

NUR 351: BIOSTATISTICS

DR. KENNETH ADU-GYAMFI

STATISTICS



➤ Statistics can be defined in two senses;

❑ Statistics (plural); is a systematic collection of numerical facts in any field

❑ Statistics (singular); It is the study of methods and procedures used in collecting, classifying, organising, analysing, and interpreting a body of numbers for information and decision making



Think
through
this

➤ From the two definitions given, what then is biostatistics?

BIOSTATISTICS



➤ From the two definitions, we can describe biostatistics as;

- 1. the body of numbers or data in the field of biological or medical sciences
 - e.g., medical sciences statistics; number of patients in a hospital, number of allied health science students in a school or a college, number of nurses or midwives in a region, categories of healthcare workers in Cape Coast Teaching Hospital
- 2. the study of the methods and procedures used in collecting, organising, analysing and interpreting a body of numbers related to biological or medication sciences for information and decision making

Think through this



➤ How significant is biostatistics to you as a healthcare worker?

Importance of biostatistics



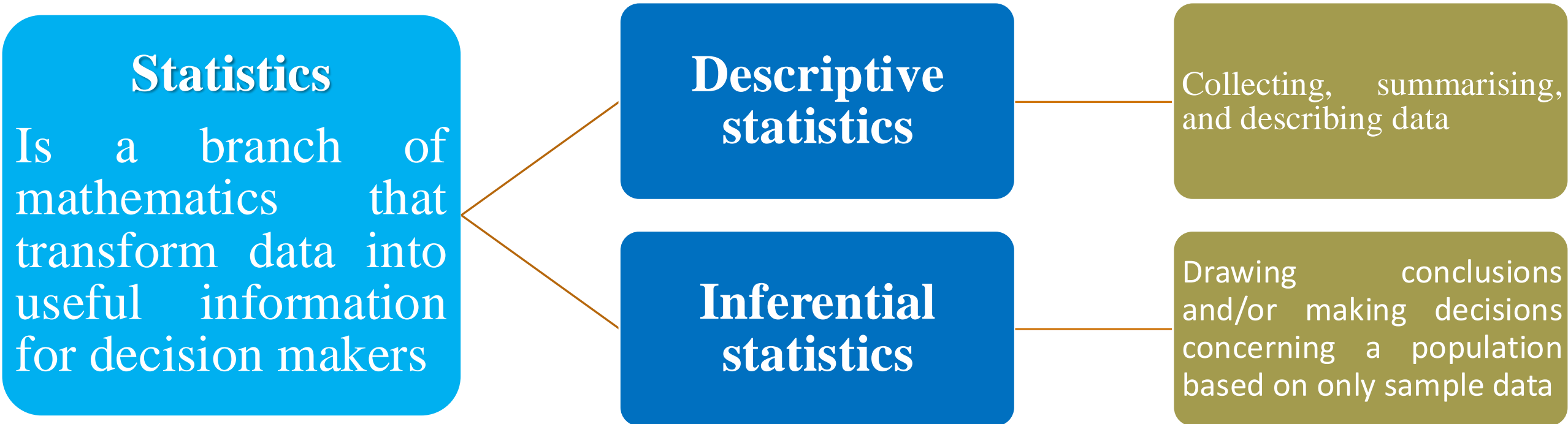
1. It helps biological or medical sciences teachers to use the appropriate statistics in describing the performance of their classes to others
2. It puts the biological or medical sciences teachers in a position to better understand the information they receive from test scores on students
3. It is useful for research purposes



4. It helps the teacher in the critical reading and understanding of professional journals in biological or medical sciences

5. It helps the biological or medical sciences teacher to understand information from standardised achievement test manuals

Types of statistics



Descriptive statistics



- ❖ Provides summary data about a group
- A single number is often used to represent a group
 - e.g., using mean (average), median, and standard deviation
- Descriptive statistics can be used to describe the actual sample you study
 - but not to extend conclusions to a broader population

Inferential Statistics



- ❖ Uses data from a small group (i.e., the sample) to make statements or generalisations about a much larger group (being the population)
- To use inferential statistics, you have to be sure the sample you are studying is representative of the group (the population) you want to generalise over

Basic Vocabulary in Statistics



- **Variable:** is any characteristic of an individual or object that can take on different values
- A value is an assigned number or label representing the attribute of a given individual or object
- e.g., *marital status* can be broken down into categories with values as:
 - a. never married – 1
 - b. married – 2
 - c. divorced - 3
 - d. widowed – 4



➤ **Data:** are the different values associated with a variable

■ Variable values are meaningless unless their variables have *operational definitions*

-universally accepted meanings that are clear to all associated with an analysis

➤ A **population:** consists of all the items or individuals about which you want to draw a conclusion

➤ A **sample:** is the portion of a population selected for analysis

➤ A **parameter:** is a numerical measure that describes a characteristic of a population

➤ A **statistic:** is a numerical measure that describes a characteristic of a sample

Sources of data



□ **Primary sources:** using the data for analysis

- e.g., *survey of general nurses*
- *an experiment on effectiveness of Covid-19 vaccine*
- *and observed data on how novice and experienced midwives treat pregnant women*
- i.e., in these cases, the one analysing the data collected them on his/her own



❑ **Secondary Sources:** the data is collected someone other than the primary user

- The common sources of secondary data are
 - analysing or examining data from census (of healthcare workers)
 - records of patient to medical doctor ratio in teaching hospitals in Ghana
 - information collected by the Ministry of Health
 - organisational records
 - data collected for other research purposes
 - data from print journals or data published on the internet



Secondary data

- The person performing data analysis is not the data collector
- Secondary data analysis can save time that would otherwise be spent collecting data
 - particularly, in the case of quantitative data
 - can provide large and high quality databases

Types of Variables



- **Categorical** (qualitative) variables: they have values that can only be placed in categories
- i.e., categorical variables represent types of data that may be divided into groups
- e.g., race of patients, sex of midwives, age group of nutritionist, educational levels of healthcare workers, marital status of patients
- NB:** frequencies are ideal descriptive statistics for categorical variables
- This will tell you how many people are in each gave each response
- e.g., 57% of the patient visiting diabetes centre in the Cape Coast Teaching Hospital are female (and 43% males)

