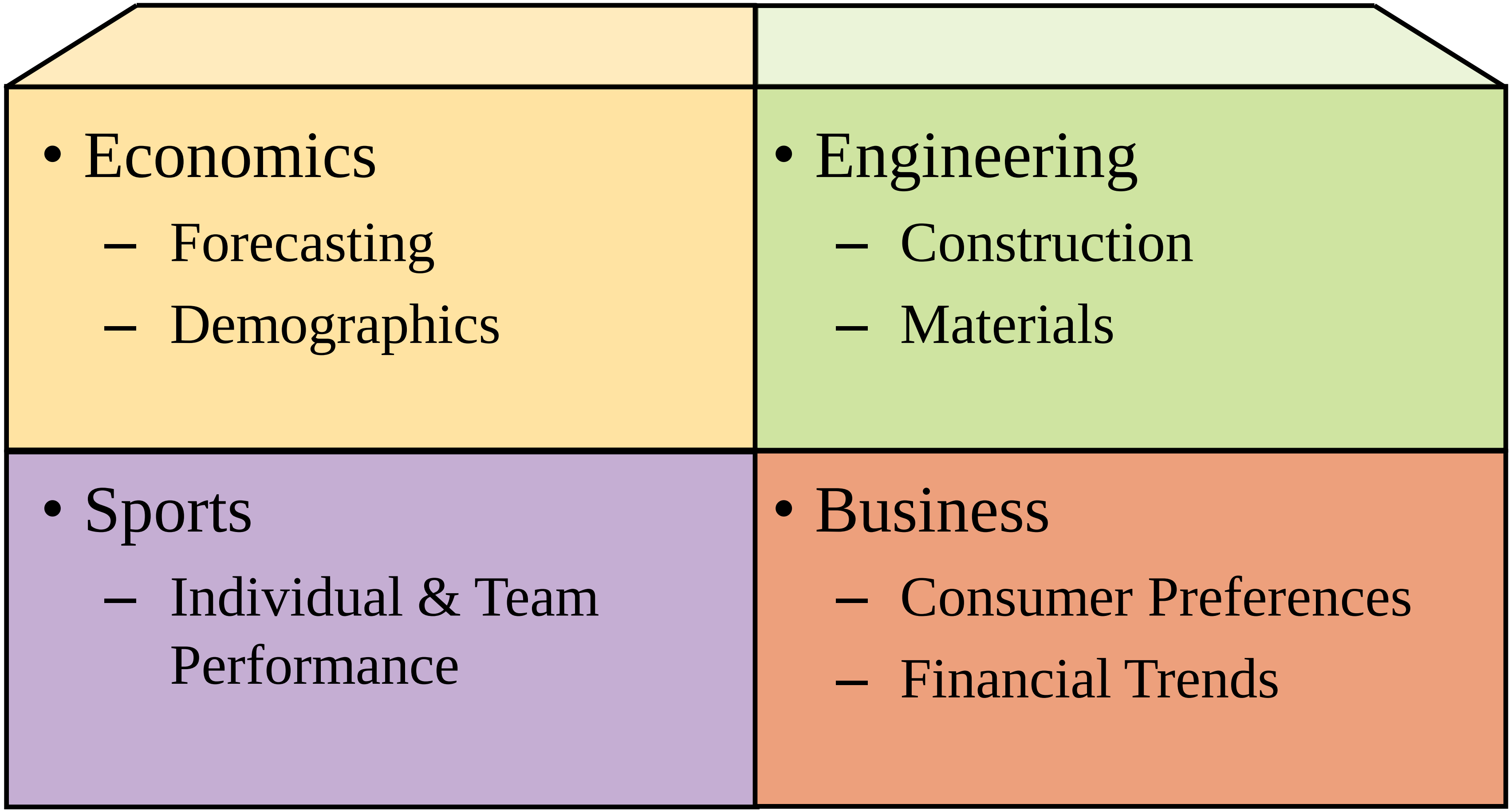
Why study statistics?

1. Data are everywhere
2. Statistical techniques are used to make many decisions that affect our lives
3. No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively

Applications of statistical concepts in the business world

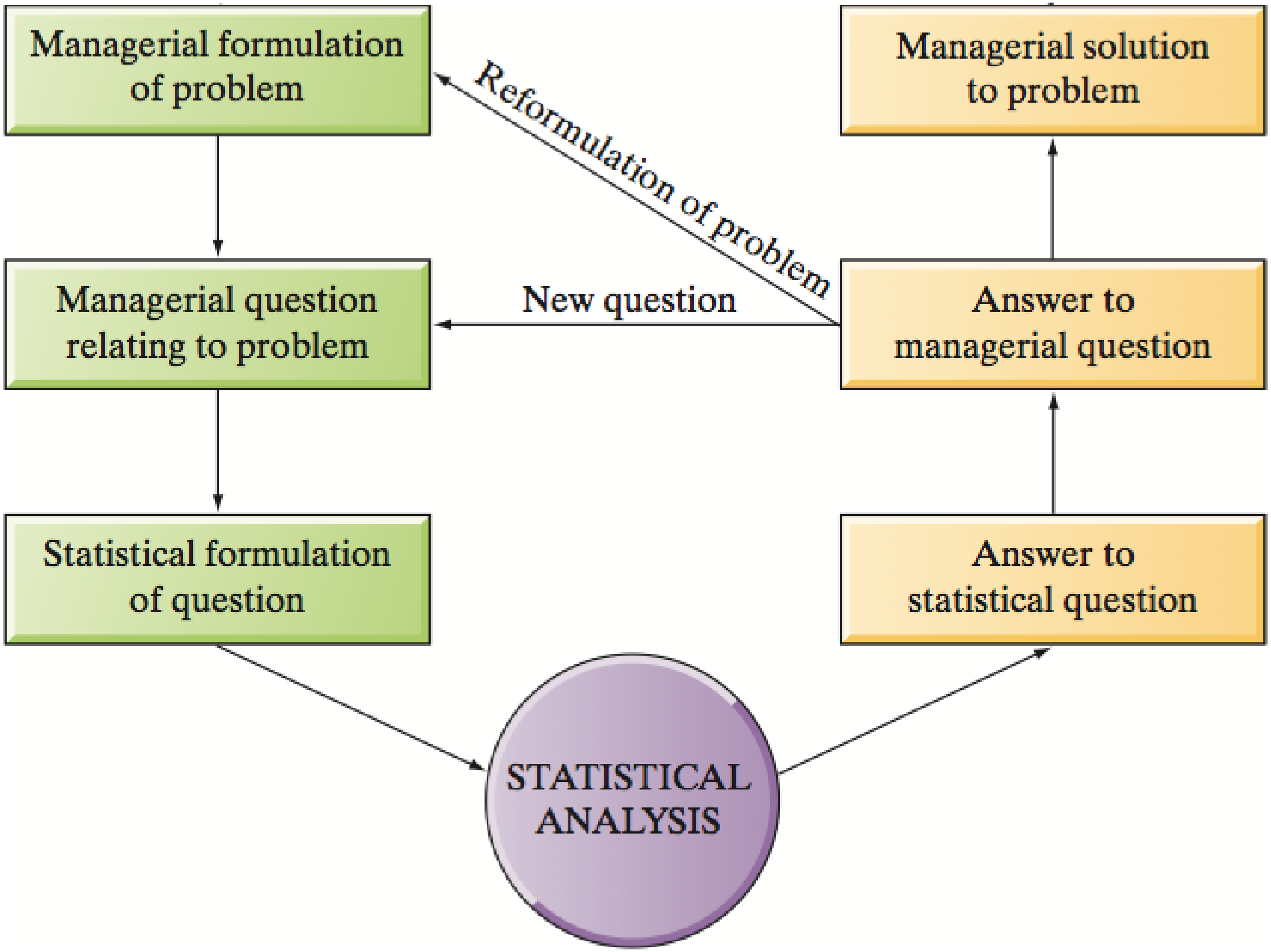
- Finance – correlation and regression, index numbers, time series analysis
- Marketing – hypothesis testing, chi-square tests, nonparametric statistics
- Personel – hypothesis testing, chi-square tests, nonparametric tests
- Operating management – hypothesis testing, estimation, analysis of variance, time series analysis

Application Areas

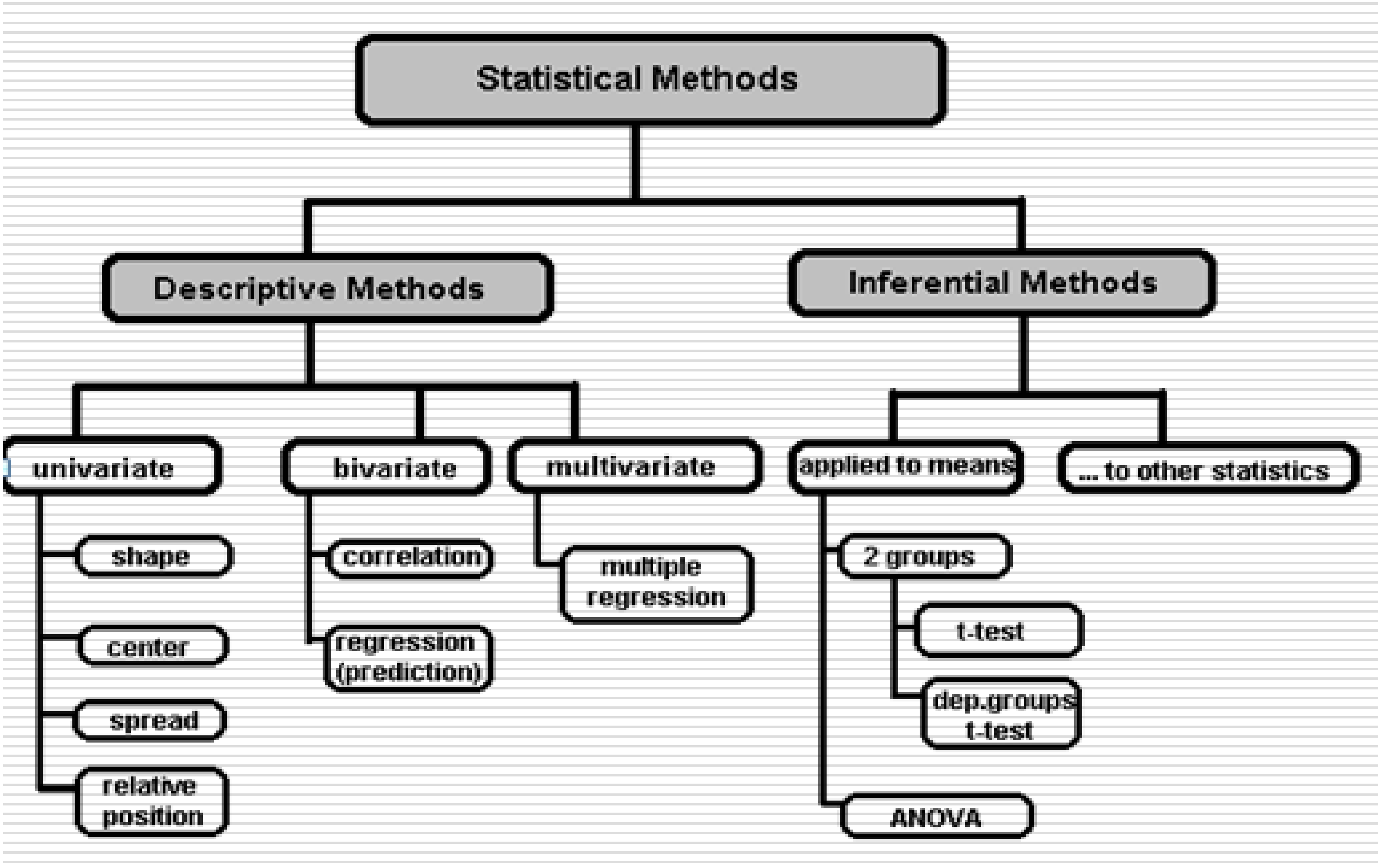


- Economics
  - Forecasting
  - Demographics
- Engineering
  - Construction
  - Materials
- Sports
  - Individual & Team Performance
- Business
  - Consumer Preferences
  - Financial Trends

Real-World Problem


Managerial formulation of problem
Reformulation of problem
Managerial solution to problem
Managerial question relating to problem
New question
Answer to managerial question
Statistical formulation of question
Answer to statistical question
STATISTICAL ANALYSIS

TAXONOMY OF STATISTICS



Statistical Methods

Descriptive Methods

Inferential Methods

univariate

bivariate

multivariate

applied to means

... to other statistics

shape

center

spread

relative position

correlation

regression (prediction)

multiple regression

2 groups

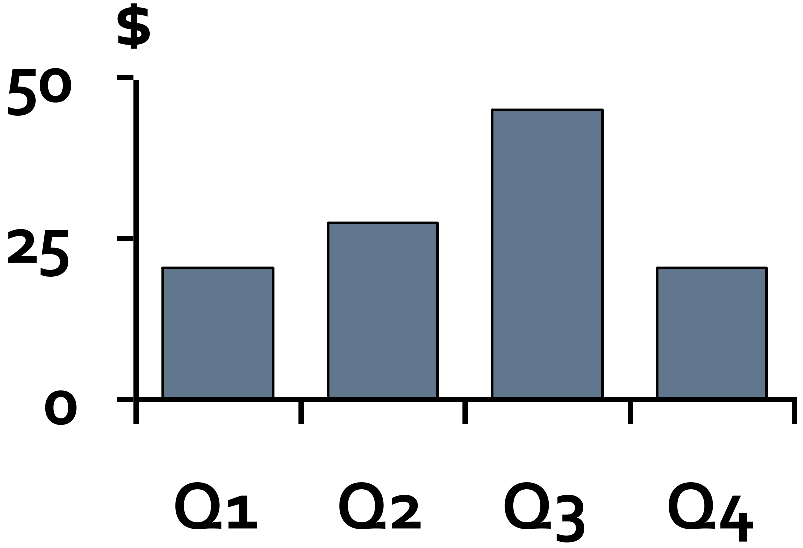
t-test

dep.groups t-test

ANOVA

Descriptive Statistics

1. Involves
   - Collecting Data
   - Presenting Data
   - Characterizing Data
2. Purpose
   - Describe Data



$\bar{X} = 30.5 \quad S^2 = 113$

Inferential Statistics

1. Involves
   - Estimation
   - Hypothesis Testing
2. Purpose
   - Make decisions about population characteristics
   - **Inference is the process of drawing conclusions or making decisions about a population based on sample results**

Types of statistics

- **Descriptive statistics** – Methods of organizing, summarizing, and presenting data in an informative way
- **Inferential statistics** – The methods used to determine something about a population on the basis of a sample
  - Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
  - Sample – A portion, or part, of the population of interest

TYPES Statistical data

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the entire research project.
- Statistical data are usually obtained by counting or measuring items.
    - **Primary data** are collected specifically for the analysis desired
    - **Secondary data** have already been compiled and are available for statistical analysis
- A **variable** is an item of interest that can take on many different numerical values.
- A **constant** has a fixed numerical value.

Data

Statistical data are usually obtained by counting or measuring items. Most data can be put into the following categories:

- **Qualitative -** data are measurements that each fail into one of several categories. (hair color, ethnic groups and other attributes of the population)
- **quantitative -** data are observations that are measured on a numerical scale (distance traveled to college, number of children in a family, etc.)

Qualitative data

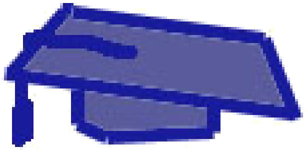
- ✓ Qualitative data are generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense tofind an average hair color or blood type.
- ✓ Qualitative data can be separated into two subgroups:

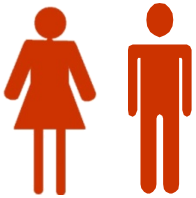
- **dichotomic** (if it takes the form of a word with two options (gender - male or female)
- **polynomic** (if it takes the form of a word with more than two options (education - primary school, secondary school and university).

Qualitative Data

Classified into categories.

- College major of each student in a class.
- Gender of each employee at a company.
- Method of payment (cash, check, credit card).





Quantitative data

- Quantitative data are always numbers and are the **result of counting or measuring** attributes of a population.
- Quantitative data can be separated into two subgroups:
- **discrete** (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf, ...)
- **continuous** (if it is the result of *measuring* (distance traveled, weight of luggage, …)

Quantitative Data

Measured on a numeric scale.

- Number of defective items in a lot.
- Salaries of CEOs of oil companies.
- Ages of employees at a company.