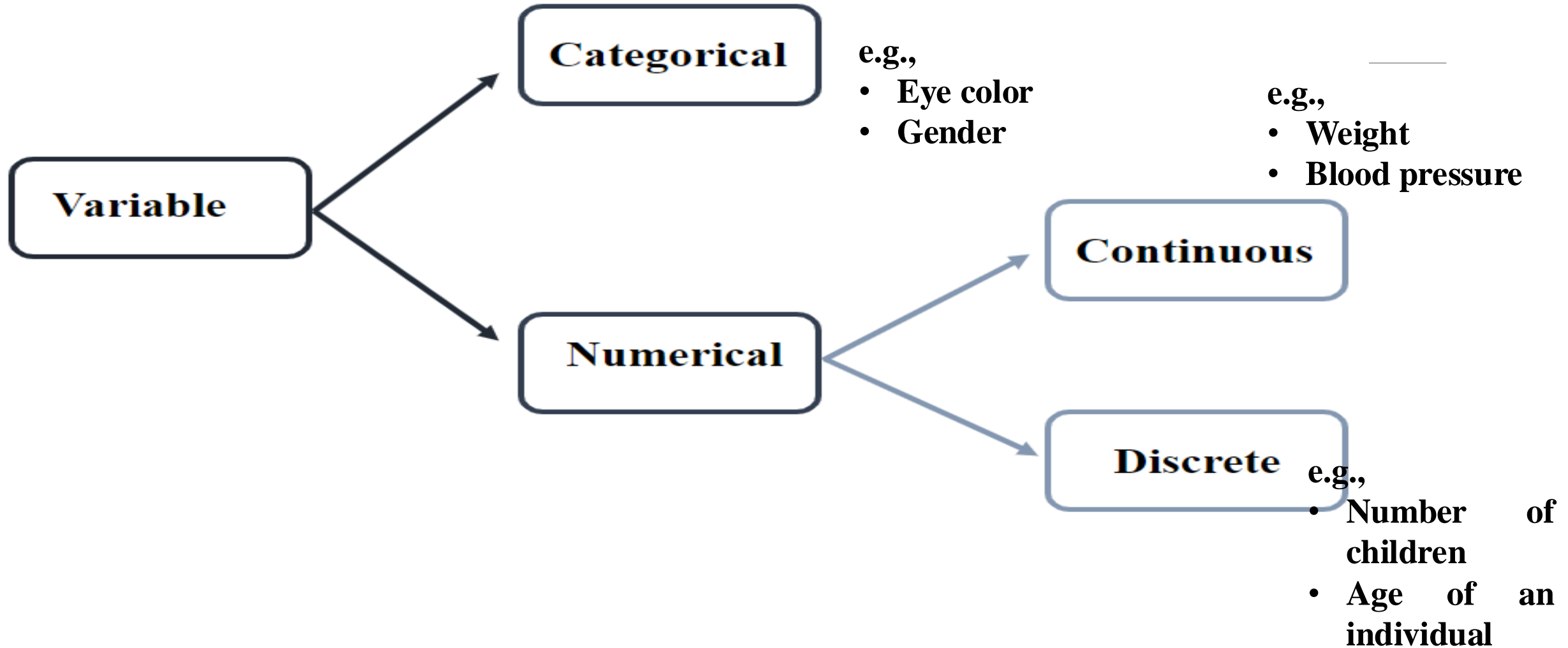
Types of Variables



- **Numerical** (quantitative) variables: they have values that represent quantities
- i.e., numerical variables are represented in numerical terms rather than in natural language descriptions
- Numerical variables provide researchers with quantitative data
- e.g., weight of patients, height of children, blood pressure of pregnant women, heart rate of new born babies
- NB:** on numerical variables measurements have numerical meaning



Types of Variables



Levels of Measurement



- ❑ **Nominal scale:** classifies persons or objects into two or more categories in which no ranking is implied
- ✓ Whatever the classification, a person can be in **one and only one** category, and members of a given category have a common set of characteristics.
- ✓ For identification purposes, categories are numbered
 - e.g., gender: Male - 1, Female – 2
- ✓ Hence, all male laboratory technologists have a common characteristic and all female laboratory technologists have a common characteristic being different from males

Levels of Measurement



□ **Ordinal scale:** *not* only classifies subjects but also ranks them in terms of the degree to which they possess a characteristic/attribute of interest

✓ i.e., an ordinal scale puts subjects in order from highest to lowest, or from most to least

-e.g., with respect to the height of patients at the OPD University of Cape Coast Hospital, 5 patients can be ranked from 1 to 5, the subject with rank 1 being the shortest

-Students' grades in Biostatistics for semester 2 of 2022/2023 academic year can be ranked as; A, B⁺, B, C⁺, C, D⁺, D, E

Levels of Measurement



❑ **Interval scale:** has all the characteristics of both nominal and ordinal scales and in addition has equal intervals

✓ The zero point is arbitrary and does not mean the absence of the characteristics/trait

-e.g., Celsius temperature of patients, academic achievement (scores) of students in Biostatistics

❑ **Ratio scale:** has all the advantages of the types of scales

✓ In addition. Ratio scale has a meaningful true zero point

-e.g., height of general nursing students, weight of new born babies less than 2 years, and time of administration of morning-after pill

Presentation 1



➤ Outline **five** examples each of variables in biological or medical sciences students under;

- a) Categorical
- b) Discrete
- c) Continuous

Measures of Central Tendency



❑ These measures are also called Averages

✓ They provide single values which are used to summarise a set of observations/data

✓ The three main measures of central tendency:

- **Mean**

- **Median**

- **Mode**

Importance of Measures of Central Tendency



- 1. They help to find representative value for a distribution
-i.e., they are used as single scores to describe data
- 2. They are used to condense data
-i.e., to collect and classify figures, say weight of 3 months old babies from 100 mothers, an average converts the whole set of figures into just one figure
- 3. They help to make comparisons of two or more than two distributions
-e.g., to know the level of performance of general nursing students in anatomy and medicine by comparing with a given standard of performance, where the average is a standard, such as the mean or median

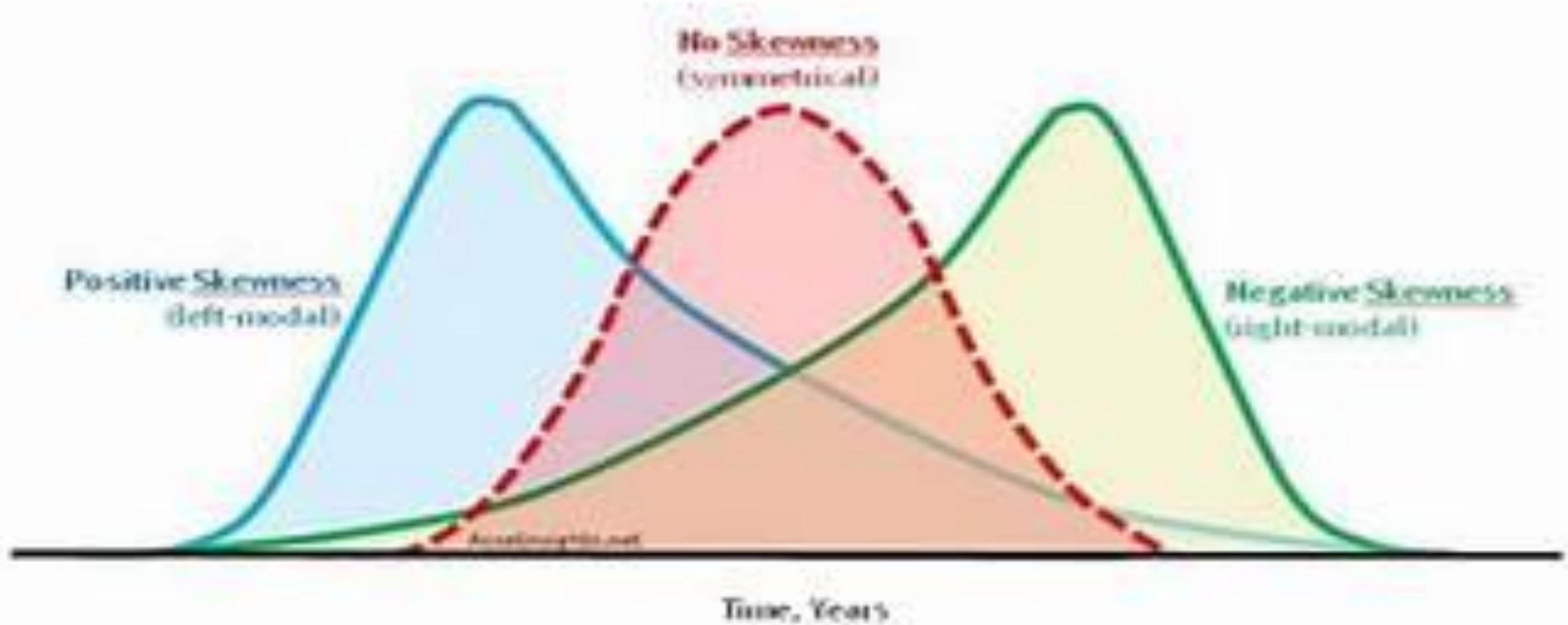
Importance of Measures of Central Tendency



- 4. they are helpful in further statistical analysis
- e.g., many techniques of statistical analysis, such as measures of dispersion, measures of skewness, measures of correlation, and index numbers are based on measures of central tendency
- being the reason why measures of central tendency are also called measures of the first order