4

943

52

21

12

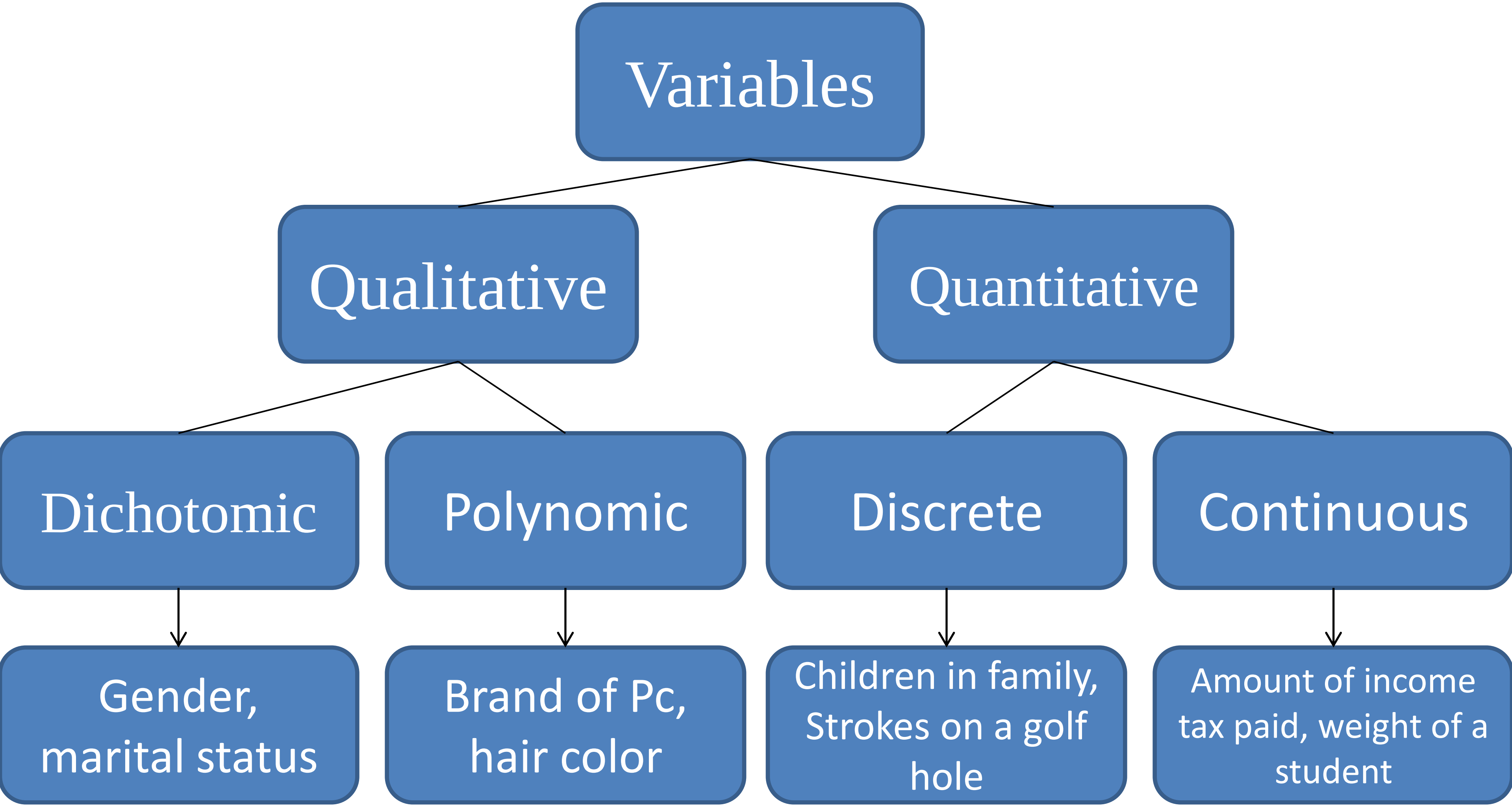
120

8

71

3

Types of variables



Variables

Qualitative

Quantitative

Dichotomic

Polynomic

Discrete

Continuous

Gender, marital status

Brand of Pc, hair color

Children in family, Strokes on a golf hole

Amount of income tax paid, weight of a student

Numerical scale of measurement:

- **Nominal** – consist of categories in each of which the number of respective observations is recorded. The categories are in no logical order and have no particular relationship. The categories are said to be ***mutually exclusive*** since an individual, object, or measurement can be included in only one of them.
- **Ordinal** – contain more information. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating-excelent, good, fair, poor)
- **Interval** – is a set of numerical measurements in which the distance between numbers is of a known, sonstant size.
- **Ratio** – consists of numerical measurements where the distance between numbers is of a known, constant size, in addition, there is a nonarbitrary zero point.

Importance of statistics in business

- There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) The planning of operations: This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) The setting up of standards: This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

(iii) The function of control: This involves comparison of actual production achieved against the norm or target set earlier. In case the production has allen short of the target, it gives remedial measures so that such a deficiency does not occur again

Limitations of statistics in business

i) There are certain phenomena or concepts where statistics cannot be used. This is because these phenomena or concepts are not amenable to measurement. For example, beauty, intelligence, courage cannot be quantified. Statistics has no place in all such cases where quantification is not possible.

ii) Statistics reveal the average behaviour, the normal or the general trend. An application of the 'average' concept if applied to an individual or a particular situation may lead to a wrong conclusion and sometimes may be disastrous.

- For example, one may be misguided when told that the average depth of a river from one bank to the other is four feet, when there may be some points in between where its depth is far more than four feet. On this understanding, one may enter those points having greater depth, which may be hazardous

Limitations of statistics in business

iii) Since statistics are collected for a particular purpose, such data may not be relevant or useful in other situations or cases. For example, secondary data (i.e., data originally collected by someone else) may not be useful for the other person.

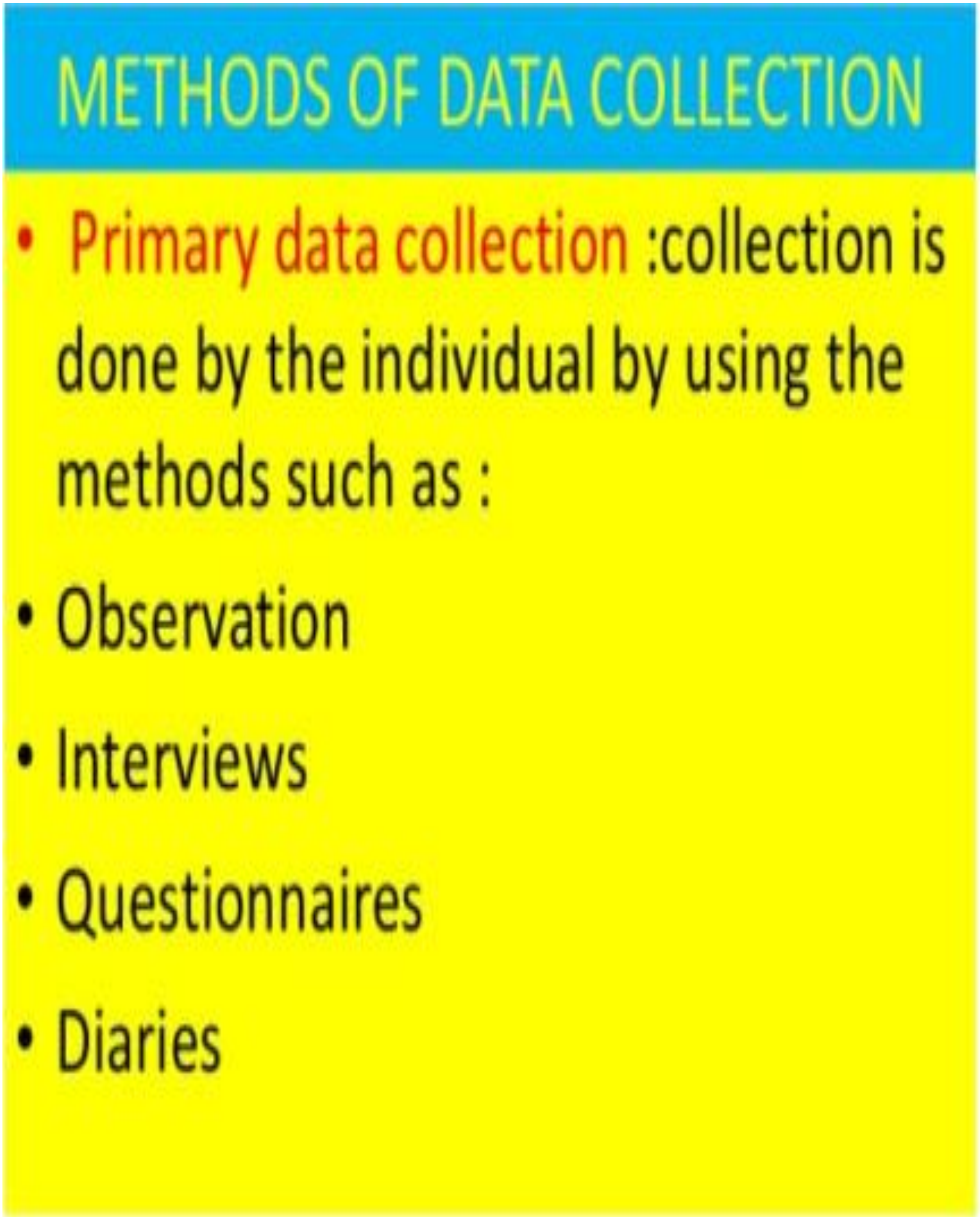
iv) Statistics are not 100 per cent precise as is Mathematics or Accountancy. Those who use statistics should be aware of this limitation.

v) In statistical surveys, sampling is generally used as it is not physically possible to cover all the units or elements comprising the universe. The results may not be appropriate as far as the universe is concerned. Moreover, different surveys based on the same size of sample but different sample units may yield different results.

UNIT 2

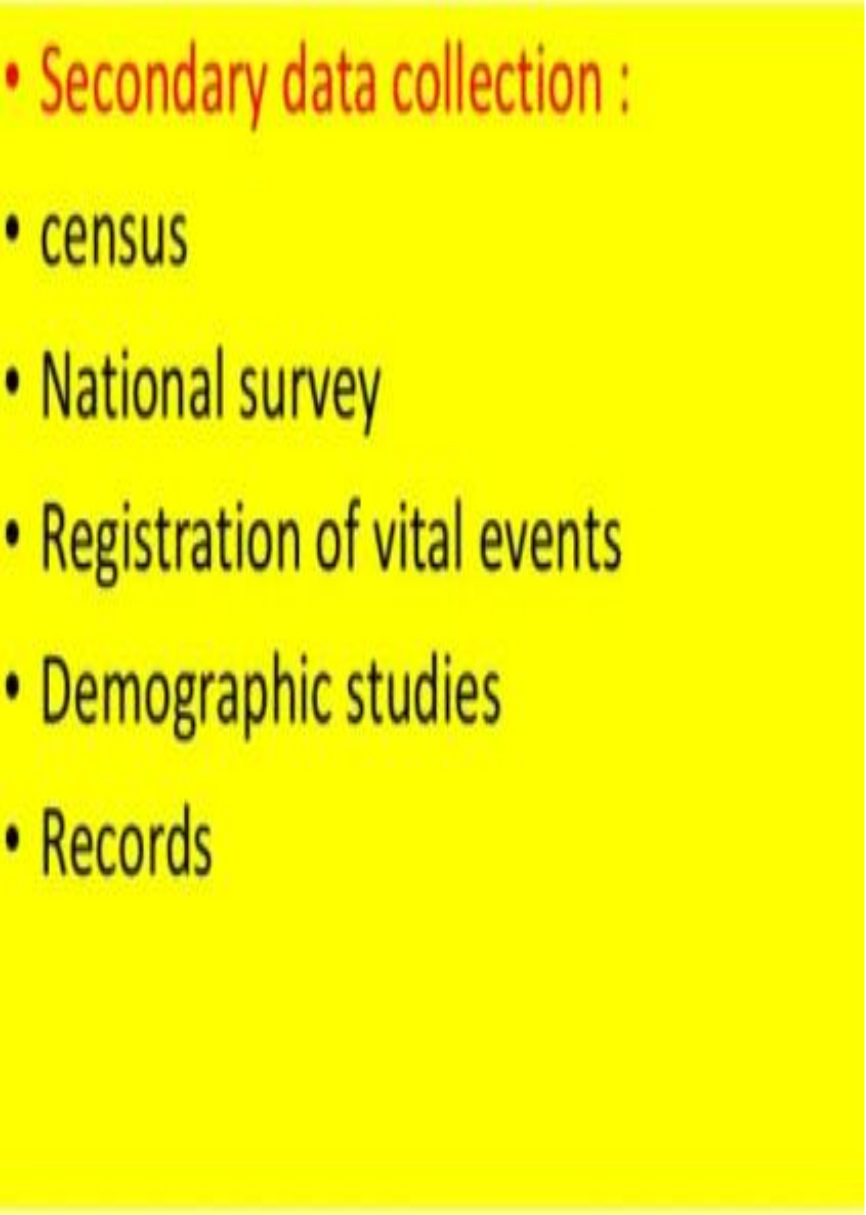
DATA COLLECTING AND PRESENTATION

Methods of data collection



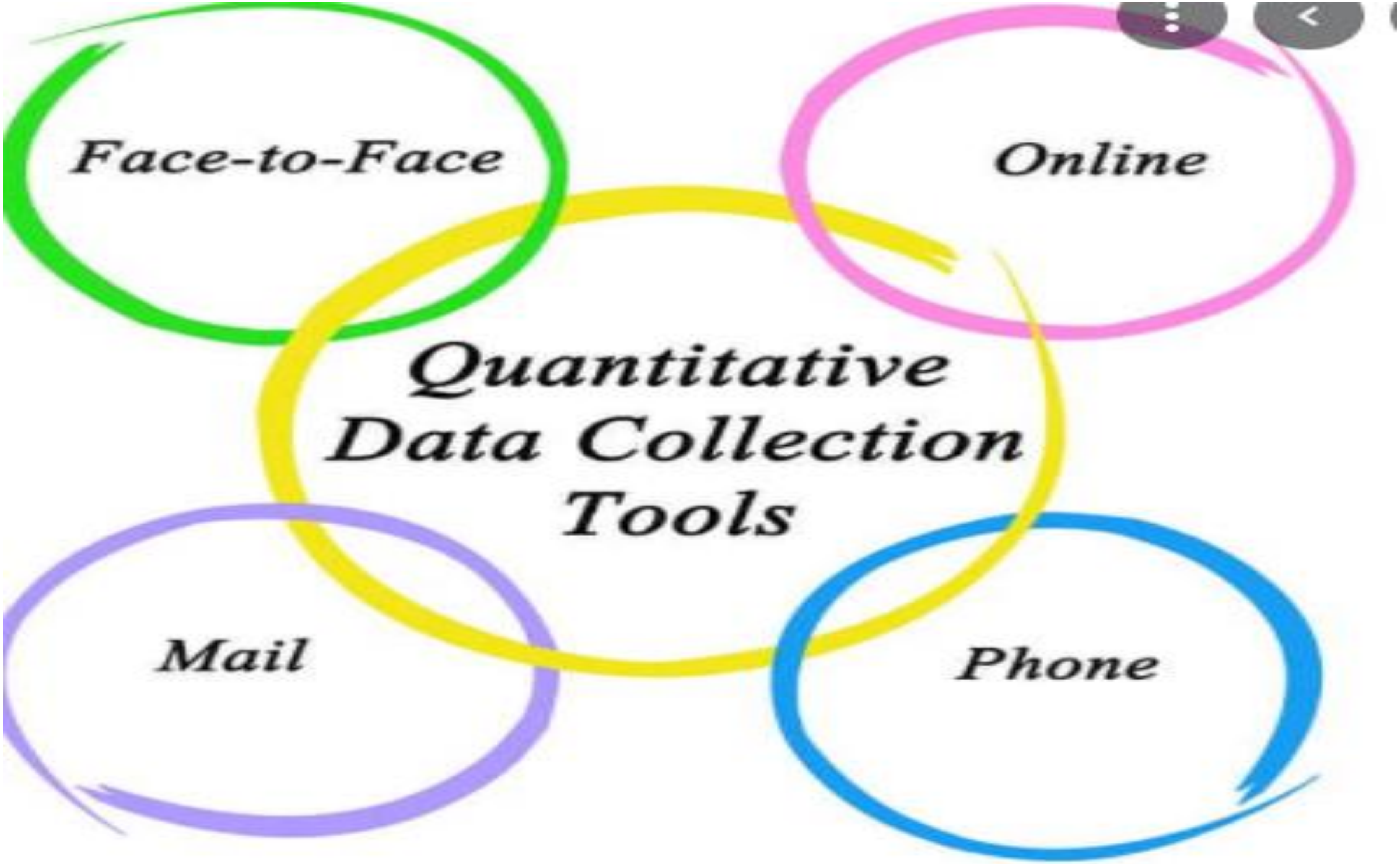
METHODS OF DATA COLLECTION

- Primary data collection :collection is done by the individual by using the methods such as :
- Observation
- Interviews
- Questionnaires
- Diaries

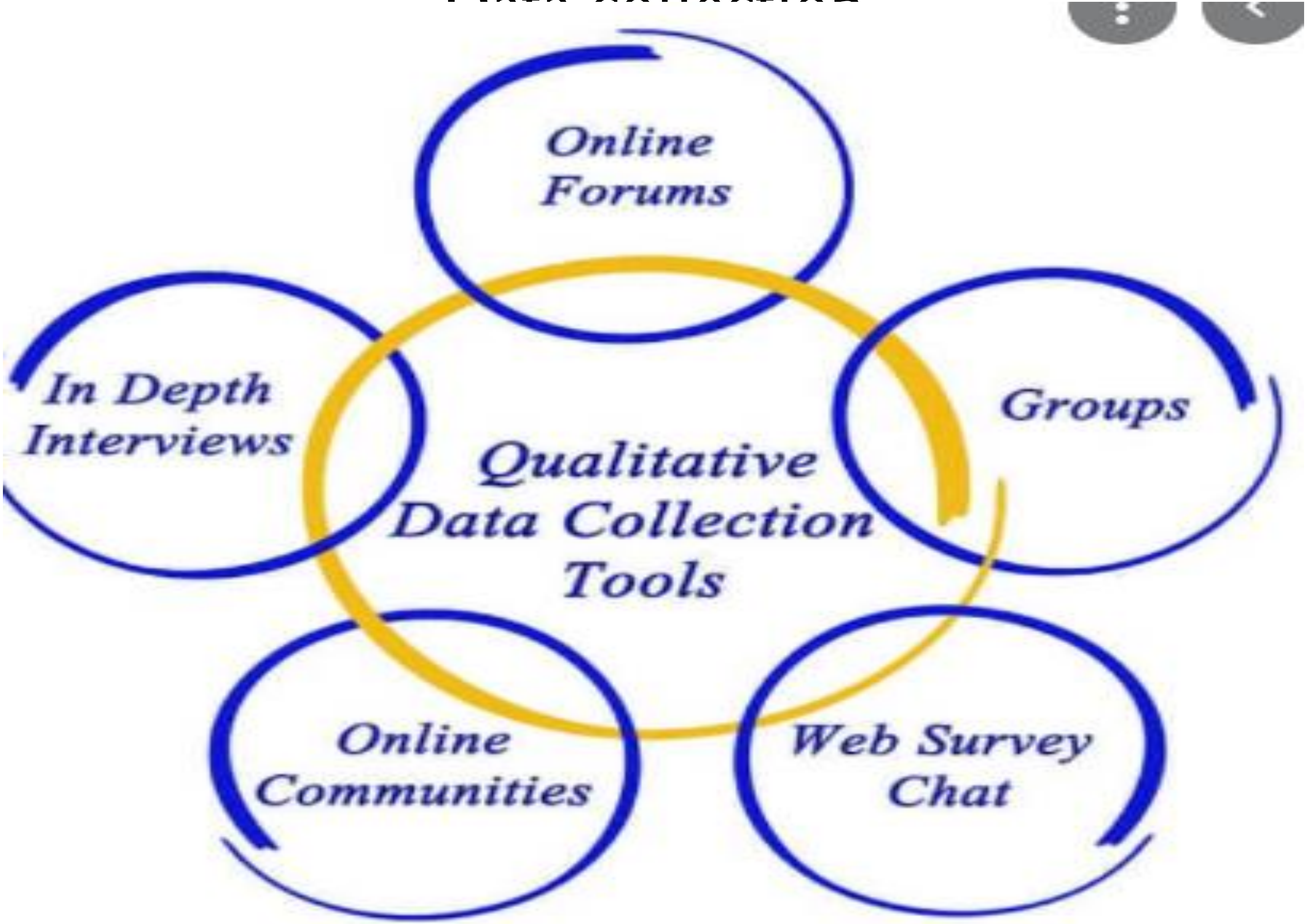


- Secondary data collection :
- census
- National survey
- Registration of vital events
- Demographic studies
- Records

Data collection



Data collection



Data presentation methods

„ The question is“ said Alice, „whether you can make words mean so many different things.“

„The question is,“ said Humpty Dumpty, „which is to be master-that´s all.“ (Lewis Carroll)

Describing categorical variables

- Table of frequency distributions
  - Frequency
  - Relative frequency
  - Cumulative frequencies
  - Relative cumulative frequency
- Charts
  - Bar charts
  - Pie charts

Frequency distributions

- Simple and effective way of summarizing categorical data
- Done by counting the number of observations falling into each of the categories or levels of the variables.