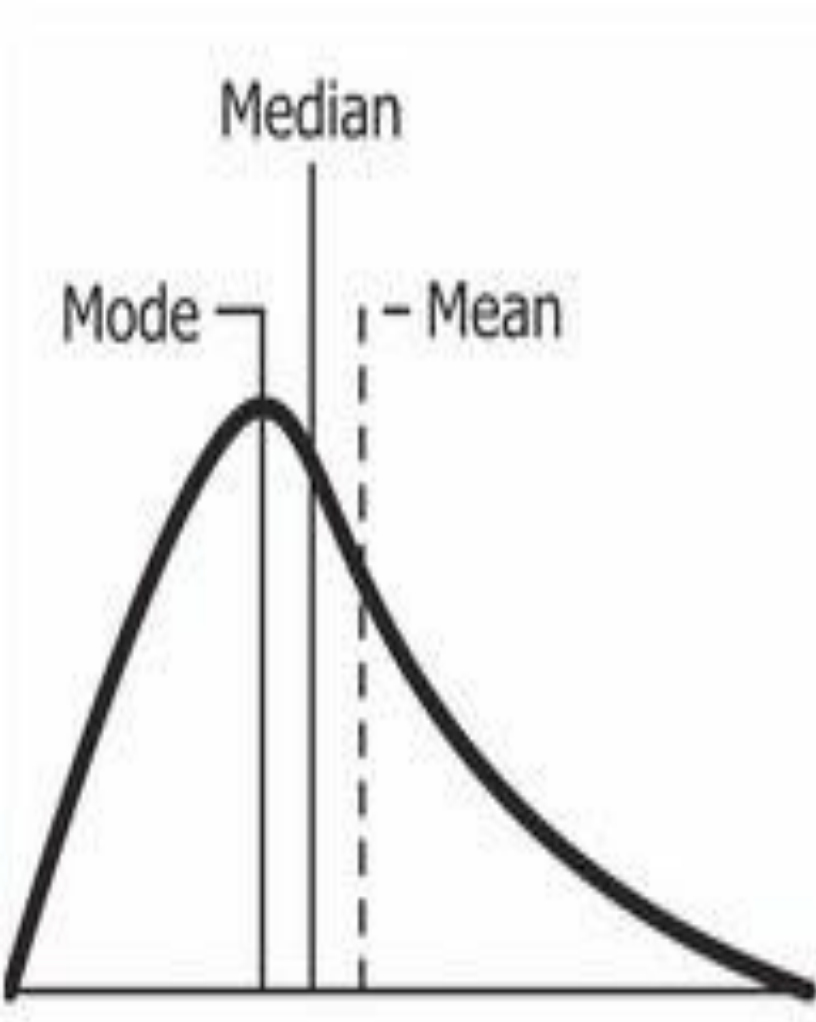


Think through these illustration of measures of central tendency

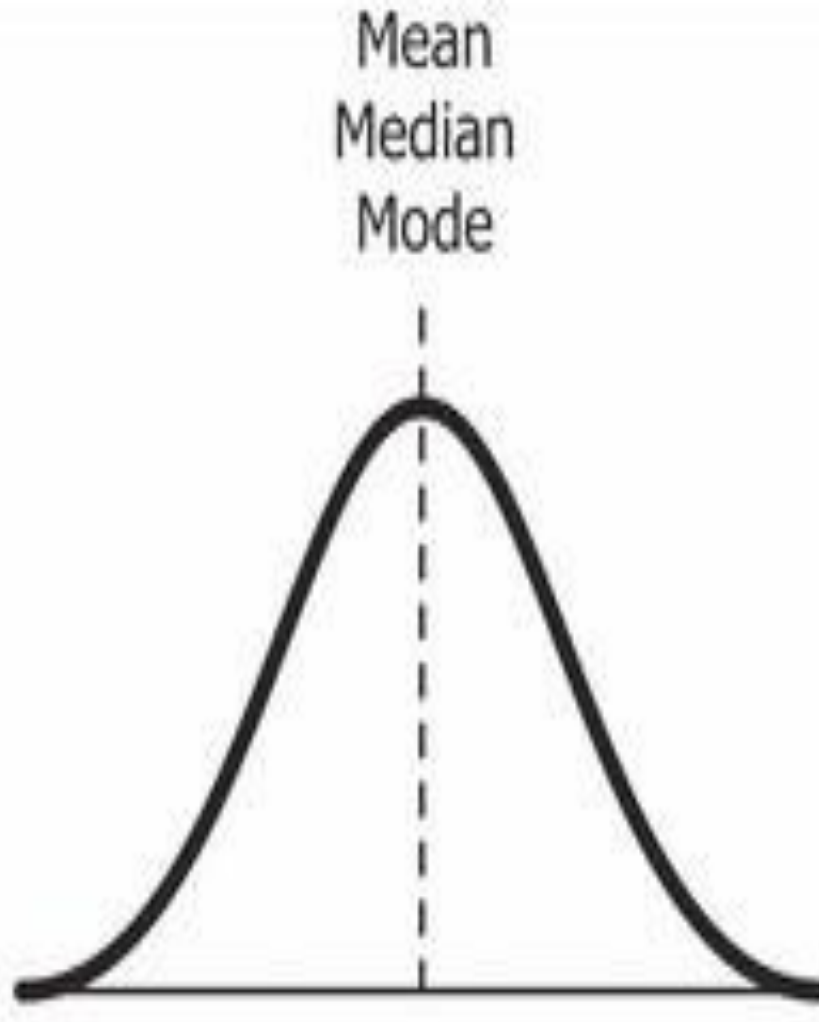




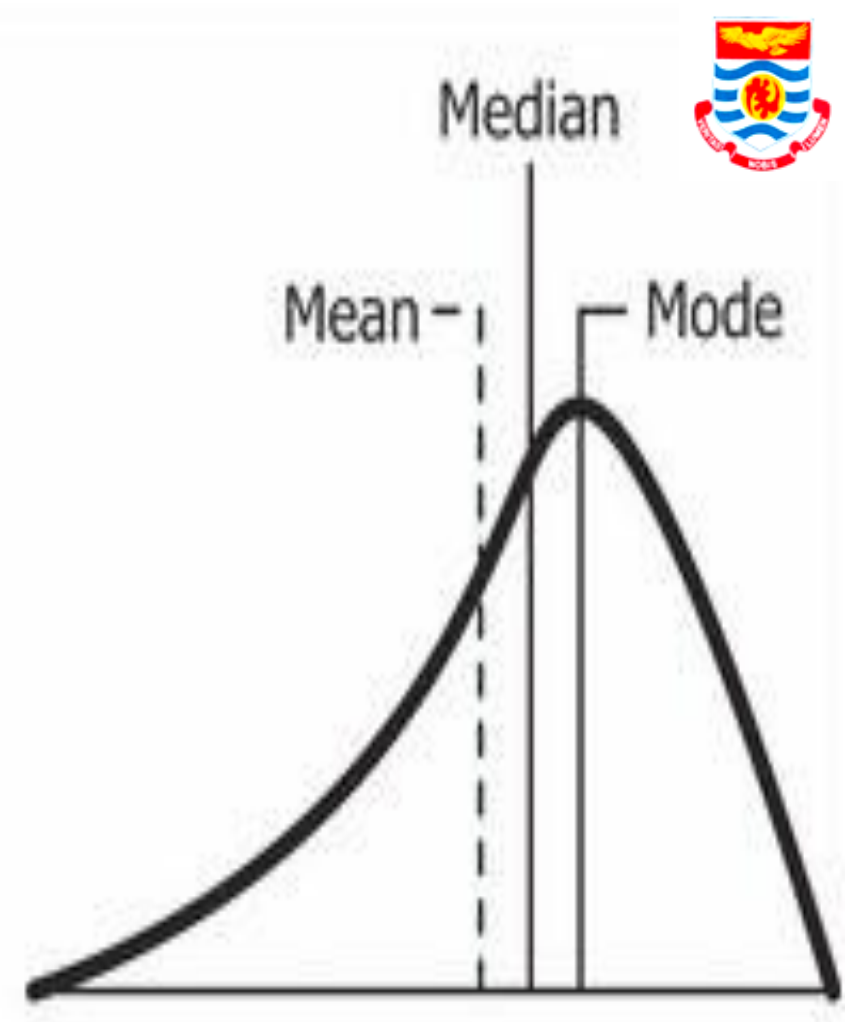
Now, let us
associate the
mean, median,
and mode with
the illustration
above



Positive
Skew



Symmetrical
Normal
Distribution



Negative
Skew



Interpretation of the single value

- Where **Mean > Median**, the distribution is skewed to the right (*positive skewness*)
 - showing that performance tends to be low
- Where **Mean < Median**, the distribution is skewed to the left (*negative skewness*)
 - showing that performance tends to be high
- **Mean = Median = Mode** shows a normal distribution



The Mean

□ The arithmetic mean of a distribution of data is simply the sum of the observations divided by the total number of observations

➤ Symbolically, the mean of a sample is given by; $\bar{X} = \frac{\sum x}{n}$

■ If the number of measures is large, it is desirable to arrange the scores in a grouped frequency distribution, where the mean is given by;

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

○ where **x** is the class midpoint and **f** is the frequency

Properties of the mean



- 1. The mean is influenced by every score or value that makes it up
- If a score is changed, the values of the mean changes.
3, 4, 2, 4, 7 Mean = 4
3, 4, 7, 4, 7 Mean = 5
- The change of the score 2 to 7 has changed the mean to 5

Properties of the mean



-2. The mean is very sensitive to extreme scores (outliers)

4, 2, 3, 6, 5 Mean = 4

4, 2, 23, 6, 5 Mean = 8

- All the scores are below 7 and the presence of 23, an outlier has moved the mean from 4 to 8

Advantages of the mean



- The mean uses every value in the data and hence, is a good representative of the data
- NB:** most of the times this value (i.e., the calculated mean) never appears in the raw data
- Repeated samples drawn from the same population tend to have similar means
- i.e., the mean is the measure of central tendency that best resists the fluctuation between different samples
- The mean is closely related to standard deviation, the most common measure of dispersion

Disadvantages of the mean



- The mean is sensitive to extreme values/outliers, especially when the sample size is small.
- i.e., the mean is not an appropriate measure of central tendency for skewed distribution
- The mean cannot be calculated for nominal or non-nominal ordinal data
- Even though mean can be calculated for numerical ordinal data, many times it does not give a meaningful value
- e.g., stage of cancer
- *You can read further on weighted mean, geometric mean, and harmonic mean*

The Median



- The median is the middle value in a dataset when scores are arranged sequentially
- i.e., to determine the median, the individual values in the dataset is arranged from the smallest to the largest and finding the middle value (score)
- i.e., the median is 50% above, 50% below
- The median is not affected by extreme values
- The median of an ordered set of data is located at the ranked value



The Median

- ✓ if the number of values is odd, the median is the middle number
- ✓ If the number of values is even, the median is the average of the two middle numbers
 - e.g., to determine the median number of the weights of 10 children being attended to by a community-health nurse

| | | | | | | | | | | |
|-----------|----|----|----|----|-----|----|----|----|----|----|
| Child | #1 | #6 | #7 | #5 | #10 | #9 | #2 | #4 | #3 | #8 |
| Weight/Kg | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 21 | 22 | 27 |

- Since we have even number of values, the median is the average of the two middle values:

$$\frac{13+14}{2} = 13.5$$

The Median



■ If we had had nine values of the weight of nine children, the arrangement would have been;

| | | | | | | | | | |
|-----------|----|----|----|----|----|----|----|----|----|
| Child | #1 | #6 | #7 | #5 | #9 | #2 | #4 | #3 | #8 |
| Weight/Kg | 8 | 9 | 11 | 12 | 14 | 15 | 21 | 22 | 27 |

- In this case, since we have an odd number of values, the median is the middle value; 14

The Median



- For a small sample, the median is found by arranging the data in order of magnitude and selecting the middle number
- With large datasets, which have been arranged in a grouped distribution table, the median is given by;

The Median



$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:

n = the **total frequency**

F = the **cumulative frequency *before*** class median

f_m = the **frequency** of the class median

i = the **class width**

L_m = the **lower boundary** of the class median

The Mode



□ The mod is the value that occurs most frequently in a distribution of a dataset

- A dataset can have no mode (if no value repeats),

- no mode, or

- multiple modes

- e.g., there is no mode among the weights of 10 children being attended to by a community-health nurse

| | | | | | | | | | | |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| Child | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 |
| Weight/Kg | 8 | 9 | 11 | 12 | 13 | 14 | 15 | 21 | 22 | 27 |

The Mode



-In the following dataset the mode among the weights of 10 children being attended to by a community-health nurse is **15**

| Child | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| Weight/Kg | 8 | 9 | 11 | 12 | 13 | 15 | 15 | 21 | 22 | 27 |

The Mode



-In the following dataset the mode among the weights of 10 children being attended to by a community-health nurse **8 15, 19**

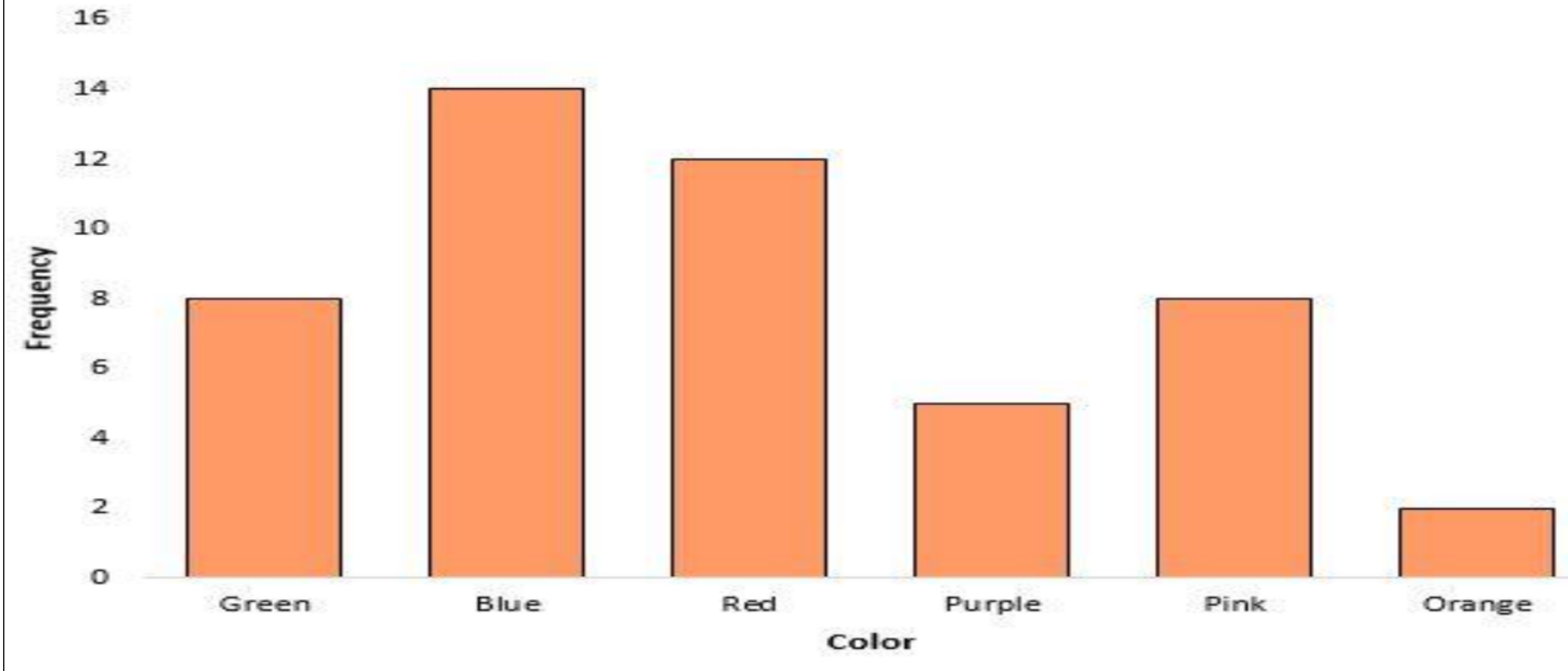
| Child | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 |
|-----------|----|----|----|----|----|----|----|----|----|-----|
| Weight/Kg | 8 | 8 | 11 | 12 | 13 | 15 | 15 | 19 | 19 | 27 |

The Mode



- The mode is not affected by extreme values
- The mode is used for either numerical or categorical data
 - because it tells us which category occurs most frequently
 - e.g., in a survey of exploring the colour blindness of first year students attending the eye clinic medicals, the results are presented as;