E.g. birth weight with levels ‘Very low ’, ‘Low’, ‘Normal’ and ‘big’.

- The frequency distribution for newborns is obtained simply by counting the number of newborns in each birth weight category.

Relative Frequency

- It is the proportion or percentages of observations in each category.
- The distribution of proportions is called the ***relative frequency*** distribution of the variable
- Given a total number of observations, the relative frequency distribution is easily derived from the frequency distribution.
- Conversion in the opposite direction is also possible, but the conversion is often inaccurate because of rounding

Cumulative frequency

- It is the number of observations in the category plus observations in all categories smaller than it.

Cumulative relative frequency

- It is the proportion of observations in the category plus observations in all categories smaller than it.
- It is obtained by dividing the cumulative frequency by the total number of observations.

Table 1. Distribution of birth weight of newborns between 1976-1996 at DBRH.

BWT	Freq.	Cum. Freq	Rel.Freq(%)	Cum.rel.freq.(%)
Very low	43			
Low	793			
Normal	8870			
Big	268			
Total	9974			

b) Describing Quantitative variable:

- Table of frequency distributions
  - Frequency
  - Relative frequency
  - Cumulative frequencies

- Select a set of continuous, non-overlapping intervals such that each value can be placed in one and only one of the intervals.

- The first consideration is how many intervals to include

To determine the number of class intervals and the corresponding width, we may use:

Sturge’s rule:

$$K = 1 + 3.322(\log n)$$

K

$$W = \frac{L - S}{K}$$

W

K

where

K = number of class intervals

n = no. of observations

W = width of the class interval

L = the largest value

S = the smallest value

Example:

Leisure time (hours) per week for 40 college students:

23 24 18 14 20 36 24 26 23 21 16 15 19 20 22 14 13

10 19 27 29 22 38 28 34 32 23 19 21 31 16 28 19 18

12 27 15 21 25 16

K **= ???????**

Maximum value = ???, Minimum value = ???

Width = ???

Time (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
10-14	5		
15-19	11		
20-24	12		
25-29	7		
30-34	3		
35-39	2		
Total	40		

- Class Limit: The range for each class
  - Upper class limit
  - Lower class limit

- Mid-point (class mark): The value of the interval which lies midway between the lower and the upper limits of a class.
- Class boundary (True limits): Are those limits that make an interval of a continuous variable continuous in both directions
  - Upper class boundary
  - Lower class boundary
- Subtract 0.5 from the lower and add it to the upper class limit

Time (Hours)	True limit(class boundary)	Mid-point	Frequency
10-14	9.5 – 14.5		
15-19	14.5 – 19.5		
20-24	19.5 – 24.5		
25-29	24.5 – 29.5		
30-34	29.5 – 34.5		
35-39	34.5 - 39.5		
Total			

Guidelines for constructing tables

- Keep them simple
- Limit the number of variables to three or less
- All tables should be self-explanatory
- Include clear title telling what, when and where
- Clearly label the rows and columns
- State clearly the unit of measurement used
- Explain codes and abbreviations in the foot-note
- Show totals
- If data is not original, indicate the source in foot-note.

Diagrammatic Representation

Pictorial representations of Statistical data

Importance of diagrammatic representation

1. Diagrams have greater attraction than mere figures.

2. They give quick overall impression of the data.

3. They have great memorizing value than mere figures.

4. They facilitate comparison

5. Used to understand patterns and trends

Specific types of graphs include:

- Bar graph
- Pie chart

Nominal, ordinal, Discrete data

- Histogram
- Frequency polygon
- Cum. Freq. polygon
- Line graph
- Others

Quantitative continuous data

Sources of Data

Published source:

book, journal, newspaper, Web site

Designed experiment:

Researcher exerts strict control over units

Survey:

A group of people are surveyed and their responses are recorded

Observation study:

Units are observed in natural setting and variables of interest are recorded

Numerical presentation of qualitative data

- **pivot table** (qualitative dichotomic statistical attributes)
- **contingency table** (qualitative statistical attributes from which at least one of them is polynomic)

You should know how to convert absolute values to relative ones (%).

Frequency distributions – numerical presentation of quantitative data

- Frequency distribution – shows the frequency, or number of occurences, in each of several categories. Frequency distributions are used to summarize large volumes of data values.
- When the raw data are measured on a qunatitative scale, either interval or ration, categories or classes must be designed for the data values before a frequency distribution can be formulated.

Steps for constructing a frequency distribution

1. Determine the number of classes $m = \sqrt{n}$
2. Determine the size of each class $h = \frac{(\max - \min)}{m}$
3. Determine the starting point for the first class
4. Tally the number of values that occur in each class
5. Prepare a table of the distribution using actual counts and/ or percentages (relative frequencies)

Frequency table

- **absolute frequency “n_i”** (Data Tab→Data Analysis→Histogram)
- **relative frequency “f_i”**

Cumulative frequency distribution shows the total number of occurrences that lie above or below certain key values.

- **cumulative frequency “N_i”**
- **cumulative relative frequency “F_i”**

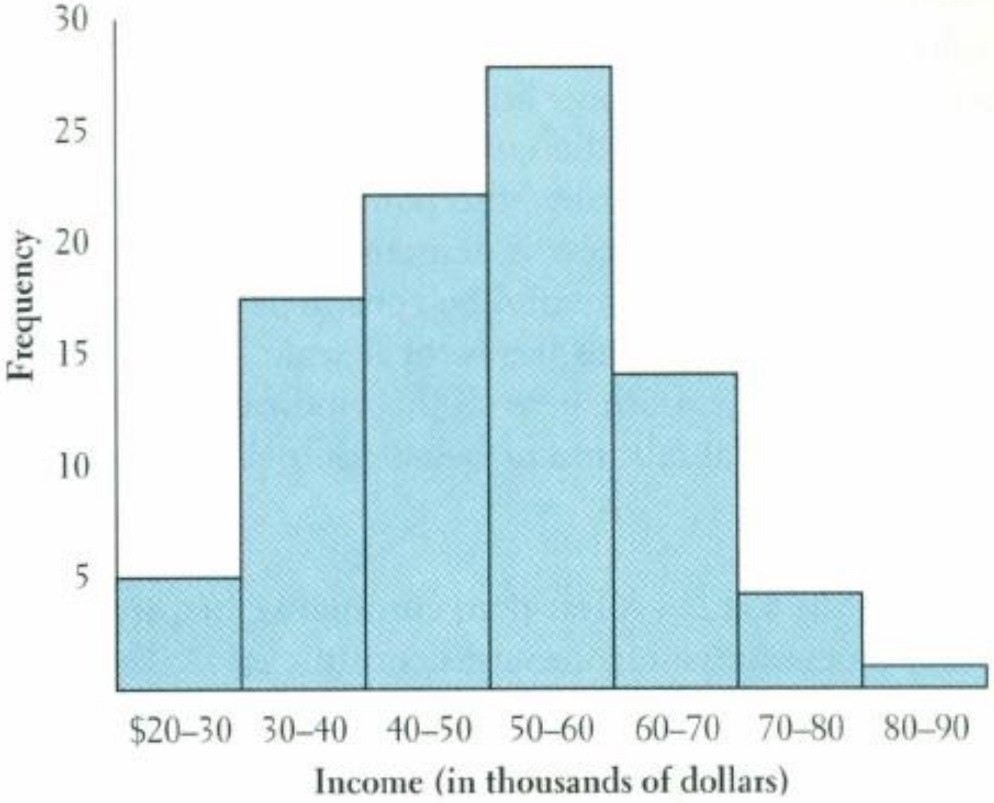
Charts and graphs

- Frequency distributions are good ways to present the essential aspects of data collections in concise and understable terms
- Pictures are always more effective in displaying large data collections

Histogram

- Frequently used to graphically present interval and ratio data
- Is often used for interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

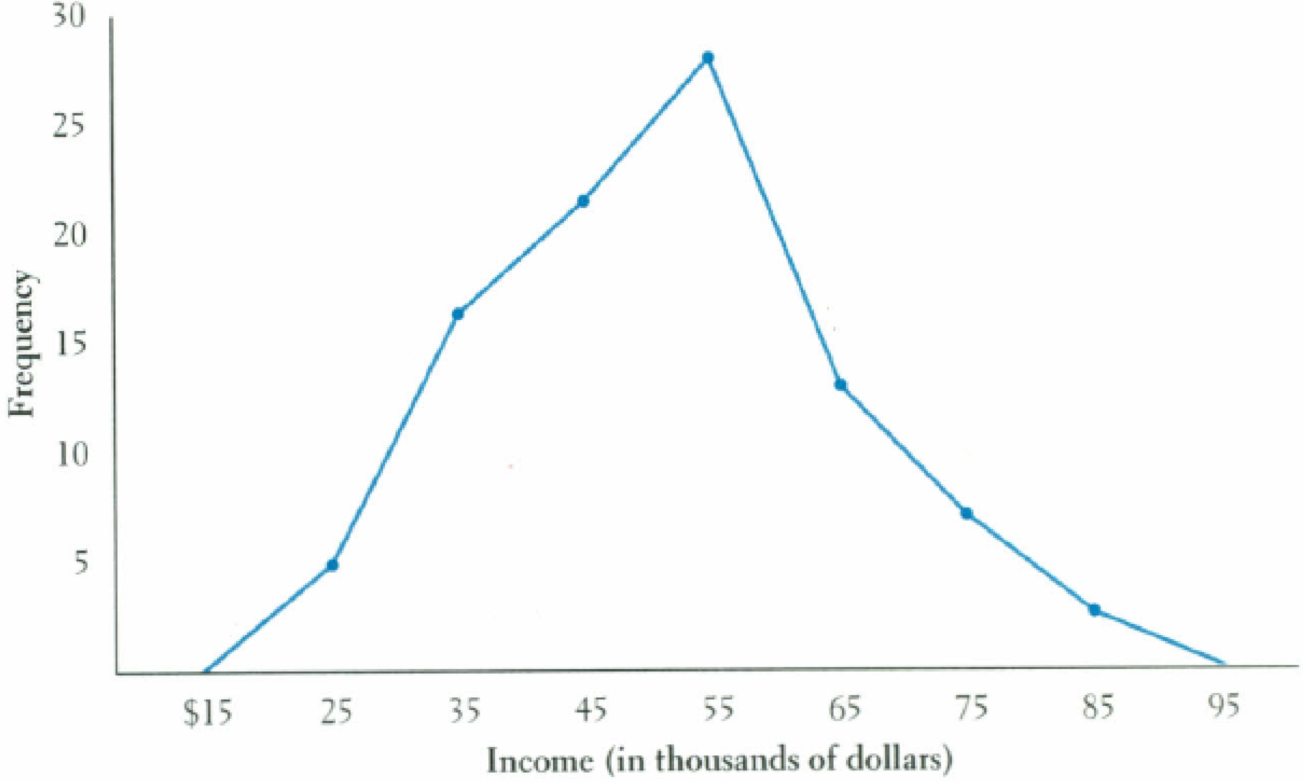
FIGURE 3.7 Histogram—Executive Incomes for the Sunrunner Corporation



Frequency polygon

- Another common method for graphically presenting interval and ratio data
- To construct a frequency polygon mark the frequencies on the vertical axis and the values of the variable being measured on the horizontal axis, as with the histogram.
- If the purpose of presenting is comparation with other distributions, the frequency polygon provides a good summary of the data

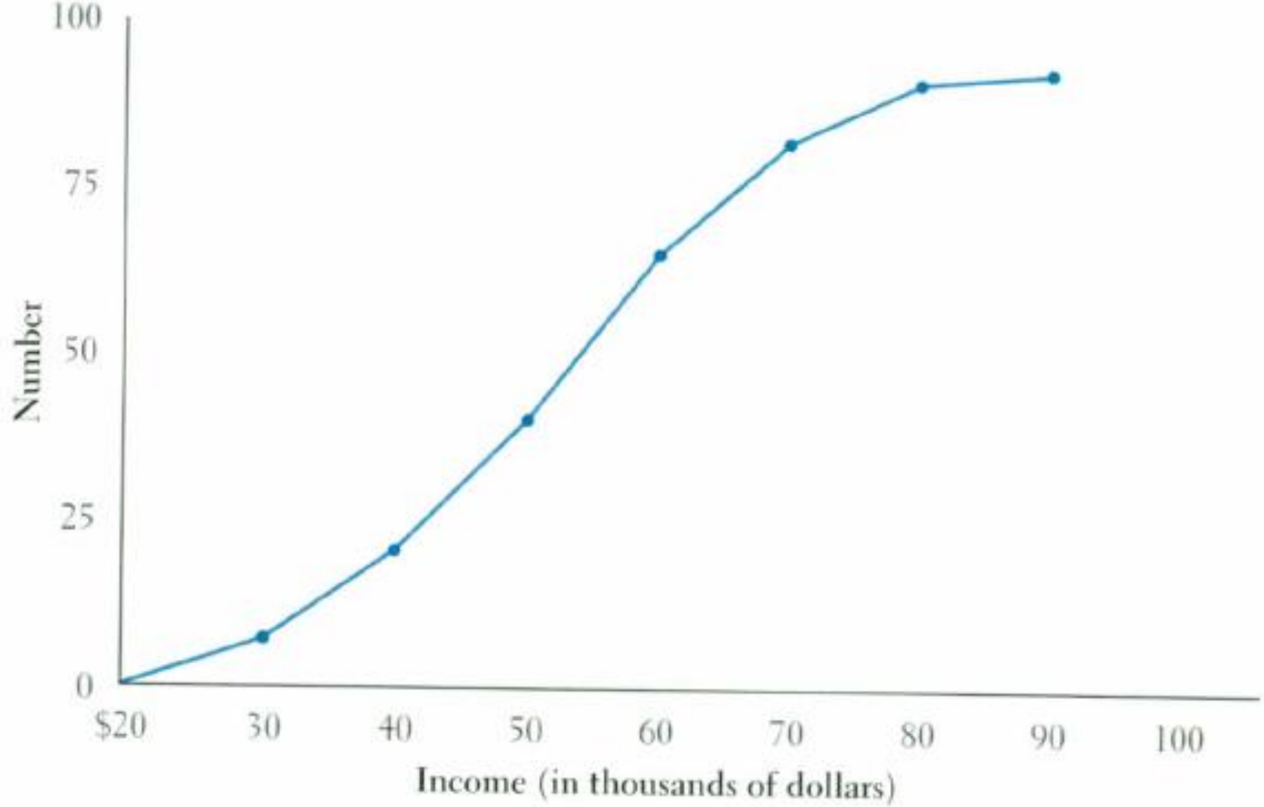
FIGURE 3.8 Frequency Polygon—Executive Incomes



Ogive **(Cumulative Frequency Polygon)**

- A graph of a cumulative frequency distribution
- Ogive is used when one wants to determine how many observations lie above or below a certain value in a distribution.
- First cumulative frequency distribution is constructed
- Cumulative frequencies are plotted at the upper class limit of each category
- Ogive can also be constructed for a relative frequency distribution.

FIGURE 3.9 Ogive—Executive Incomes (frequencies)

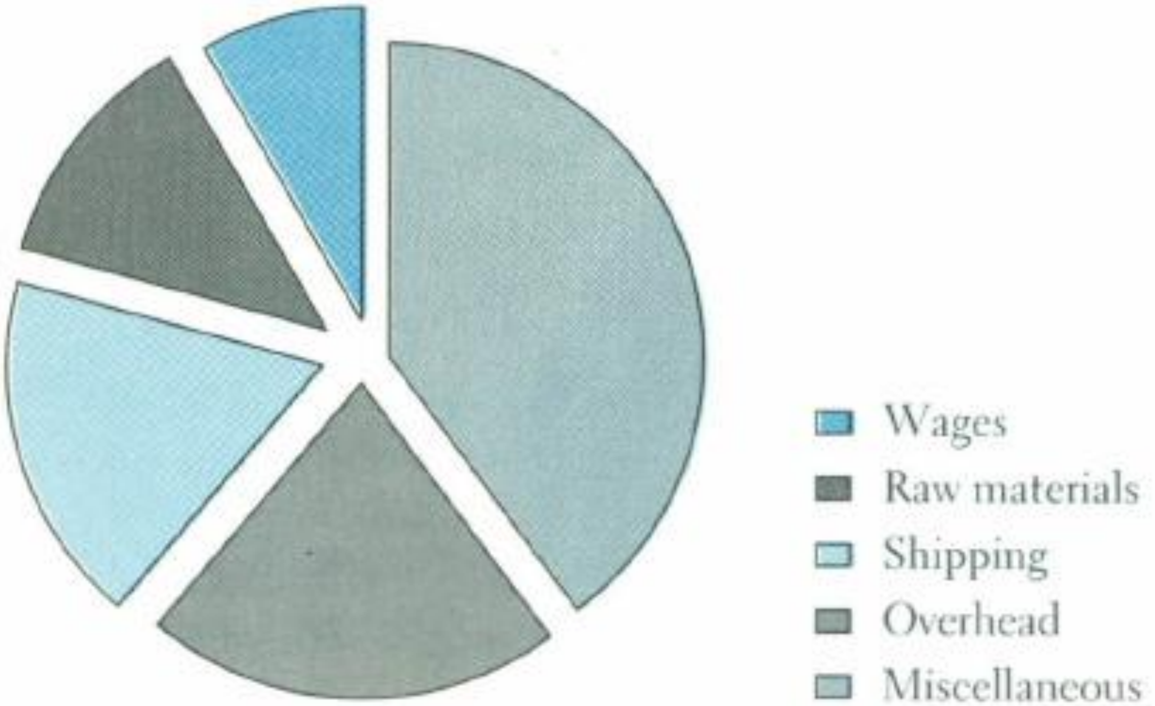


Pie Chart

- The pie chart is an effective way of displaying the percentage breakdown of data by category.
- Useful if the relative sizes of the data components are to be emphasized
- Pie charts also provide an effective way of presenting ratio- or interval-scaled data after they have been organized into categories