The Mode

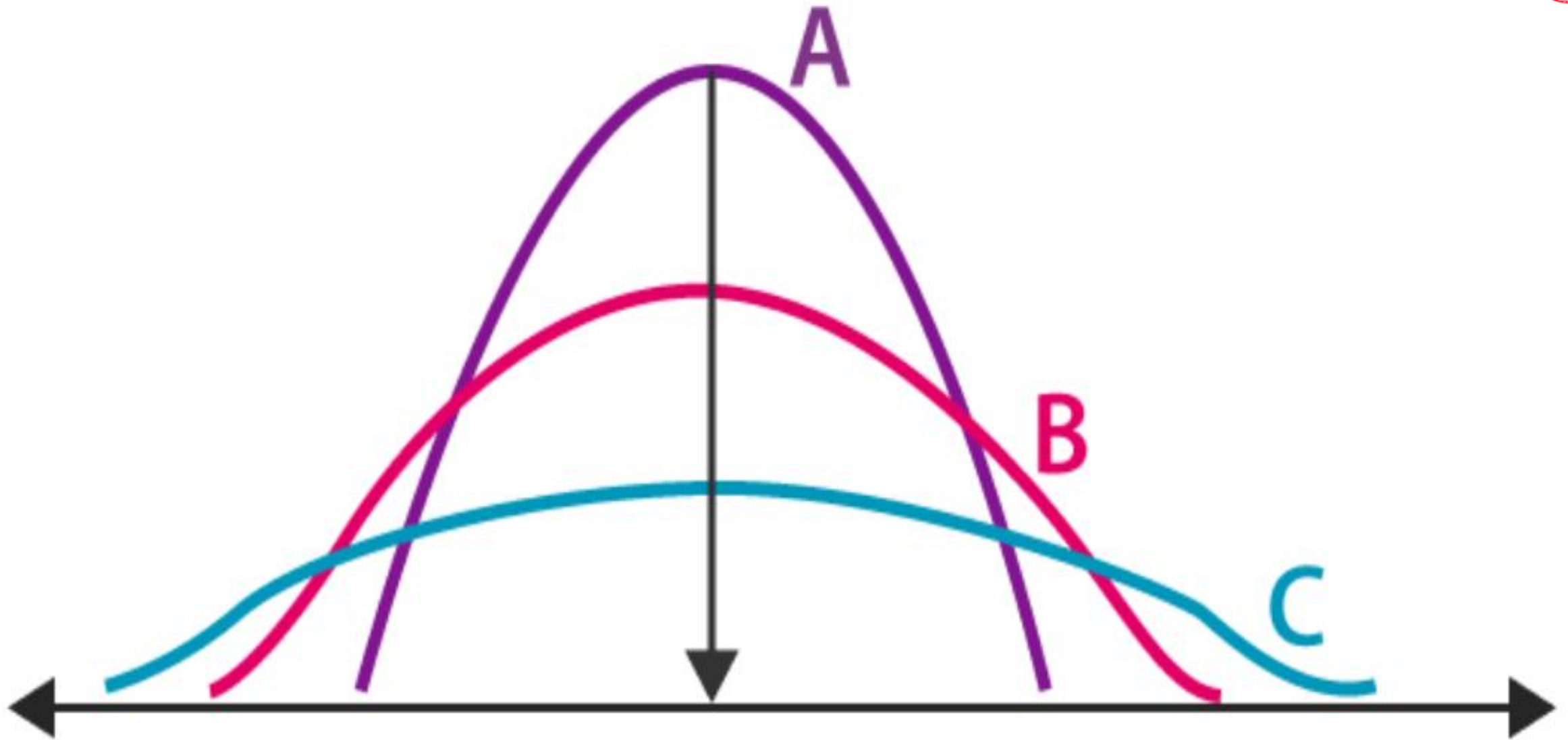


Note the following about Measures of central tendency



- ❖ The mean is generally used unless extreme values (outliers) exist
- ❖ The median is often used since the median is not sensitive to extreme values
- ❖ For data on a nominal scale, only the mode is useful
- ❖ On an ordinal scale, the median and the mode are appropriate
- ❖ For interval and ratio scales, all three could be used

Observe this figure carefully



Measures of Dispersion



- ❑ In statistics, the measures of dispersion help to interpret the variability of data
- i.e., the measures of dispersion help us to know how much homogeneous or heterogeneous the data is
- In simple terms the measures of dispersion tells us how squeezed or scattered the variables is; hence
- ✓ these are also called measures of variation, or scatter
- ✓ they are used as single scores to describe individual differences in terms of achievement

Measures of Dispersion



❖ Categories of Measures of Dispersion

- 1. absolute measures of dispersion
- 2. relative measures of dispersion
- The absolute measures of dispersion are range, variation, standard deviation, and mean deviation
- The relative measures of dispersion are coefficients of dispersion
- In this the measures of dispersion to be learnt are the range, standard deviation, variances, and coefficient of variation

The Range



□ The range is the difference between the highest and the lowest values in a dataset

-i.e., the range is the difference between the maximum value and the minimum value of a dataset

-e.g., 48, 51, 47, 50

-Mean = 49 Range = $51 - 47 = 4$

-1, 3, 5, 6, 7

-Mean = 22 Range = $7 - 1 = 6$

➤ The range does not consider the typical observations in the distribution but concentrates only on the extreme values

The Interquartile Range



- Problems caused by outliers can be eliminated by using the **interquartile range**.
- **Interquartile range** = 3rd quartile – 1st quartile
$$= Q_3 - Q_1$$
- Q_3 = 75% of the dataset
- Q_1 = 25% of the dataset
- Quartiles are values that divide a list of numbers into quarters
- Quartile deviations is half of the distance between the third and first quartiles

Uses of the range



- When data is too scanty or too scattered to justify the computation of a more precise measure
- When knowledge of extreme scores or total spread is all that is needed

Variance and Standard Deviation



- The variance is always considered together with the standard deviation
- ❖ The variance deduct the mean from each data in the set, square each of them and add each square and finally divide them by the total number of values in the dataset
- $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
- The variance is the squared of the standard deviation
- ❖ Standard Deviation (S, σ) is the square root of the variances
- i.e., Std. = $\sqrt{\sigma}$



1. The standard deviation is used as the most appropriate measure of variation/dispersion when there is reason to believe that the distribution is normal
2. The standard deviation helps to find out the variation in dataset (e.g., in achievement among a group of students)
-i.e. the standard deviation determines if a group is homogeneous or heterogeneous

Uses of standard deviation



- Where the standard deviation is **relatively small**, the group is believed to be **homogeneous**

- i.e., all values (scores of students' performance) at about the same level
- Where the standard deviation is **relatively large**, the group is believed to be **heterogeneous**
- i.e. performing at different levels
- To be more precise, the coefficient of variation (CV) is computed

Coefficient of Variation



- The **CV** is a statistical measure of the dispersion of data points in a data series around the mean
- The CV is a simple ratio, involving the division of the standard deviation by the mean

$$CV = \frac{\text{Standard deviation}}{\text{Mean}}$$

- i.e., the value of CV tell us the relative size of the standard deviation compared to the mean of the distribution of measurements taken
- The CV is often reported by researchers as a percentage
- e.g., 25.0% of the size of the mean if the standard deviation was 5 and the calculated mean was 20

Coefficient of Variation



- If the value of CV equals **1** or **100%**, the standard deviation equals the mean
- CV values less than **1** indicate that the standard deviation is smaller than the mean
- This is very common result to obtained from a distribution of a dataset
- when the value of CV is close to zero, it becomes sensitive to small changes in the mean
- While CV values greater than **1** occur when the standard deviation is greater than the mean
- In general, *higher CV values represent a greater degree of relative variability*