Pie Chart

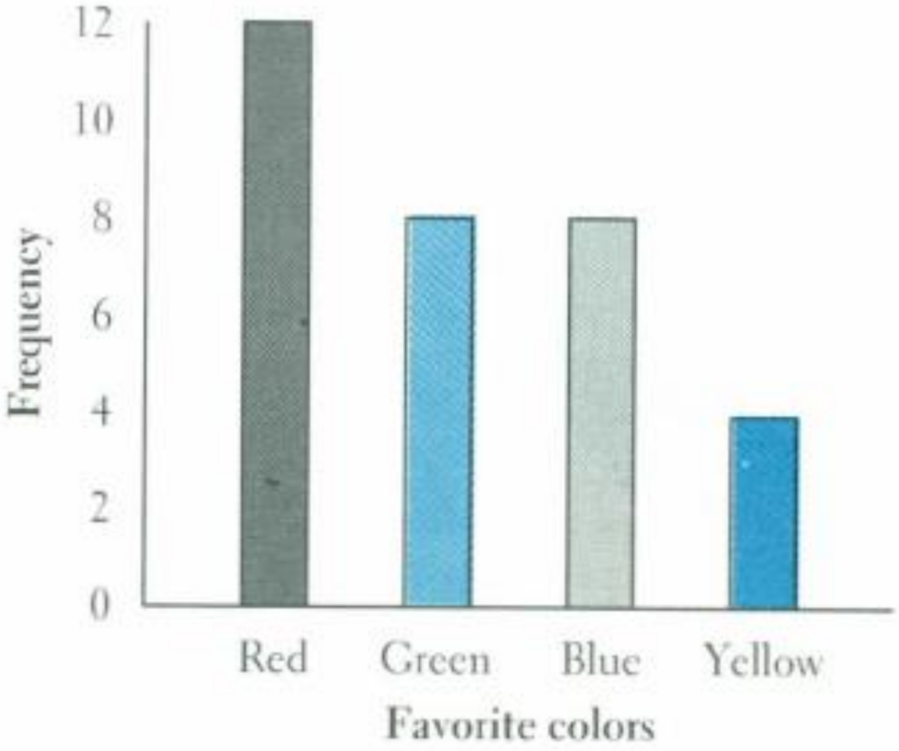
FIGURE 3.3 Pie Chart—Expenditures of Funds for Itrex Company



Bar chart

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data – the separation emphasize the plotting of frequencies for distinct categories

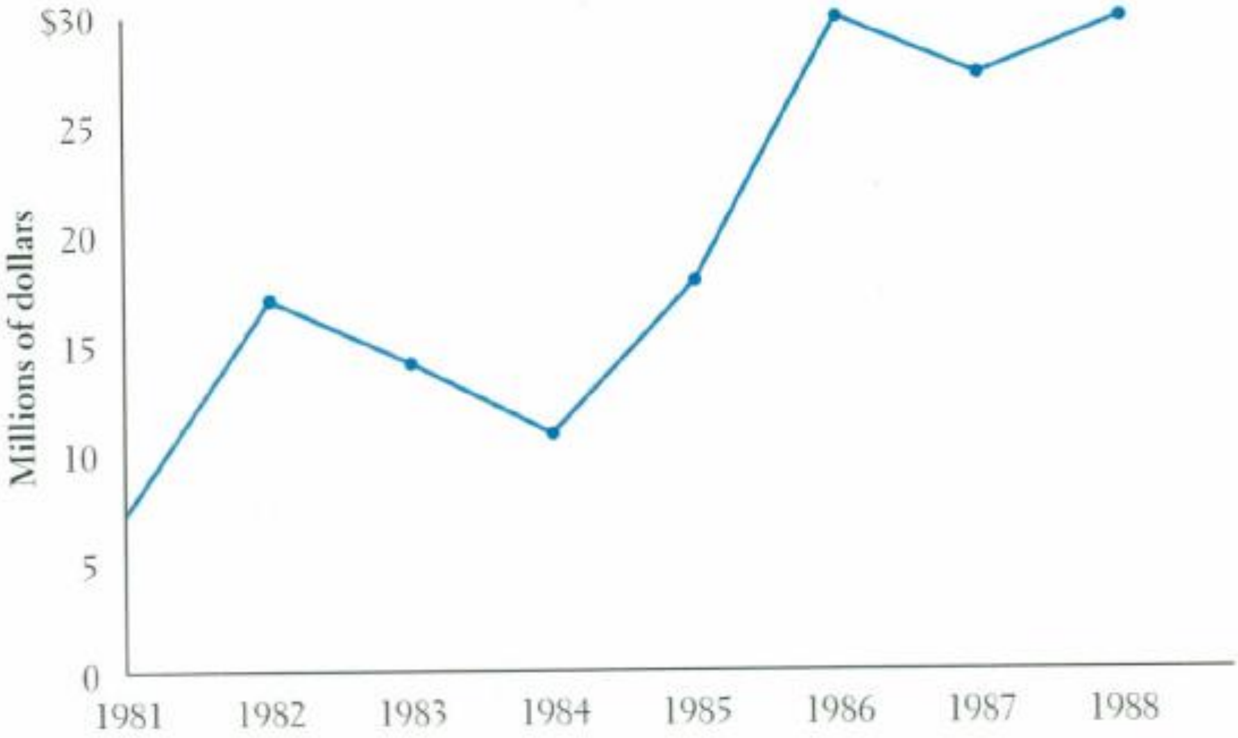
FIGURE 3.4 Bar Chart—Favorite Colors of 32 People



Time Series Graph

- The time series graph is a graph of data that have been measured over time.
- The horizontal axis of this graph represents time periods and the vertical axis shows the numerical values corresponding to these time periods

FIGURE 3.13 Time Series Graph—Corporate Revenue, Flightcraft Corp.



Statistical graphs of data

- A picture is worth a thousand words!

- Graphs for numerical data:

  Histograms

  Frequency polygons

  Pie

- Graphs for categorical data

  Bar graphs

  Pie

UNIT 3

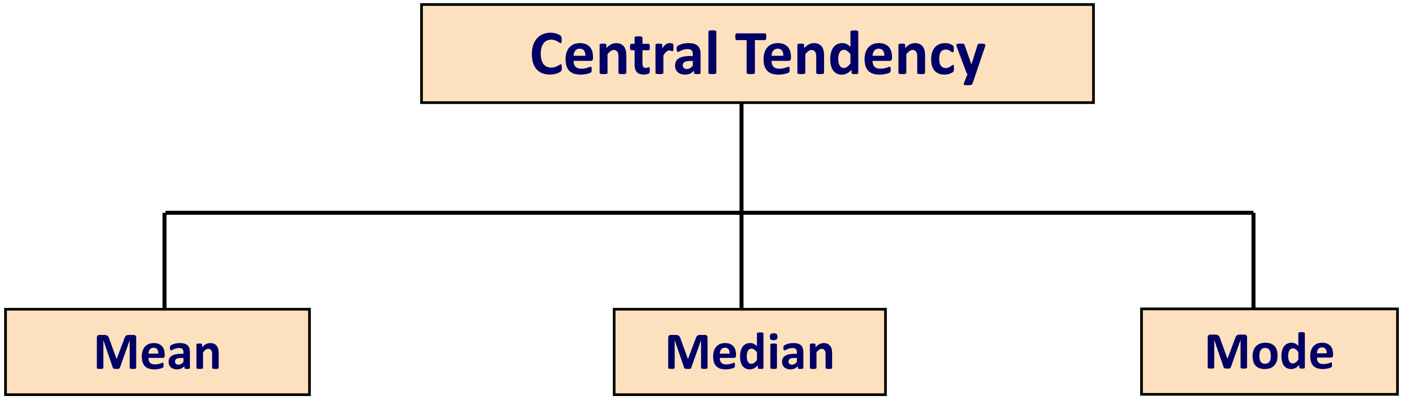
Methods of Centeral cendency and Measurement

Central value

- Center measurement - is a summary measure of the overall level of a dataset
- Give information concerning the average or typical score of a number of scores
  - mean
  - median
  - mode

Measures of Central Tendency

Overview



Central Tendency

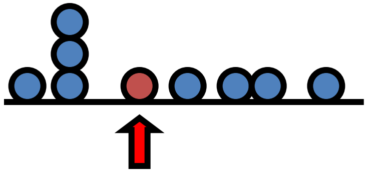
Mean

Median

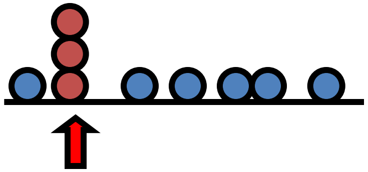
Mode

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

Arithmetic average



Midpoint of ranked values



Most frequently observed value

Central value: The Mean

- The Mean is a measure of *central value*
  - What most people mean by “average”
  - Sum of a set of numbers divided by the number of numbers in the set

$$\frac{1+2+3+4+5+6+7+8+9+10}{10}=\frac{55}{10}=5.5$$

Central value: The Mean

Arithmetic average:

Sample

$$\overline{X} = \frac{\Sigma \mathrm{x}}{\mathrm{n}}$$

Population

$$\mu = \frac{\Sigma \mathrm{x}}{\mathrm{N}}$$

$$X = [1,2,3,4,5,6,7,8,9,10]$$

$$\sum X / n = 5.5$$

– The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions

Properties of the arithmetic mean

- Can be used for both discrete and continuous data. However, it is not appropriate for either nominal or ordinal data.
- For given set of data there is one and only one arithmetic mean.
- It is easily understood and easy to compute.
- Algebraic sum of the deviations of the given values from their arithmetic mean is always zero.
- It is greatly affected by the extreme values.
- In grouped data if any class interval is open, arithmetic mean can not be calculated.

Central value: The Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If *odd n*, middle value of sequence
  - if X = [9, 1,14,10,6,1,4,12,2,17]
  - then --------- is the median
- If *even n*, average of 2 middle values
  - if X = [1,12,4,17,9,10,11,2,14,6]
  - then ---------- is the median; i.e., how
- Median is not affected by extreme values

Properties of median

- Can be used for ordinal, discrete and continuous data. However, it is not appropriate for nominal data.
- There is only one median for a given set of data
- The median is easy to calculate
- Median is a positional average and hence it is not drastically affected by extreme values
- Can be calculated even in the case of open end intervals
- It is not a good representative of data if the number of items is small

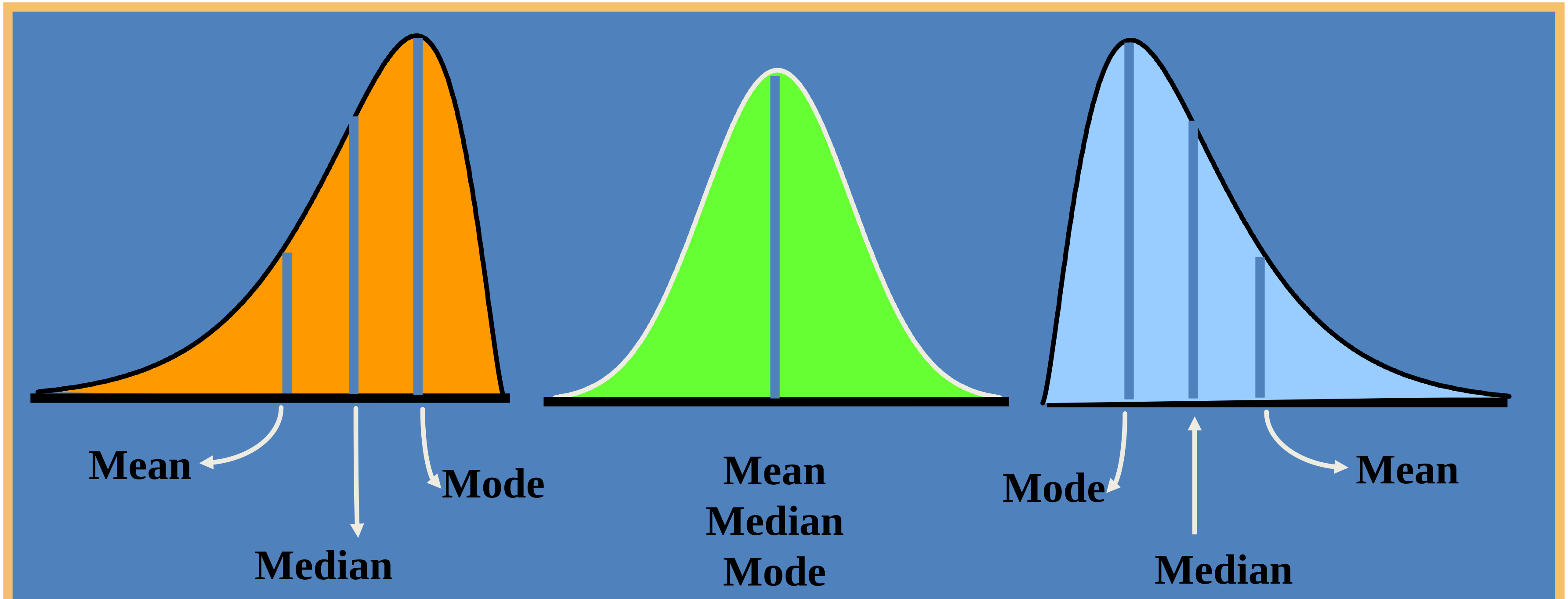
Central value: The Mode

- The mode is the most frequently occurring number in a distribution
  - if $X = [17,2,4,7, 12,7,8,10, 7,14, 1]$
  - then ------- is the mode
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
  - *bimodal* and *multimodal*
- Don’t have to be exactly equal frequency
  - *major mode, minor mode*
- Mode is not affected by extreme values

Properties of mode

- Can be used for nominal, ordinal, discrete and continuous data. However, it is more appropriate for nominal and ordinal data.
- It is not affected by extreme values
- It can be calculated for distributions with open end classes
- Often its value is not unique
- The main drawback of mode is that often it does not exist

Mean, Median, Mode



Negatively Skewed

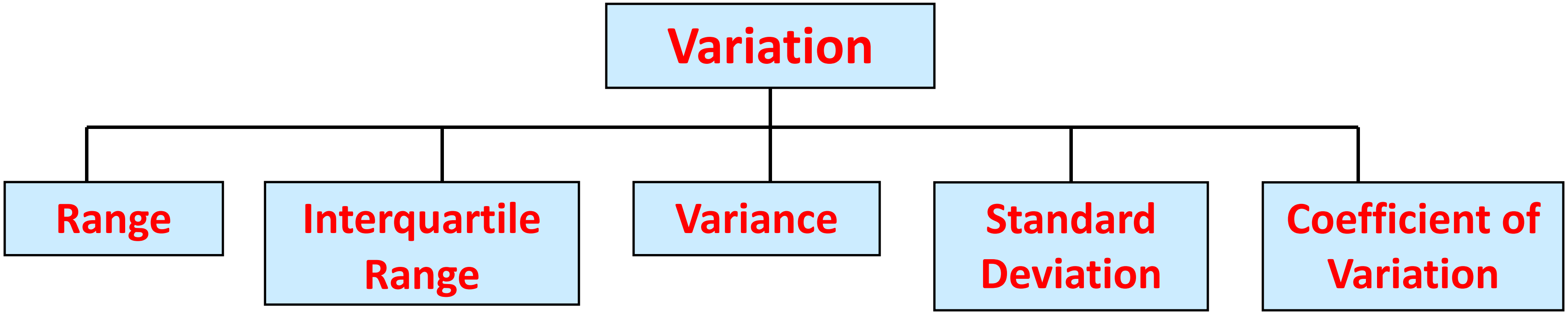
Symmetric (Not Skewed)

Positively Skewed

Measures of Dispersion

- Measures that quantify the variation or dispersion of a set of data from its central location
- Dispersion refers to the variety exhibited by the values of the data.
- The amount may be small when the values are close together.
- If all the values are the same, no dispersion

Measures of Variability - measures the amount of scatter in a dataset.



Variation

Range

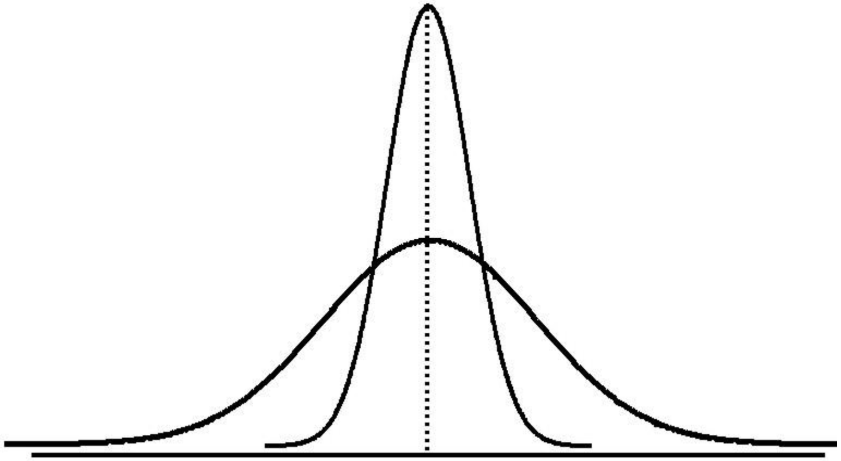
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

- Measures of variation give information on the **spread** or **variability** of the data values.



Same center, different variation

Population Variance

- Average of squared deviations of values from the mean
  - Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}$$

Where μ = population mean

N = population size

x_i = ith value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - – Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = ith value of the variable X

Properties of Variance:

- The main disadvantage of variance is that its unit is the square of the unite of the original measurement values
- The variance gives more weight to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.
- The drawbacks of variance are overcome by the standard deviation.