➤ Since CV is a ratio, it has no unit and this provides it with some advantages; $CV = \frac{\sigma}{\mu}$

- 1. CV facilitates meaningful comparisons in scenarios where absolute measures cannot
- 2. CV is useful when we want to compare variability between groups that have means of very different magnitudes
- 3. CV is useful when we want to compare variability between characteristics that use different units of measurements

➤ In most score interpretations in biological or medical sciences and for descriptive statistics, the standard deviation is preferred to variance because

1. the standard deviation (σ), is the natural measure of spread or variation for normal distributions
2. the variance (σ^2) involves squaring the deviations and does not have the same unit of measurement as the original observations

Inferential Statistics



- ❑ Inferential statistics infer from the sample to the population
- ❑ Inferential statistics help us determine probability of characteristics of population based on the characteristics of our sample
- ❑ Inferential statistics help us to assess strength of the relationship between our independent variables and the dependent variables

Reasons for Using Inferential Statistics



- The idea of inferential statistics is to be able to make conclusions and make predictions based on a dataset
- ✓ Hence, there are **two** reasons for using inferential statistics in biological and medical sciences;
 - 1. Inferential statistics will allow us to make estimates about populations
 - e.g., the mean score of 50 undergraduate students on medical nursing can be used to estimate the performance of a population of 205 students

Reasons for Using Inferential Statistics



- 2. Inferential statistics will allow us to test hypotheses to draw conclusions about populations
- i.e., to make generalisation of our findings to the larger population
- to assess strength of the relationship between our independent variables and the dependent variables
- Many top-tiered journals will not publish articles that do not use inferential statistics



➤ There are two forms of inferential statistics

- 1. parametric tests

- 2. non-parametric test

- NB:** parametric tests are used when sampling is achieved through probability sampling procedures and the scores of samples are normally distributed

- When assumptions underpinning some parametric tests are violated, the alternative non-parametric tests should be selected

Examples of Inferential Statistics



- Examples of inferential statistics are
- 1. t-test – *parametric*
- 2. Analysis of variance – *parametric*
- 3. Mood's median - *non-parametric*
- 4. Wilcoxon signed-rank – *non-parametric*
- 5. Mann-Whitney U – *non-parametric*

Examples of Inferential Statistics



- 6. Kruskal-Wallis test – *non-parametric*
- 7. Pearson's r – *parametric*
- 8. Spearman's r – *non-parametric*
- 9. Chi square test of independence – *non-parametric*
- 10. Simple linear regression – *parametric*
- 11. Multiple linear regression – *parametric*
- 12. Logistic regression – *parametric*



t-test

- t-tests are statistical tests used when you have two groups and you want to compare on their mean scores

- T-tests are often used in testing hypotheses to determine whether a process or a treatment actually has an effect on the population of interest or whether two groups are different from one another
 - e.g., male and female or married and single and their attitude towards attending medical facility for treatment of ill health
- or to compare means of two sets of data (before and after)
 - e.g., pre-test and post-test scores
- or to compare the mean scores on some continuous variable
 - e.g., perceptions of patient on taking medication before meals

Types of t-test



➤ There are three main types of t-test

- 1. independent-samples t-test
- 2. paired-samples (related-samples) t-test
- 3. one-sample t-test

➤ Selection of a particular t-test depends on whether the groups being compared come from

- 1. the groups being compared come from two different populations
- 2. the groups being compared come from a single population
- 3. to test the difference in a specific direction

t-test formula



➤ The formula for the two-samples (independent-samples) t-test is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- Where t is the t value
- \bar{x}_1 and \bar{x}_2 are the means of the two groups being compares
- S^2 is the variance of the two groups
- n_1 and n_2 are the number of observations in each of the groups

t-test formula



- A large t value shows that the difference between group means is greater than the variance
- Also, indicating that the difference between the two groups is significant

Standard Scores (Z, T)



- Standard scores indicate the number of standard deviation units an individual score is above or below the mean of each group
- The standard score represents an individual score that has been transformed into a common standard using the mean and the standard deviation
- The standard score (commonly referred to as z-score) is a very useful statistic because it;
 - 1. it will allow us to calculate the probability of a score occurring within our normal distribution
 - 2. it will allow us to compare two scores that are from different normal distributions

The Z standard score



➤ The Z-scores are expressed in terms of standard deviations from their means

■ The Z-scores have a distribution with a mean of zero and a standard deviation of **1**

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

- where Z is the standard score

X is the score

μ is the mean

σ is the standard deviation

Z-Score



➤ Worked example

■ How well did Dankwa, a first year student, perform in his Human Physiology 1 course compared to her other 50 colleagues, if he scored 70 out of 100, of a mean score of 60 and standard deviation of 15?