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Population standard deviation:

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}}}$$

Sample Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data
  - Sample standard deviation:

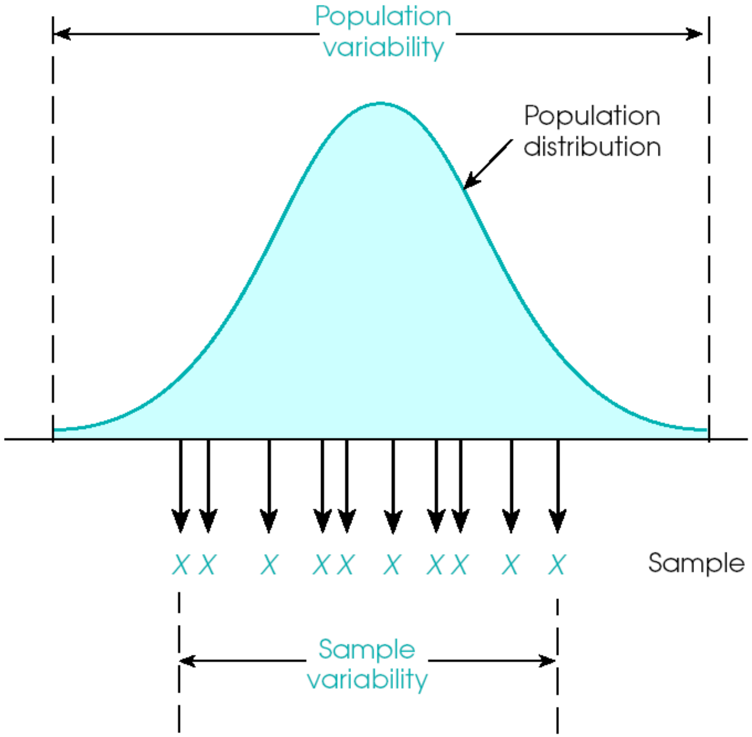
$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Properties of SD

- Has the advantage of being expressed in the same units of measurement as the mean
- The best measure of dispersion and is used widely because of the properties of the theoretical normal curve.
- However, if the units of measurements of variables of two data sets is not the same, then there variability can’t be compared by comparing the values of SD.

Dispersion

- Example
  - Data set 1: [0,25,50,75,100]
  - Data set 2: [48,49,50,51,52]
  - Both have a mean of ----------, but data set 1 clearly has greater *Variability* than data set 2.



Variance and standard deviation

Variance:

$$s^2 = \frac{\sum (X - \overline{X})^2}{n-1} =$$

Standard Deviation of sample:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}} =$$

Standard Deviation for whole population:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Dispersion: Standard Deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

- let $X = [3, 4, 5\ ,6, 7]$
- $\overline{\mathrm{X}} = ????$
- $(X - \overline{X}) = [\text{-----}, \text{-----}, \text{-----}, \text{-----}, \text{-----}]$

  ⇧ subtract $\overline{x}$ from each number in X
- $(X - \overline{X})^2 = [\text{-----}, \text{-----}, \text{------}, \text{------}, \text{-----}]$

  ⇧ squared deviations from the mean
- $\Sigma\ (X - \overline{X})^2 = ????$

  ⇧ sum of squared deviations from the mean (SS)
- $\Sigma\ (X - \underline{X})^2 /\mathrm{n}\text{-}1 = ???? = ????$

  ⇧ average squared deviation from the mean
- $\sqrt{\Sigma\ (X - \underline{X})^2 /\mathrm{n}\text{-}1} = \sqrt{??} = ???$

  ⇧ square root of averaged squared deviation

Advantages of Variance and Standard Deviation

- Each value in the data set is used in the calculation

- Values far from the mean are given extra weight

  (because deviations from the mean are squared)

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

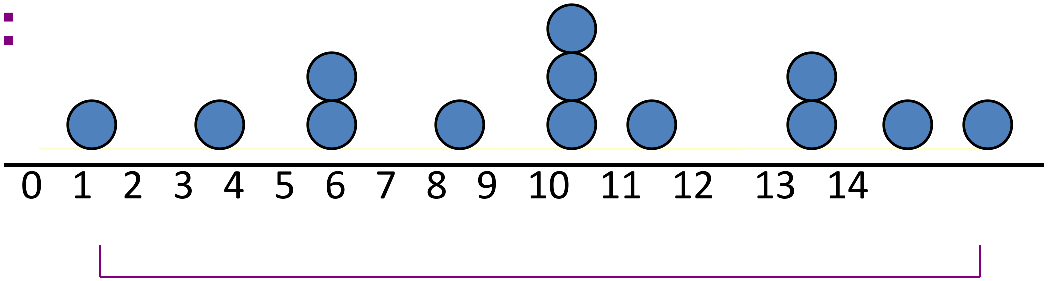
$$\boxed{CV = \left(\frac{s}{\overline{x}}\right) \cdot 100\%}$$

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

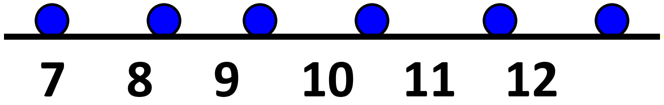
Example:



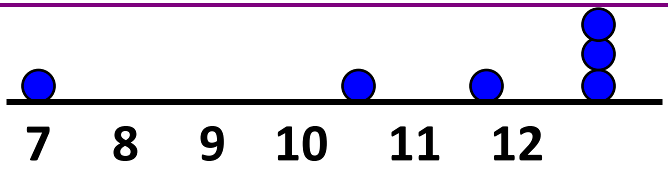
Range = ???

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = ???



Range = ?????

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = ???

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = ???

Properties of range

- It is the simplest crude measure and can be easily understood
- It takes into account only two values which causes it to be a poor measure of dispersion
- Very sensitive to extreme observations
- The larger the sample size, the larger the range

Unit 4

Introduction to Probability

3. Probability and Probability Distribution

Definition

- The concept of probability is frequently encountered in everyday communication.
- **For example**, a physician may say that a patient has a 50-50 chance of surviving a certain operation. Another physician may say that she is 95% certain that a patient has a particular disease.

Probability…cont'd

- Most people express probabilities in terms of **percentages**.
- But, it is more convenient to express probabilities as fractions. (Percentages result from multiplying the fractions by 100).
- Thus, we may measure the probability of the occurrence of some event by a number between 0 and 1.
- The more likely the event, the closer the number is to one. An event that can't occur has a probability of zero, and an event that is certain to occur has a probability of one.

- **Probability** as a general concept can be defined as
  ∩
  the chance of an event occurring.

Some definitions:

Equally likely outcomes:

- Are the outcomes that have the same chance of occurring.

Mutually exclusive:

- Two events are said to be mutually exclusive if they cannot occur simultaneously such that $A\cap B = \Phi$.

- **The universal Set** (S): The set all possible outcomes.
- **The empty set Φ** : Contain no elements.
- **The event,** E : is a set of outcomes in S which has a certain characteristic.
- **Classical Probability** : If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a triat, E, the probability of the occurrence of event E is equal to m/ N .
- ***For Example:*** in the rolling of the die , each of the six sides is equally likely to be observed . So, the probability that a 4 will be observed is equal to 1/6.

Relative Frequency Probability:

- **Def:** If some process is repeated a large number of times, n , and if some resulting event E occurs, *m* times , the relative frequency of occurrence of E , m/n will be approximately equal to probability of E . P(E) = m/n .

Subjective Probability:

- Probability measures the confidence that a particular individual has in the truth of a particular proposition.
- ***For Example :*** the probability that a cure for cancer will be discovered within the next 10 years.

Elementary Properties of Probability:

1. Given some process (or experiment) with n mutually exclusive events $E_1, E_2, E_3, \ldots, E_n$, then the probability of any event E_i is assigned a nonnegative number. That is, $P(E_i) \geq 0$

2. The sum of the probabilities of the mutually exclusive outcomes is equal to 1. $P(E_1) + P(E_2) + \ldots\ldots + P(E_n) = 1$

3. Consider any two mutually exclusive events, E_i and E_J. The probability of the occurrence of either E_i or E_J is equal to the sum of their individual probabilities.

$$P(E_i + E_J) = P(E_i) + P(E_J)$$

Rules of Probability

1. Addition Rule

- Given two events *A* and *B*, the probability that event *A*, or event *B*, or both occur is equal to the probability that event *A* occurs, plus the probability that event *B* occurs, minus the probability that the events occur simultaneously.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2. If A and B are mutually exclusive (disjoint), then

$P(A \cap B) = 0$ Then , addition rule is $P(A \cup B) = P(A) + P(B)$.

Rules of Probability

3. Complementary Rule

- The probability of an event *A* is equal to 1 minus the probability of its complement, which is written $\bar{A}$, and

$$P(\bar{A}) = 1 - P(A)$$

- This follows from the third property of probability since the **event, *A*, and its complement**, are **mutually exclusive**.
- **EXAMPLE:** Suppose that of 1200 admissions to a general hospital during a certain period of time, 750 are private admissions. If we designate these as set *A*, then $\bar{A}$ is equal to 1200 minus 750, or 450.

Rules of Probability

- We may compute
- P(A) = ???   and
- P(Ā) = ???   , and see that

➢ P (Ā) = 1 – P(A)

?????????

????????

Example: Frequency of Family History of Mood Disorder by Age Group

Family history of Mood Disorders	Early = 18 (E)	**Later >18** (L)	**Total**
Negative (A)	28	35	63
Bipolar Disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and Bipolar(D)	53	60	113
Total	141	177	318

Answer the following questions:

Suppose we select a person at random from the sample in the previous slide

1. The probability that this person will be 18-years old or younger?
2. The probability that this person has family history of mood orders Unipolar (C)?
3. The probability that this person has no family history of mood orders Unipolar ($\overline{C}$)?
4. The probability that this person is 18-years old or younger **or** has no family history of mood orders Negative (A)?
5. The probability that this person is more than18-years old and has family history of mood orders Unipolar and Bipolar (D)?

Conditional Probability:

- Sometimes, the set of “all possible outcomes” may constitute a subset of the total group.
- In other words, the size of the group of interest may be reduced **by conditions** not applicable to the total group.
- When probabilities are calculated with a subset of the total group as the denominator, the result is a ***conditional probability***.
- P(A\B) is the probability of A assuming that B has happened.

■ P(A\B)= $\frac{P(A \cap B)}{P(B)}$, P(B)≠ 0

■ P(B\A)= $\frac{P(A \cap B)}{P(A)}$, P(A)≠ 0

Example

From previous example, answer

- Suppose we pick a person at random and find he is 18 years or younger (E),what is the probability that this person will be one who has no family history of mood disorders (A)?

$P(A/E) = ???$ $= ???$

Multiplicative Rule:

- A probability may be computed from other probabilities. For example, a joint probability may be computed as the product of an appropriate marginal probability and an appropriate conditional probability. This relationship is known as the *multiplication rule* of probability.

- $P(A\cap B)= P(B)P(A\backslash B)$
- $P(A\cap B)= P(A)P(B\backslash A)$

Where,

$P(A)$: marginal probability of A.

$P(B)$: marginal probability of B.

$P(B\backslash A)$:The conditional probability.

EXAMPLE

- We wish to compute the joint probability of Early age at onset (E) and a negative family history of mood disorders (A) from knowledge of an appropriate marginal probability and an appropriate conditional probability.
- **Solution:** The probability we seek is
-
-
-
-

Independent Events:

- If A has no effect on B, we said that A and B are independent events.

- Thus, we see that if two events **are independent**, the probability of their joint occurrence is equal to the product of the probabilities of their individual occurrences.

  1. $P(A \cap B) = P(B)P(A)$

- Note that when two events with nonzero probabilities are independent, each of the following statements is true:

  2. $P(A \backslash B) = P(A)$

  3. $P(B \backslash A) = P(B)$

Example

- In a certain high school class consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses . If a student is picked at random from this class, the probability that the student wears eyeglasses , P(E), is 40/100 or 0.4 .

A) What is the probability that a student picked at random wears eyeglasses given that the student is a boy?

$$P(E/B) = \frac{P(E \cap B)}{P(B)}$$

???

B) What is the probability of the joint occurrence of the events of wearing eye glasses and being a boy?

Solution

$$P(E \cap B) = P(B)P(E/B)$$

but, since we have shown that events *E and B are independent we may r*eplace *P(E/B)* by P(E) to obtain

$$P(E \cap B) = P(B)P(E)$$

$$= ??????$$

Probability Distributions

- The relationship between the values of a random variable and the probabilities of their occurrence may be summarized by means of a device called a *probability distribution.*
- Knowledge of the probability distribution of a random variable provides a researcher with a powerful tool for summarizing and describing a set of data and for reaching conclusions about a population of data on the basis of a sample of data drawn from the population.

Probability Distributions of Discrete Variables

- The *probability distribution of a discrete random variable is a table,* graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.

Examples

TABLE 4.2.1 Number of Assistance Programs Utilized by Families with Children in Head Start Programs in Southern Ohio

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

TABLE 4.2.2 Probability Distribution of Programs Utilized by Families Among in Example 4.2.1

Number of Programs (x)	$P(X = x\}$
1	.2088
2	.1582
3	.1313
4	.1313
5	.1953
6	.1246
7	.0135
8	.0370
Total	1.0000

Mean and Variance of Discrete Probability Distributions

$$\mu = \sum xp(x)$$

$$\sigma^2 = \sum (x - \mu)^2 p(x) = \sum x^2 p(x) - \mu^2$$

Where *p(x)* is the relative frequency of a given random variable *X.*

The standard deviation is simply the positive square root of the variance.

The Binomial Distribution:

- The binomial distribution is one of the most widely encountered probability distributions in applied statistics. It is derived from a process known as a Bernoulli trial.
- Bernoulli trial is :

- When a random process or experiment called a trial can result in only one of two mutually exclusive outcomes, such as dead or alive, sick or well, the trial is called a Bernoulli trial.

The Bernoulli Process

- A sequence of Bernoulli trials forms a Bernoulli process under the following conditions

1. Each trial results in one of two possible, mutually exclusive, outcomes. One of the possible outcomes is denoted (arbitrarily) as a success, and the other is denoted a failure.

2. The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure, 1-p, is denoted by q.

3. The trials are independent, that is the outcome of any particular trial is not affected by the outcome of any other trial.

Example

- We are interested in being able to compute the probability of *x* successes in *n* Bernoulli trials.
- For example, if we examine all birth records from the Black Lion Hospital for the calendar year 2001, we find that 85.8 percent of the pregnancies had delivery in week 37 or later. We will refer to this as a full-term birth. With that percentage, we can interpret the probability of a recorded birth in week 37 or later as .858.
- If we randomly select five birth records from this population, what is the probability that exactly three of the records will be for full-term births?

Solution

- It will also be convenient to assign the number 1 to a success (record for a full-term birth) and the number 0 to a failure (record of a premature birth).
- Suppose the five birth records selected resulted in this sequence of full-term births: 10110
- The resulting probability is that of obtaining the specific sequence of outcomes in the order shown. We are not, however, interested in the order of occurrence of records for full-term and premature births but the probability of the occurrence of exactly three records of full-term births out of five randomly selected records.

- Instead of occurring in the sequence shown above (call it sequence number 1), three successes and two failures could occur in any one of the following additional sequences as well:

Number	Sequence
2	11100
3	10011
4	11010
5	11001
6	10101
7	01110
8	00111
9	01011
10	01101

- Each of these sequences has the same probability of occurring, and this probability is equal to $q^2 p^3$.
- What is the probability of getting sequence number 1 or sequence number 2 . . . or sequence number 10? From the addition rule we know that this probability is equal to the sum of the individual probabilities.
- In the present example we need to sum the 10 $q^2 p^3$ or, equivalently, multiply $q^2 p^3$ by 10.
- What is the probability, in a random sample of size 5, drawn from the specified population, of observing three successes (record of a full-term birth) and two failures (record of a premature birth)?

- Since in the population,
-
-
-
-

- As the size of the sample increases, listing the number of sequences becomes more and more difficult and tedious. What is needed is an easy method of counting the number of sequences.
- Such a method is provided by means of a counting formula that allows us to determine quickly how many subsets of objects can be formed.
- When the order of the objects in a subset is immaterial, the subset is called **a combination of objects**.
- When the order of objects in a subset does matter, we refer to the subset as **a permutation of objects**.

Permutation rule

- We consider rearrangement of the same items to be different (the permutation of **ABC** is **different** from **CBA** and is counted separately)

$$_nC_x = \frac{n!}{(n-x)!}$$

Combination rule

- We consider rearrangement of the same items to be the same (the combination of **ABC** is the **same** as **CBA**)

$$_nC_x = \frac{n!}{x!(n-x)!}$$

- The probability distribution of the binomial random variable **X**, the number of successes in **n** independent trials is:

$$f(x)=P(X=x)=\begin{pmatrix} n \\ x \end{pmatrix} p^{X} q^{n-X} \quad , \quad x=0,1,2,\ldots.,n$$

- Where $\begin{pmatrix} n \\ x \end{pmatrix}$ is the number of combinations of **n** distinct objects taken **x** of them at **a** time.

$$\begin{pmatrix} n \\ x \end{pmatrix} = \frac{n!}{x!(n-x)!}$$

$$x! = x(x-1)(x-2)\ldots.(1)$$

* Note: 0! =1

Example

- 14% of women admitted to Debre Berhan Referral Hospital drink *tella* during pregnancy. If 10 women are selected randomly, what is the probability that it will contain exactly 4 mothers who drink *tella* during pregnancy?

  ????????????????

  ?????

Properties of the binomial distribution

1. $f(x) \geq 0$
2. $\sum f(x) = 1$
3. The parameters of the binomial distribution are *n* and *p.* They are sufficient to specify a binomial distribution.

The Poisson Distribution

- If the random variable X is the number of occurrences of some random event in a certain period of time or space (or some volume of matter).
- The probability distribution of X is given by:

f (x) =P(X=x) = $\frac{e^{-\lambda}\lambda^{x}}{x\,!}$,x = 0,1,.....

- The symbol e is the constant equal to 2.7183. λ (Lambda) is called the parameter of the distribution and is the average number of occurrences of the random event in the interval (or volume)

Properties of the Poisson distribution

1. $f(x) \geq 0$, for every x
2. $\sum f(x) = 1$, so that the distribution satisfies the requirements for a probability distribution.
3. $\mu = E(X) = \lambda$
4. $\sigma^2 = \mathrm{var}(X) = \lambda$

Example

- In a study of a drug-induced anaphylaxis among patients taking rocuronium bromide as part of their anesthesia, Laake and Rottingen found that the occurrence of anaphylaxis followed a Poisson model withλ = 12 incidents per year in Norway .Find

1. The probability that in the next year, among patients receiving rocuronium, exactly three will experience anaphylaxis?

2. The probability that less than two patients receiving rocuronium, in the next year will experience anaphylaxis?
3. The probability that more than two patients receiving rocuronium, in the next year will experience anaphylaxis?
4. The variance of patients receiving rocuronium, in the next year who will experience anaphylaxis
5. The standard deviation of patients receiving rocuronium, in the next year who will experience anaphylaxis

Continuous Probability Distribution

Properties of continuous probability Distributions:

1. Area under the curve = 1.

2. P(X = a) = 0 , where a is a constant.

3. Area between two points a , b = P(a < x < b).

The normal distribution:

- It is one of the most important probability distributions in statistics. The distribution is frequently called the *Gaussian distribution* in recognition of his contributions.
- The normal density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty$$

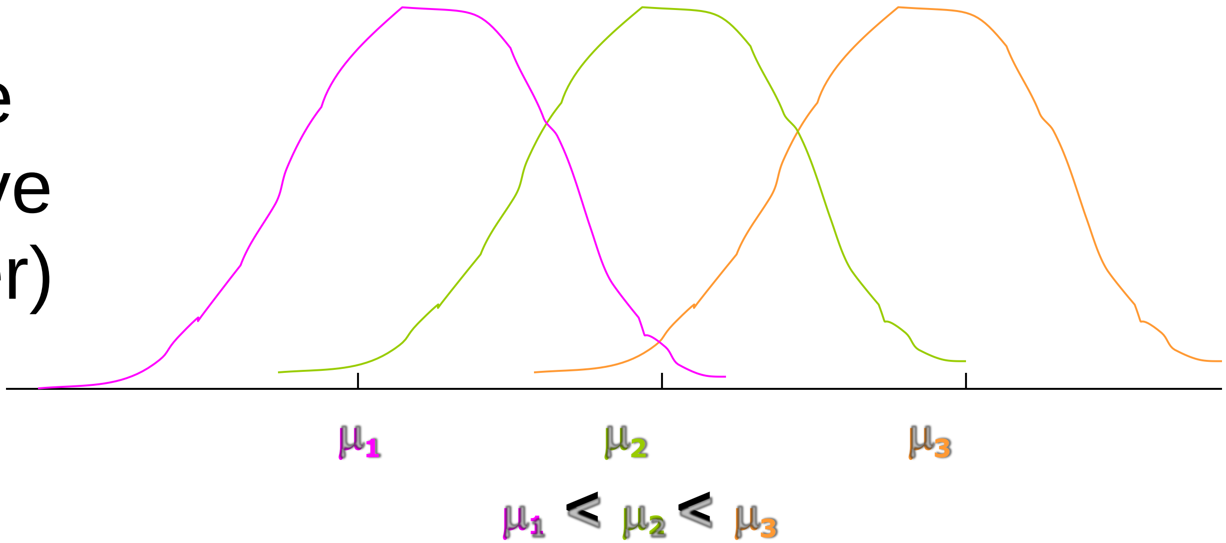
- π & e, are the familiar constants, 3.14 and 2.7183, respectively
- µ: population mean.
- σ : Population standard deviation.

Characteristics of the normal distribution:

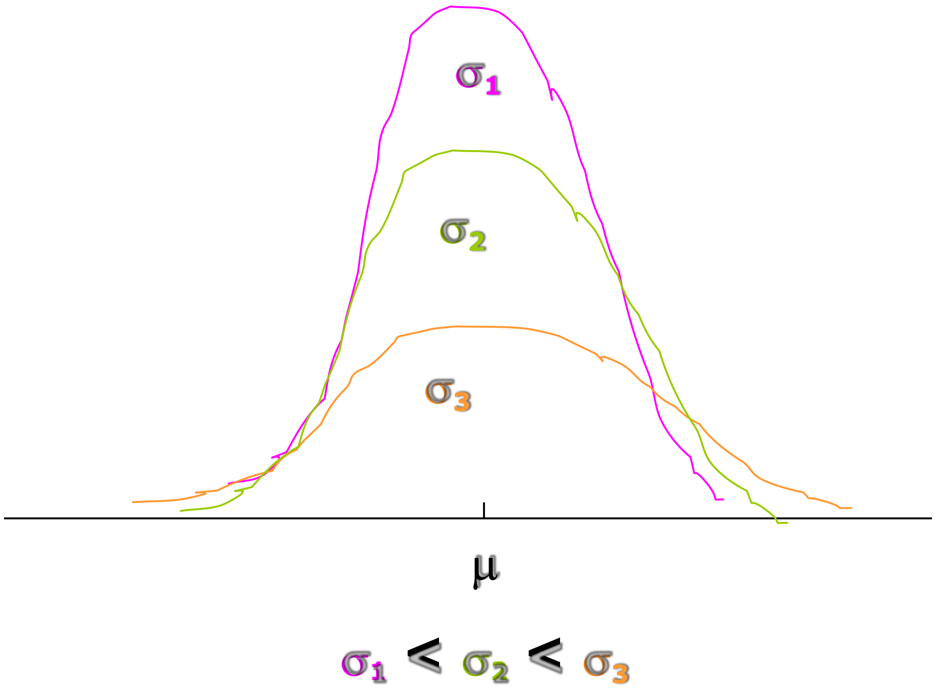
The following are some important characteristics of the normal distribution:

1. It is symmetrical about its mean, μ ,i.e., the curve on either side of μ is a mirror image of the other side
2. The mean, the median, and the mode are all equal.
3. The total area under the curve above the x-axis is one.
4. The normal distribution is completely determined by the parameters μ and σ.

- μ determines the location of the curve (location parameter)



But, σ determines the scale of the curve, i.e. the degree of flatness or peakedness of the curve (shape parameter)



5. If we erect perpendiculars a distance of 1σ from the mean in both directions, the area enclosed by these perpendiculars, the *x*-axis, and the curve will be approximately 68% of the total area. If we extend these lateral boundaries a distance of 2σ on either side of the mean, approximately 95% of the area will be enclosed, and extending them a distance of 3σ will cause approximately 99.7% of the total area to be enclosed.

1. $P(\mu - \sigma < x < \mu + \sigma) = 0.68$
2. $P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.95$
3. $P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.997$

The Standard normal distribution:

- Is a special case of normal distribution with mean equal 0 and a standard deviation of 1.
- The equation for the standard normal distribution is written as

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

Characteristics of the standard normal distribution

1. It is symmetrical about 0.
2. The total area under the curve above the x-axis is one.
3. We can use table (D) to find the probabilities and areas.

“How to use tables of Z”

Note that

The cumulative probabilities P(Z ≤ z) are given in tables for -3.49 < z < 3.49. Thus,

P (-3.49 < Z < 3.49) ≅ 1.

For standard normal distribution,

P (Z > 0) = P (Z < 0) = 0.5

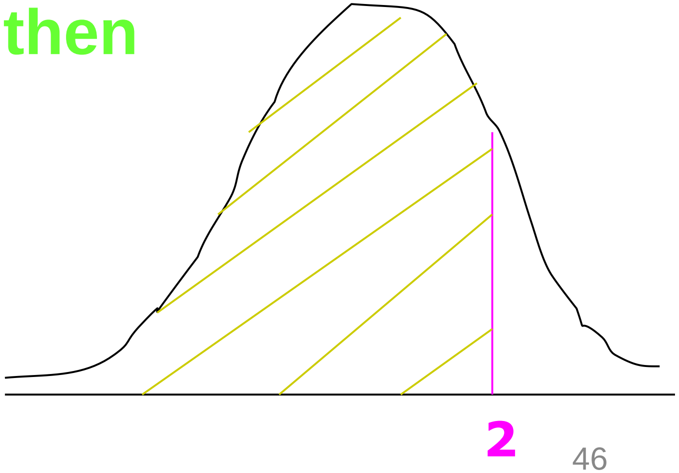
<u>Example 4.6.1:</u>

If Z is a standard normal distribution, then

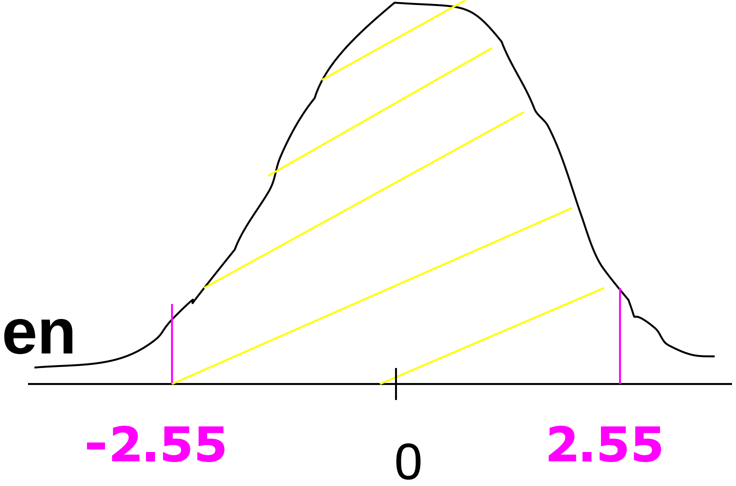
1) P(Z < 2) = 0.9772

is the area <u>to the left to</u> 2

and it equals 0.9772.



Example 4.6.2:



P(-2.55 < Z < 2.55) is the area between -2.55 and 2.55, Then it equals

P(-2.55 < Z < 2.55) =0.9946 – 0.0054

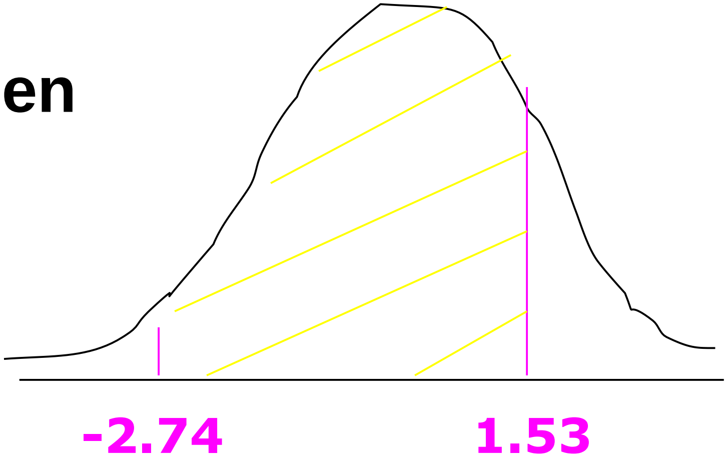
= 0.9892.

Example 4.6.2:

P(-2.74 < Z < 1.53) is the area between -2.74 and 1.53.

P(-2.74 < Z < 1.53) =0.9370 – 0.0031

= 0.9339.

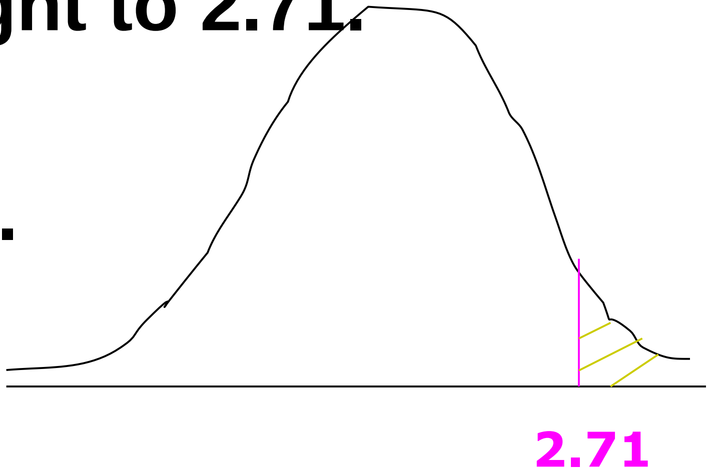


Example 4.6.3:

$P(Z > 2.71)$ is the area to the right to 2.71.

So,

$P(Z > 2.71) = 1 - 0.9966 = 0.0034$.

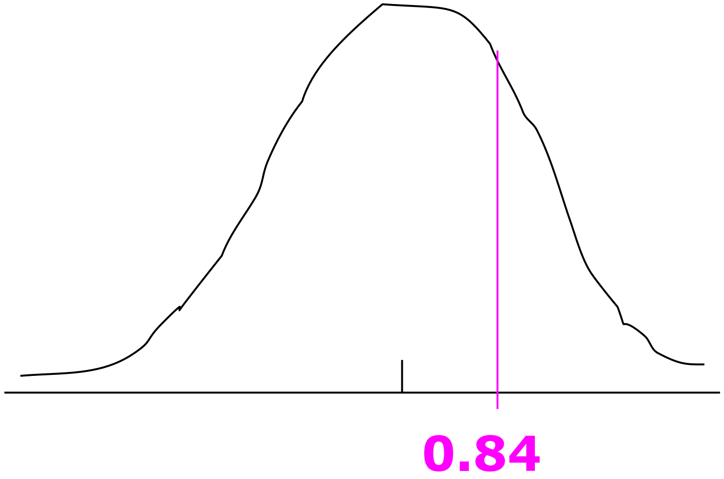


Example :

$P(.84 \leq z \leq 2.71)$

$P(z \leq 2.71) - P(Z \leq .84)$

$= .9966 - 7995 = .1971$



How to transform normal distribution (X) to standard normal distribution (Z)?

- This is done by the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Example:

- If X is normal with µ = 3, σ = 2. Find the value of standard normal Z, If X= 6?

Answer:

Normal Distribution Applications
Example

- The average daily emergency patients number in Leghare General Hospital is 28 and standard deviation is 2.4.

A) What is the probability that the number of patients is to be less than 26?

B) What is the probability that the number of patients is to be greater than 26?

???????????????

C) What is the probability that the number of patients is to be greater than 36?

C) What is the probability that the number of patients is to be 29 and 33?

And

The area between the two is

UNIT – 5

SAMPLING AND SAMPLING TECHNIQUES

4. Sampling concepts and estimation

4.1. Sampling Methods

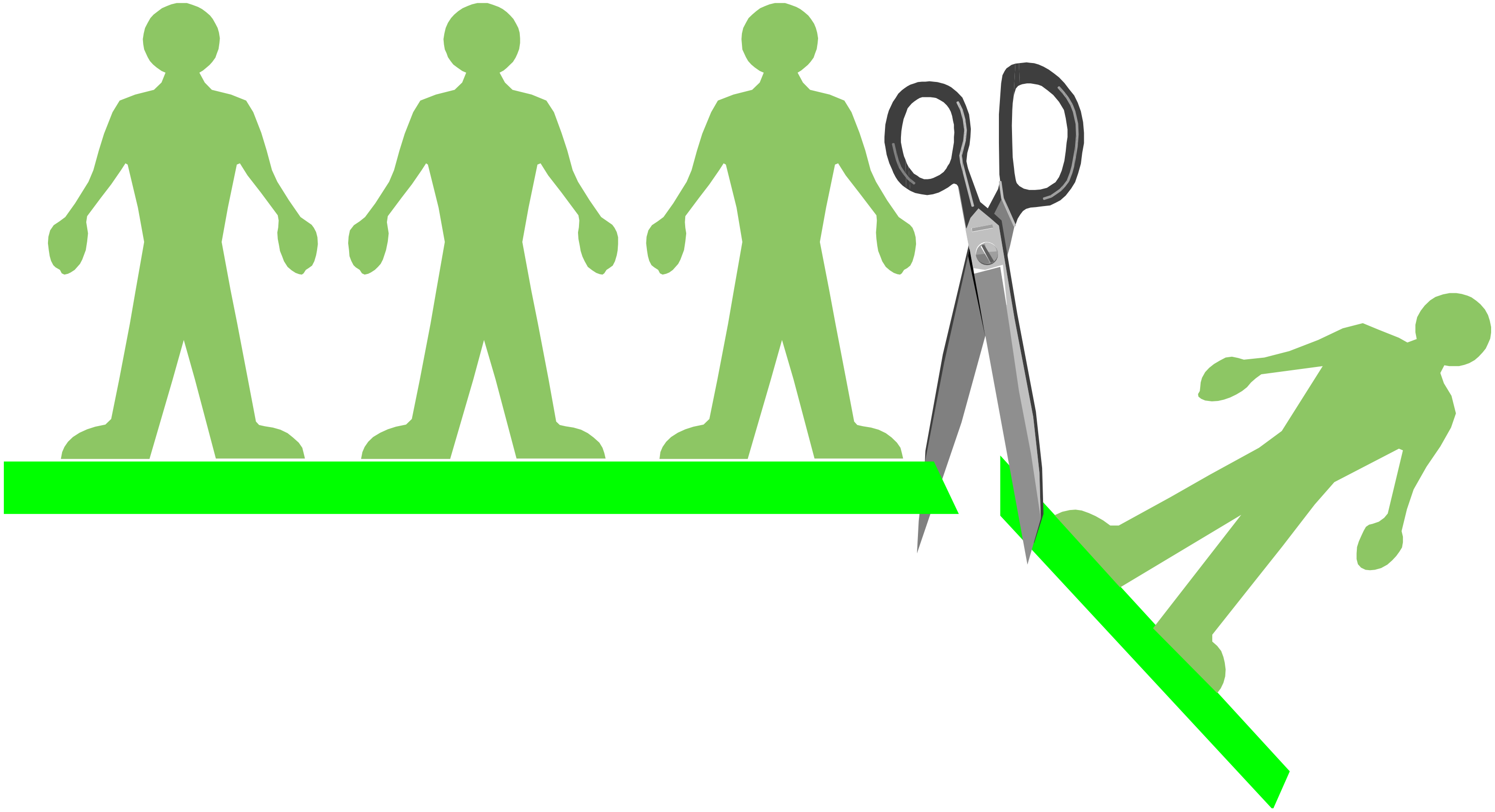
Sampling why?

- It is very difficult to study each and every unit of the population
- Time Consuming
- Requires huge amount of finance

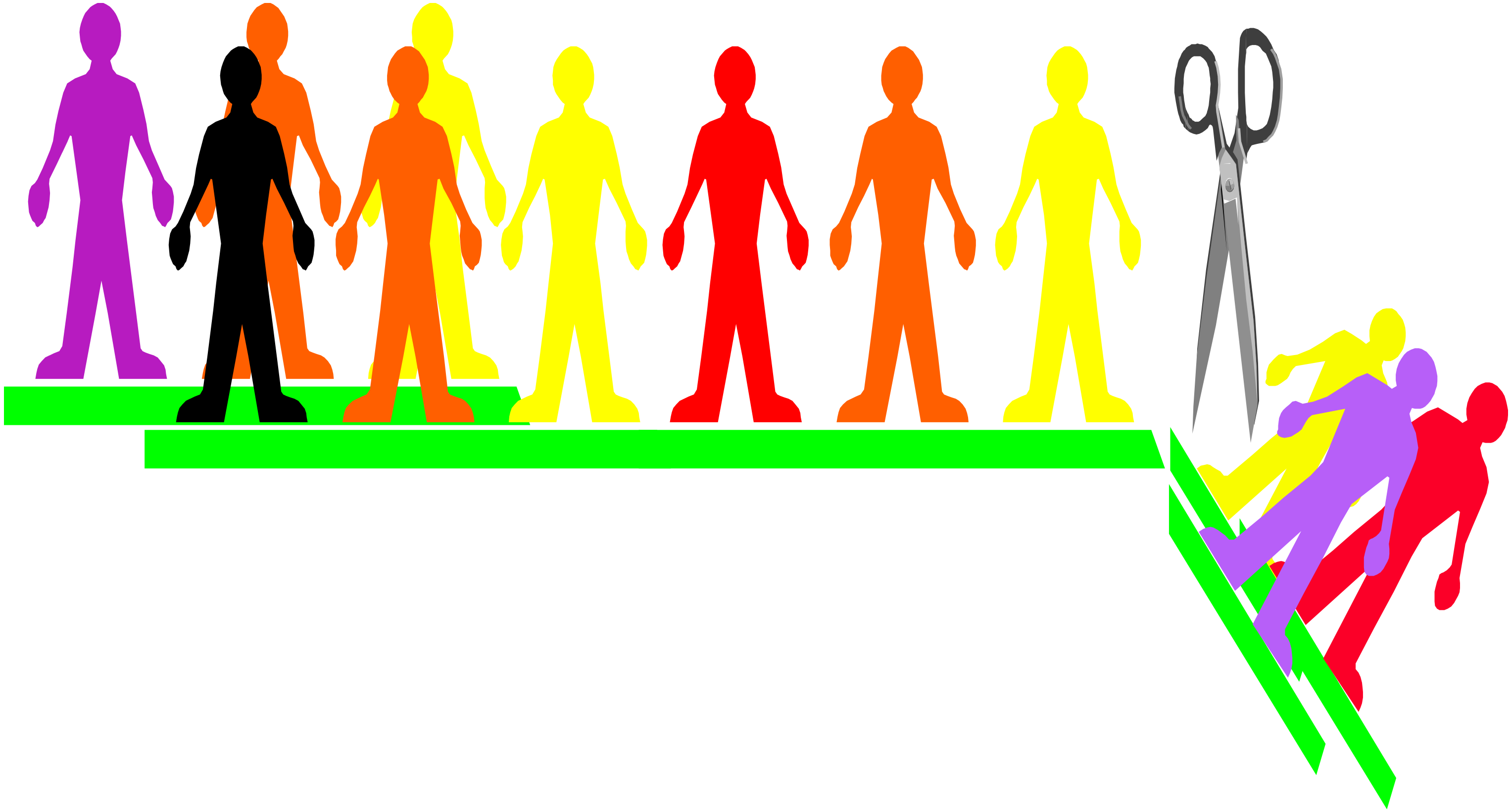
Sampling what?

- It is a unit selected from the population, i.e., subset of the population
- Representatives of the population
- Easy to draw the inferences

- It is very easy and convenient to draw the sample from homogenous population



- The population having significant variations (Heterogeneous), observation of multiple individual needed to find all possible characteristics that may exist



Some Characteristics of Good Samples

- Representative
- Accessible
- Low cost

Sampling Error

- Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample

- Increasing the sample size will reduce this type of error

Types of Sampling Error

- **Sample Errors**

- **Non Sample Errors**

Sample Errors

- Error caused by the act of taking a sample
- They cause sample results to be different from the results of census
- Differences between the sample and the population that exist only because of the observations that happened to be selected for the sample
- Statistical errors are sample error
- We have no control over

Non Sample Errors

- Refers to an error that occurs during data collection, causing the data to differ from the true values
- Not Controlled by Sample Size

❖Non Response Error

❖Response Error

Non Response Error

- Non-response errors occur when the survey fails to get a response to one, or possibly all, of the questions.

- Non-response errors result from a failure to collect complete information on all units in the selected sample.

Response Errors

- A response or data error is any systematic bias that occurs during data collection, analysis or interpretation

- Respondent error (e.g., lying, forgetting, etc.)
- Interviewer bias
- Recording errors
- Poorly designed questionnaires
- Measurement error

Respondent error

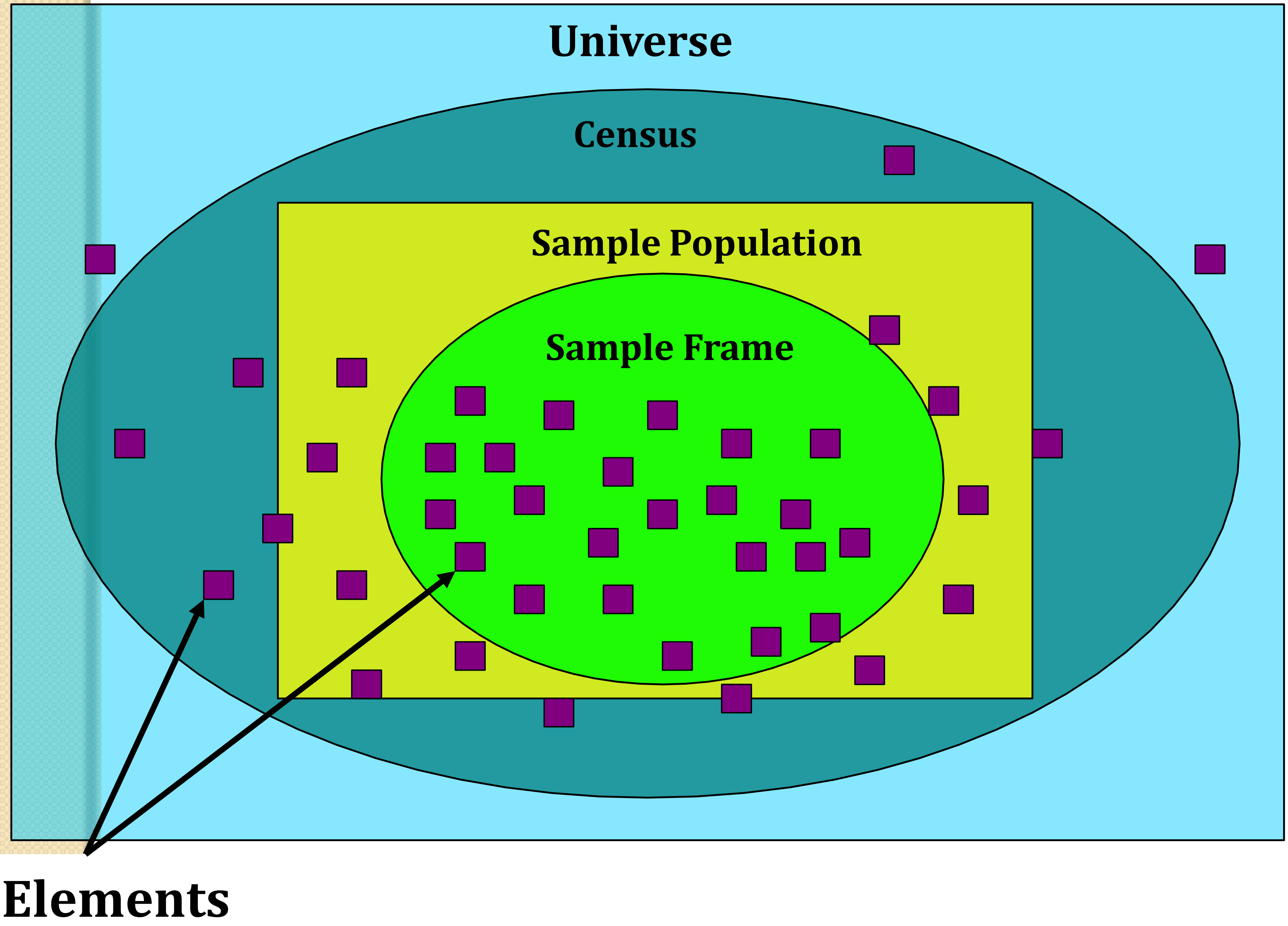
- Respondent gives an incorrect answer, e.g. due to competence implications, or due to sensitivity or social undesirability of question
- Respondent misunderstands the requirements
- Lack of motivation to give an accurate answer
- “Lazy” respondent gives an “average” answer
- Question requires memory/recall
- Proxy respondents are used, i.e. taking answers from someone other than the respondent

Interviewer bias

- Different interviewers administer a survey in different ways
- Differences occur in reactions of respondents to different interviewers, e.g. to interviewers of their own sex or own ethnic group
- Inadequate training of interviewers
- Inadequate attention to the selection of interviewers
- Too high workload for the interviewer

Measurement Error

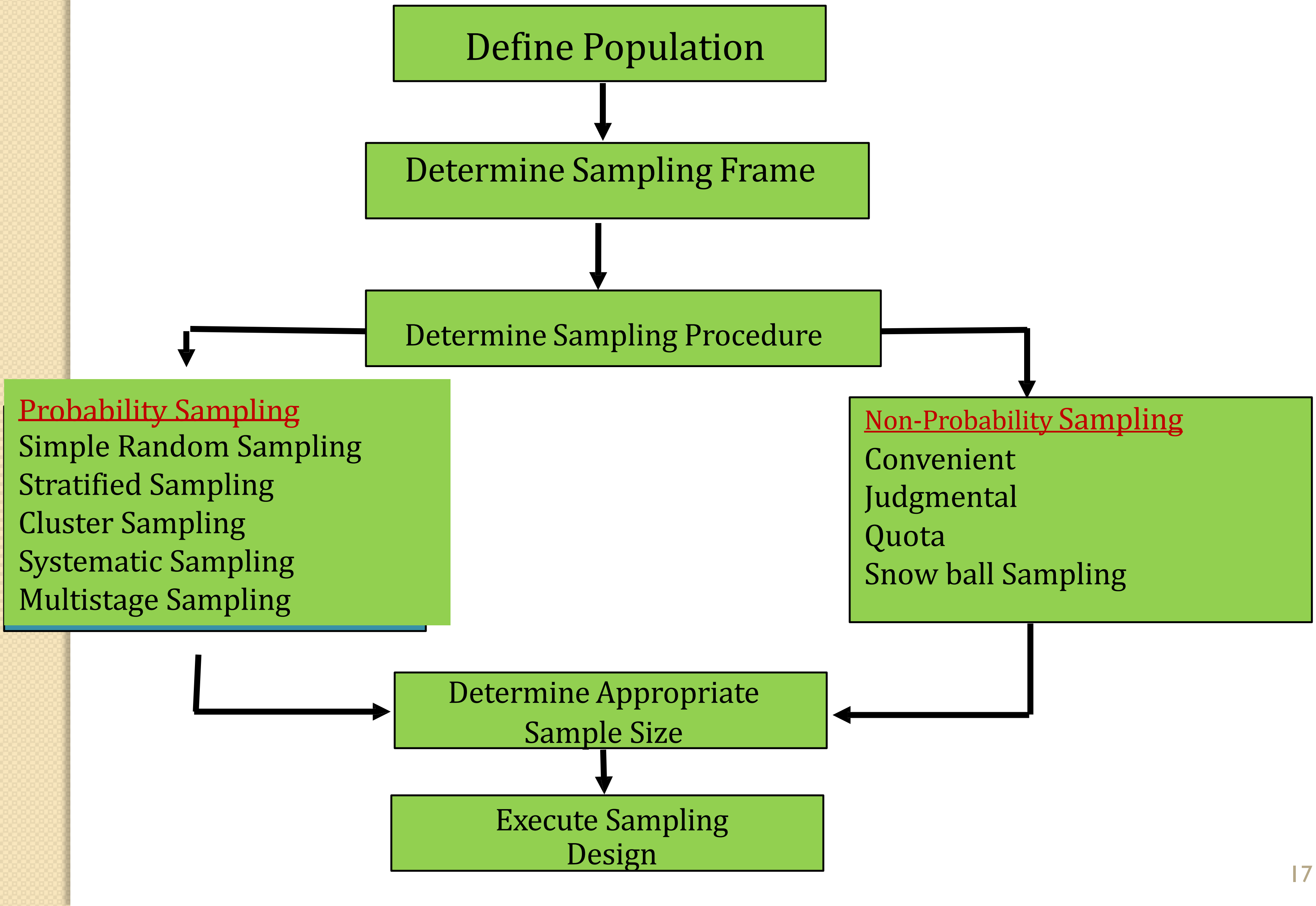
- The question is unclear, ambiguous or difficult to answer
- The list of possible answers suggested in the recording instrument is incomplete
- Requested information assumes a framework unfamiliar to the respondent
- The definitions used by the survey are different from those used by the respondent (e.g. how many part-time employees do you have? See next slide for an example)


Universe
Census
Sample Population
Sample Frame
Elements

Steps in Sampling Process

- Define the population
- Identify the sampling frame
- Select a sampling design or procedure
- Determine the sample size
- Draw the sample

Sampling Design Process



Define Population

Determine Sampling Frame

Determine Sampling Procedure

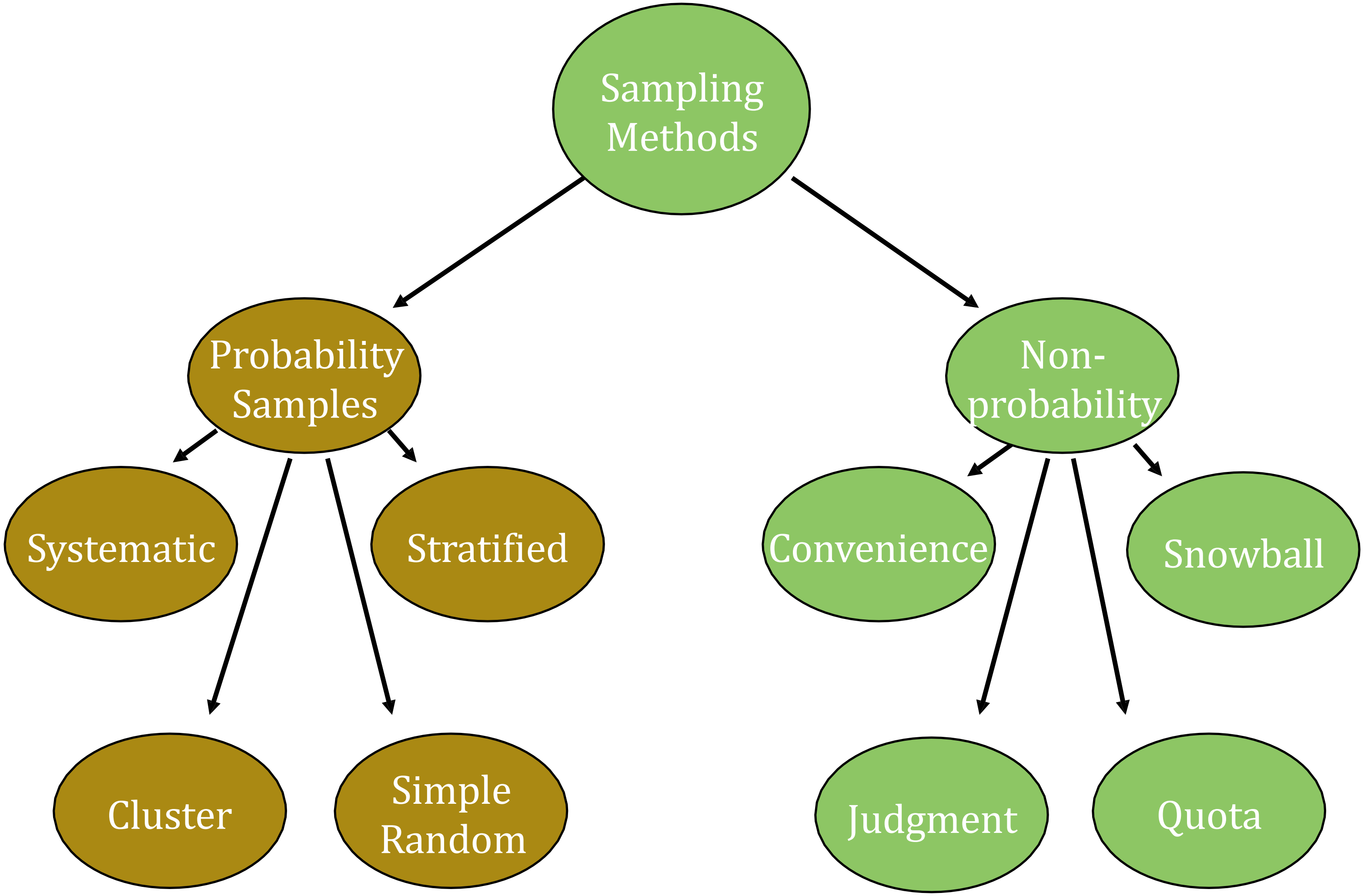
Probability Sampling
Simple Random Sampling
Stratified Sampling
Cluster Sampling
Systematic Sampling
Multistage Sampling

Non-Probability Sampling
Convenient
Judgmental
Quota
Snow ball Sampling

Determine Appropriate Sample Size

Execute Sampling Design

Sampling how? Classification of Sampling Methods



1. Probability Sampling

- Each and every unit of the population has a chance of being selected as a sampling unit
- It is mainly used in quantitative research
- Enables to produce results that are representative of the whole population
- Probability samples allow us to estimate the accuracy of the sample

Types of Probability Sampling

- **There are four main types of probability sample**

A. Simple Random Sampling

B. Stratified Sampling

C. Cluster Sampling

D. Systematic Sampling

A) Simple Random Sampling

- The purest form of probability sampling
- Assures each element in the population has an equal chance of being included in the sample
- Random samples are selected by using chance methods or random numbers

Types of Simple Random Sample

- With replacement

- Without replacement

With replacement

- The unit once selected has the chance to be selected again

Without replacement

- The unit once selected can not be selected again

Methods of SRS

- Lottery Method
- Random number Table method

Advantages of SRS

- There is less chance for personal bias
- Sampling error can be measured
- This method is economical as it saves time, money and labour
- Easy to analyze data

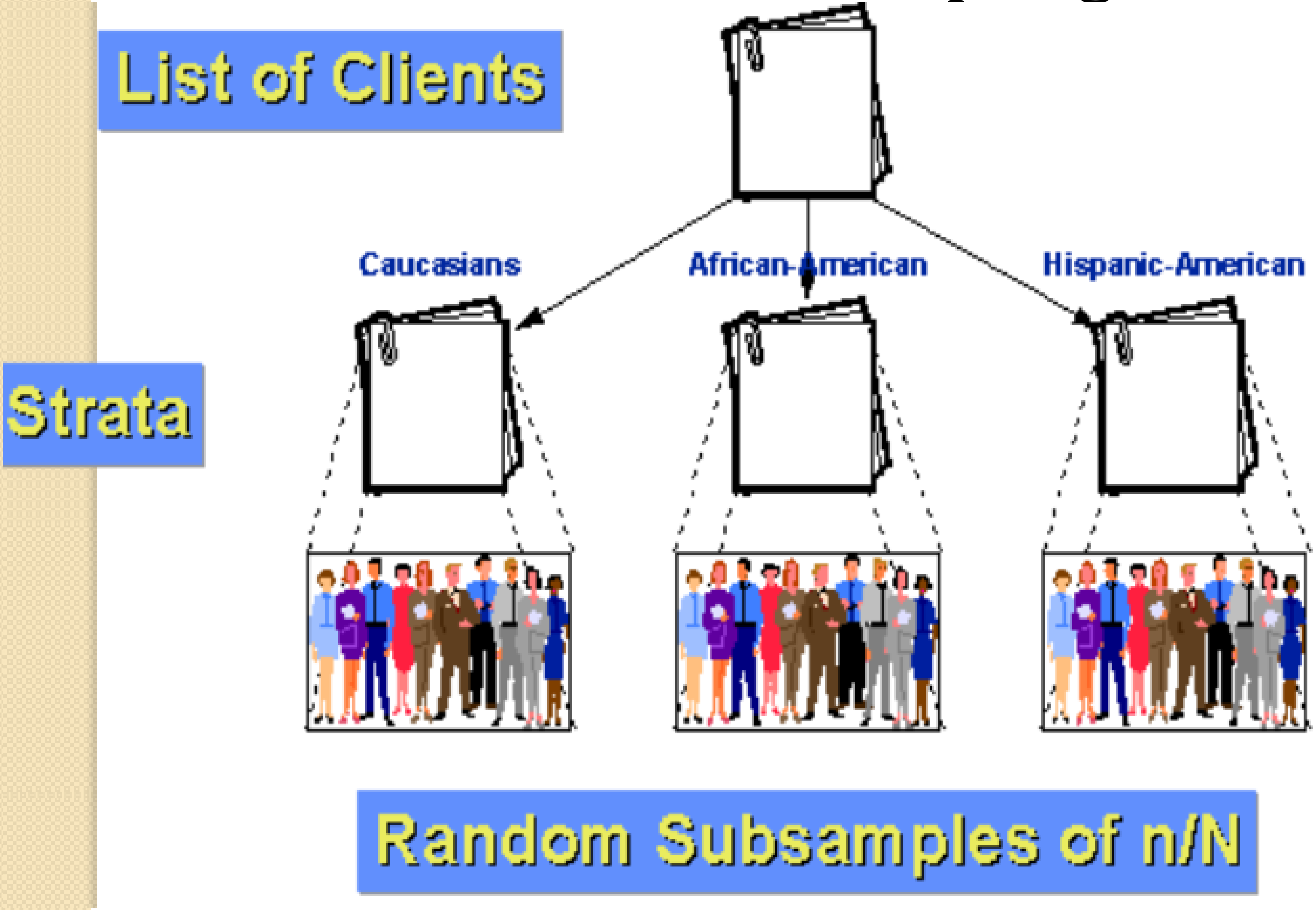
Disadvantage

- It cannot be applied if the population is heterogeneous
- Requires sampling frame
- Does not use researchers’ expertise
- If the size of the sample is small, then it will not be a representative of the population.

B) Stratified Random Sampling

- When the population is heterogeneous with respect to the characteristic in which we are interested, we adopt stratified sampling
- The population is divided into two or more groups called strata, according to some criterion, such as geographic location, grade level, age, or income, and subsamples are randomly selected from each strata.
- Elements within each strata are homogeneous, but are heterogeneous across strata

Stratified Random Sampling



List of Clients

Strata

Caucasians

African-American

Hispanic-American

Random Subsamples of n/N

Types of Stratified Random Sampling

- Proportionate Stratified Random Sampling

  Equal proportion of sample unit are selected from each strata

- Disproportionate Stratified Random Sampling

- If the number of units to be selected from a stratum is proportional to the size of the stratum

Advantages

- Assures representation of all groups in sample population needed
- Characteristics of each stratum can be estimated and comparisons made
- It ensures greater accuracy

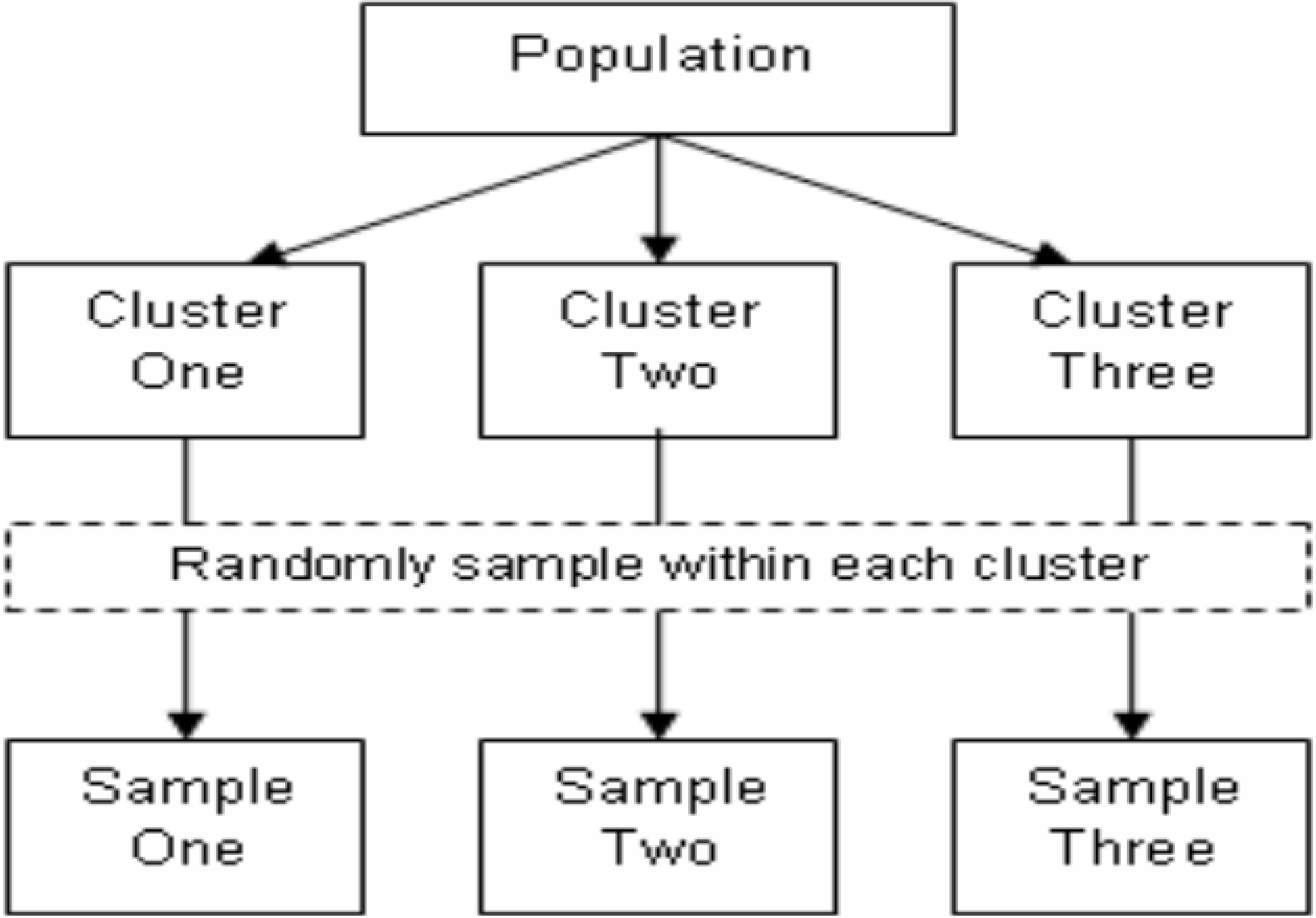
Disadvantages

- To divide the population into homogeneous strata, it requires more money, time and statistical experience which is a difficult one.
- If proper stratification is not done, the sample will have an effect of bias.

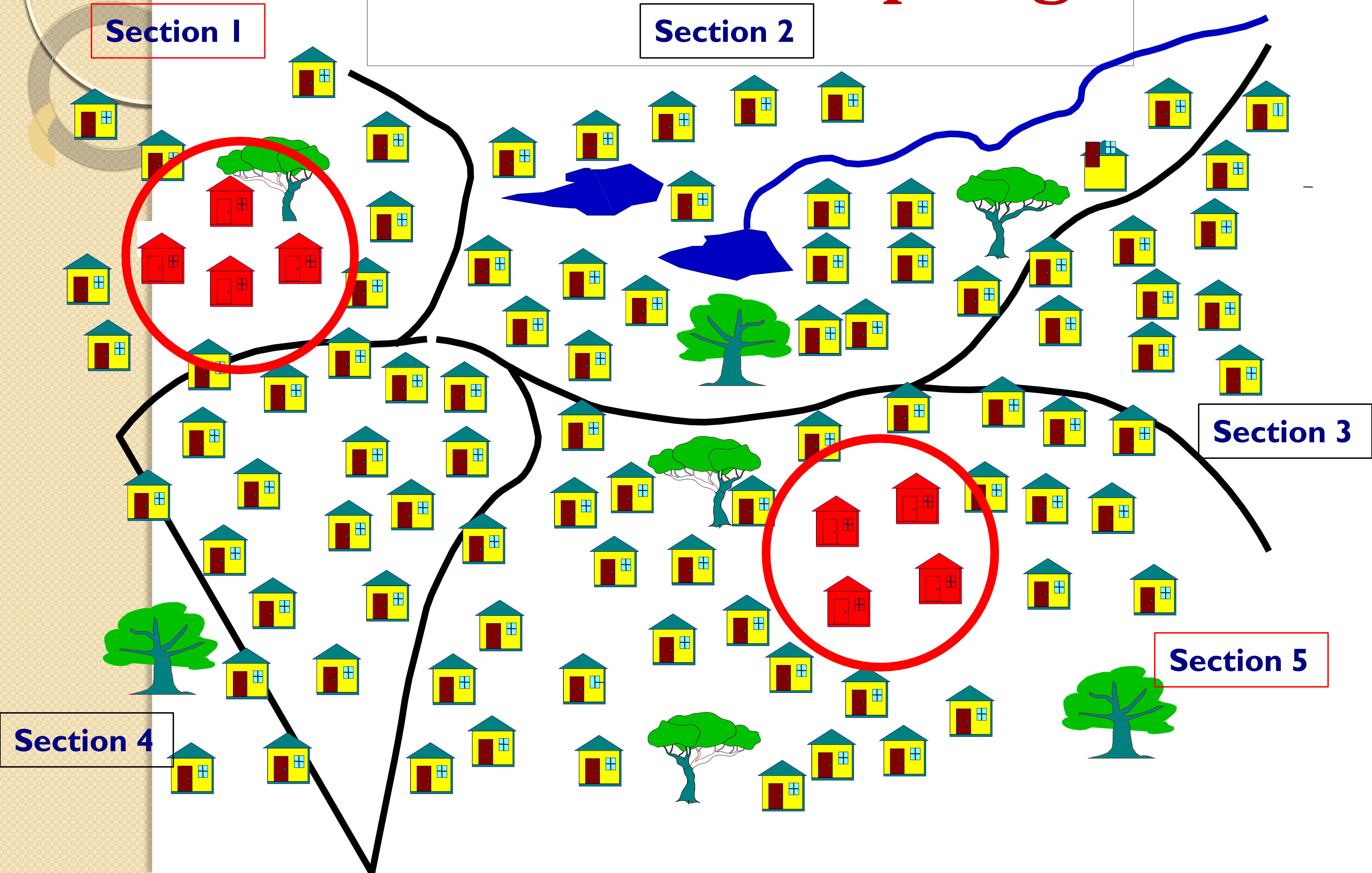
C) Cluster Sampling

- The population is divided into subgroups (clusters) like families.

  A simple random sample is taken of the subgroups and then all members of the cluster selected are surveyed.


Population
Cluster One
Cluster Two
Cluster Three
Randomly sample within each cluster
Sample One
Sample Two
Sample Three

Cluster sampling



Advantages

- Low cost/high frequency of use
- Requires list of all clusters, but only of individuals within chosen clusters
- Can estimate characteristics of both cluster and population
- Researchers lack a good sampling frame for a dispersed population

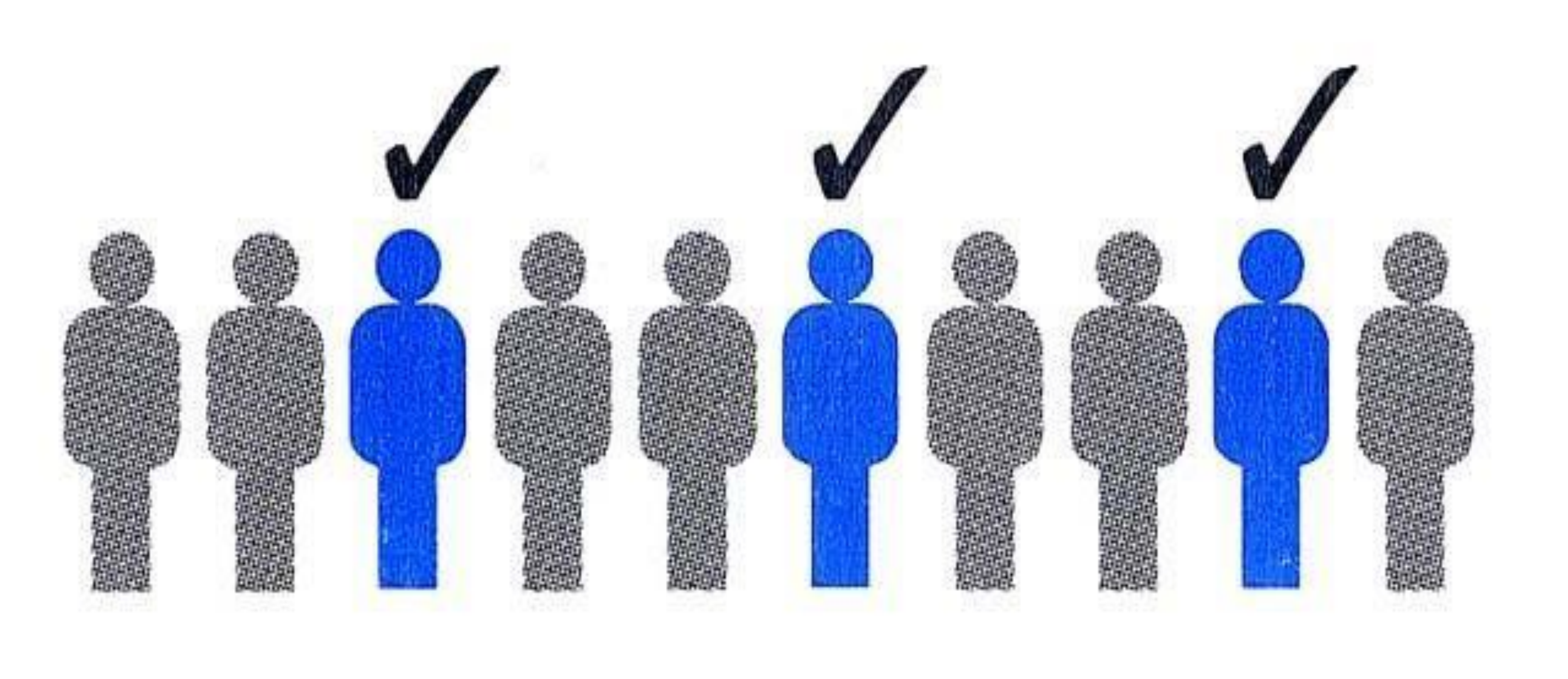
Disadvantages

- Usually less expensive than SRS but not as accurate
- Each stage in cluster sampling introduces sampling error—the more stages there are, the more error there tends to be

D) Systematic Random Sampling

- Order all units in the sampling frame based on some variable and then every nth number on the list is selected
- Gaps between elements are equal and Constant → There is periodicity.
- N= Sampling Interval

Systematic Random Sampling



Advantages

- Moderate cost; moderate usage
- External validity high; internal validity high; statistical estimation of error
- Simple to draw sample; easy to verify

Disadvantage

- Periodic ordering
- Requires sampling frame

2. Non Probability Sampling

- The probability of each case being selected from the total population is not known
- Units of the sample are chosen on the basis of personal judgment or convenience
- This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias, and you can’t use it to make valid statistical inferences about the whole population.
- Non-probability sampling techniques are often appropriate for exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population

Non Probability Sampling

- Involves non random methods in selection of sample
- All have not equal chance of being selected
- Selection depends upon situation
- Considerably less expensive
- Convenient
- Sample chosen in many ways

Types of Non probability Sampling

- Purposive Sampling
- Quota sampling (larger populations)
- Snowball sampling
- Self-selection sampling
- Convenience sampling

A) Purposive Sampling

- Also called judgment Sampling
- It involves the researcher using his judgment to select a sample that is most useful to the purposes of the research.
- When taking sample reject people who do not fit for a particular profile
- It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences

Advantages

Samples are chosen well based on the some criteria

Meet the specific objective

Demerit

- Bias selection of sample may occur

- Time consuming process

B) Quota Sampling

- The population is divided into cells on the basis of relevant control characteristics.
- A quota of sample units is established for each cell
- A convenience sample is drawn for each cell until the quota is met
- It is entirely non random and it is normally used for interview surveys

Advantages

- Used when research budget limited
- Very extensively used/understood
- No need for list of population elements
- Introduces some elements of stratification

Demerit

- Variability and bias cannot be measured or controlled
- Time Consuming
- Projecting data beyond sample not justified

C) Snowball Sampling

- The research starts with a key person and introduce the next one to become a chain
- Make contact with one or two cases in the population
- Ask these cases to identify further cases.
- Stop when either no new cases are given or the sample is as large as manageable

Advantages

- low cost
- Useful in specific circumstances
- Useful for locating rare populations

Demerit

- Bias because sampling units not independent
- Projecting data beyond sample not justified

D) Self selection Sampling

- It occurs when you allow each case usually individuals, to identify their desire to take part in the research you therefore
- Publicize your need for cases, either by advertising through appropriate media or by asking them to take part
- Collect data from those who respond

Advantages

- More accurate
- Useful in specific circumstances to serve the purpose

Demerit

- More costly due to Advertizing
- Mass are left

E) Convenience Sampling

- Called as Accidental / Incidental Sampling
- Selecting individuals who happen to be most accessible to the researcher
- Easy and inexpensive way to gather initial data
- There is no way to tell if the sample is representative of the population, so it can't produce generalizable results
- It is done at the "convenience" of the researcher

Merits

- Very low cost
- Extensively used/understood
- No need for list of population elements

Demerit

- Variability and bias cannot be measured or controlled
- Projecting data beyond sample not justified
- Restriction of Generalization

End of part one of the course

Thanks all for your tolerance!