■ Solution

$$■ Z = \frac{x - \mu}{\sigma}$$

$$■ \text{If } x = 70, \mu = 60, \sigma = 15$$

$$■ Z = \frac{70 - 60}{15} = 0.6667$$

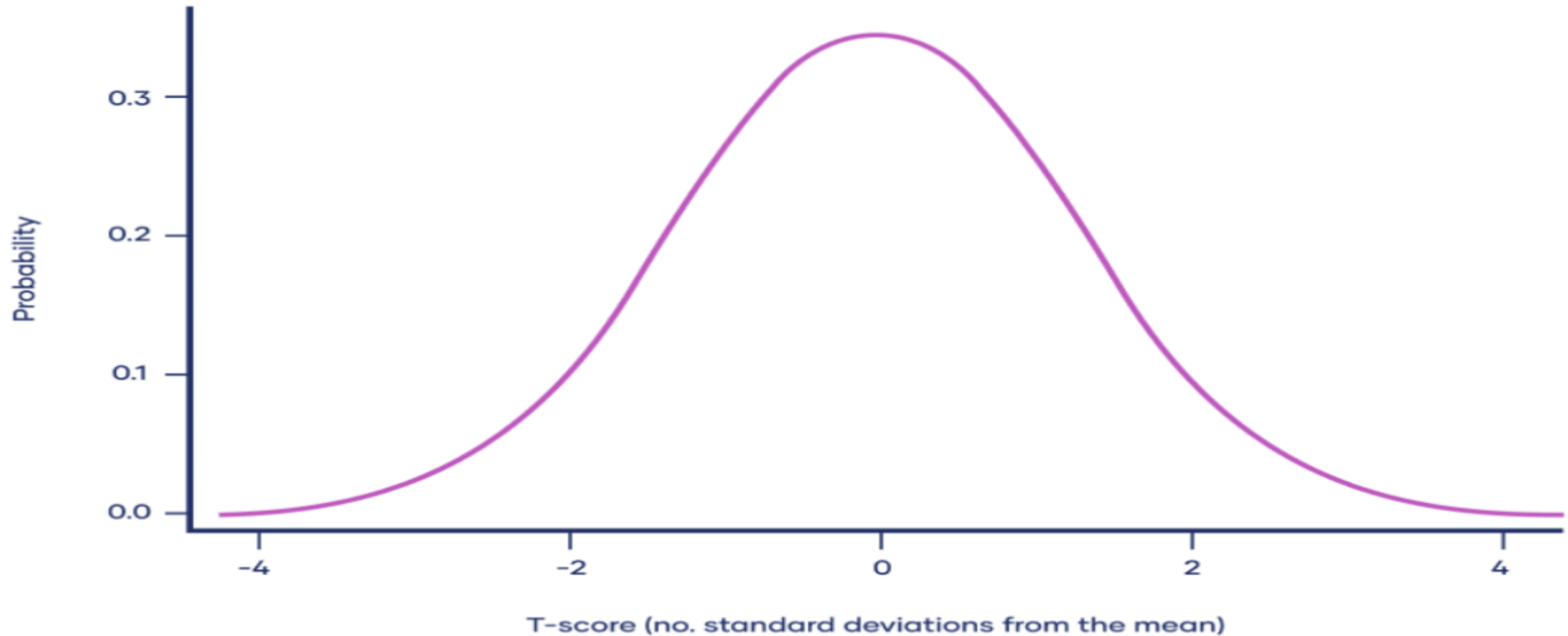
$$■ Z = 0.67$$

T-Standard Scores



- T-score is a way of describing data that follow a bell curve when plotted on a graph, with greatest number of observations close to the mean and fewer observations in the tails
- i.e., T-standard scores give a distribution used for smaller sample sizes, where the variance in the data is unknown
- **NB:** A normal distribution of observations (or values) form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tail

A normal distribution



T- standard score



➤ T-standard score can be calculated from

$$T = 50 + 10Z$$

■ where mean is 50 and standard deviation is 10

-e.g., what are the Z- and T-standard scores when an undergraduate student obtained 15 scores in Introduction to Mental Health Nursing quiz 1 with a mean of 12 and a standard deviation of 2?

-Solution

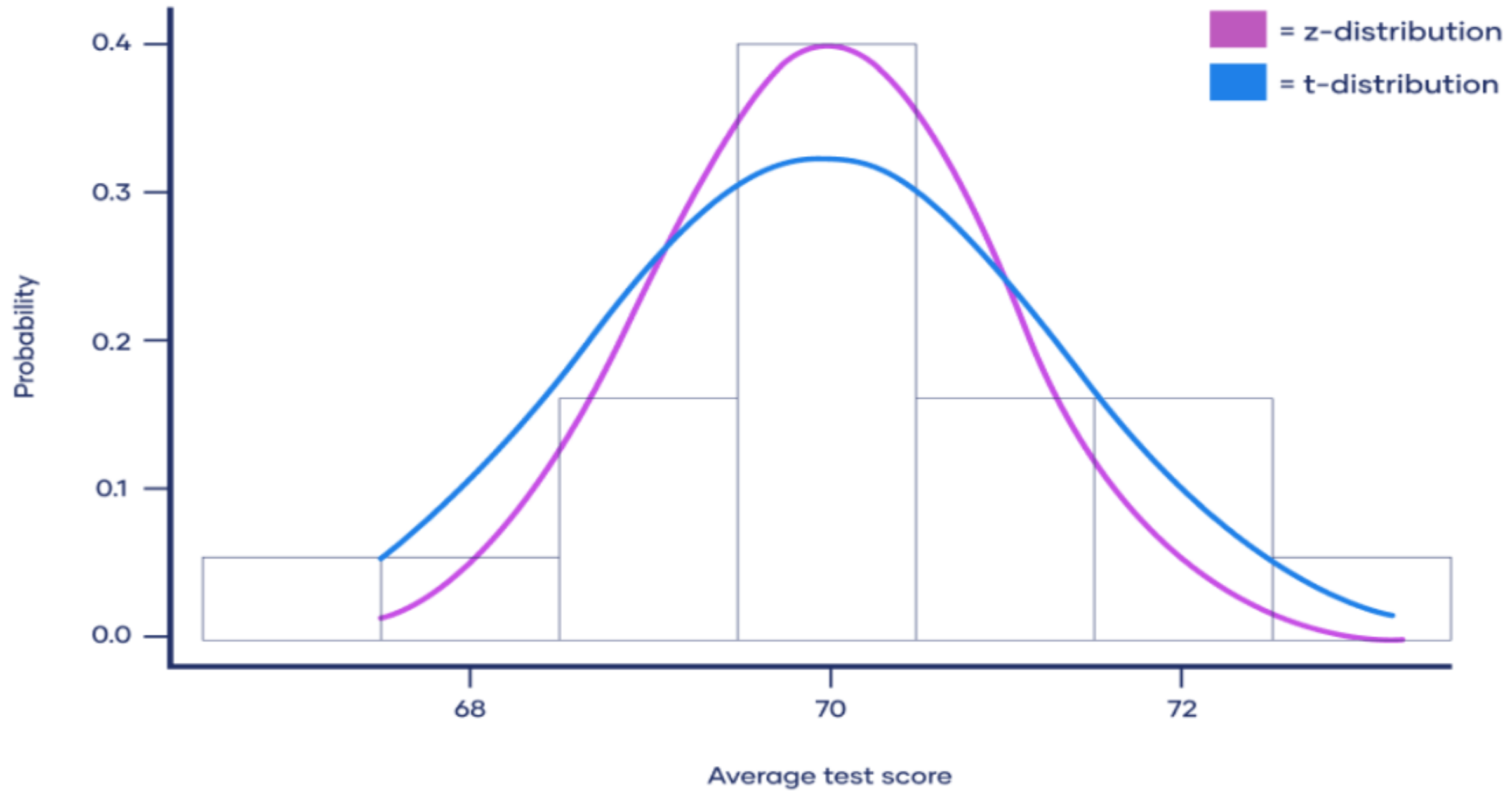
$$-Z = \frac{x - \mu}{\sigma} = \frac{15 - 12}{2} = 1.5$$

$$-T = 50 + 10Z = 50 + 10(1.5) = 65$$

T-standard scores



- The T-distribution is a more conservative form of the standard normal distribution (z-distribution)
- i.e., the T-distribution gives a lower probability to the centre and a higher probability to the tails than the z-distribution
- If we measure the average test score in Fundamentals of Mental Health Nursing from a sample of only 20 students, we will use the t-distribution to estimate the confidence interval around the mean
- If we use the z-distribution, the confidence interval will be artificially precise



Comparing T-distribution and z-distribution

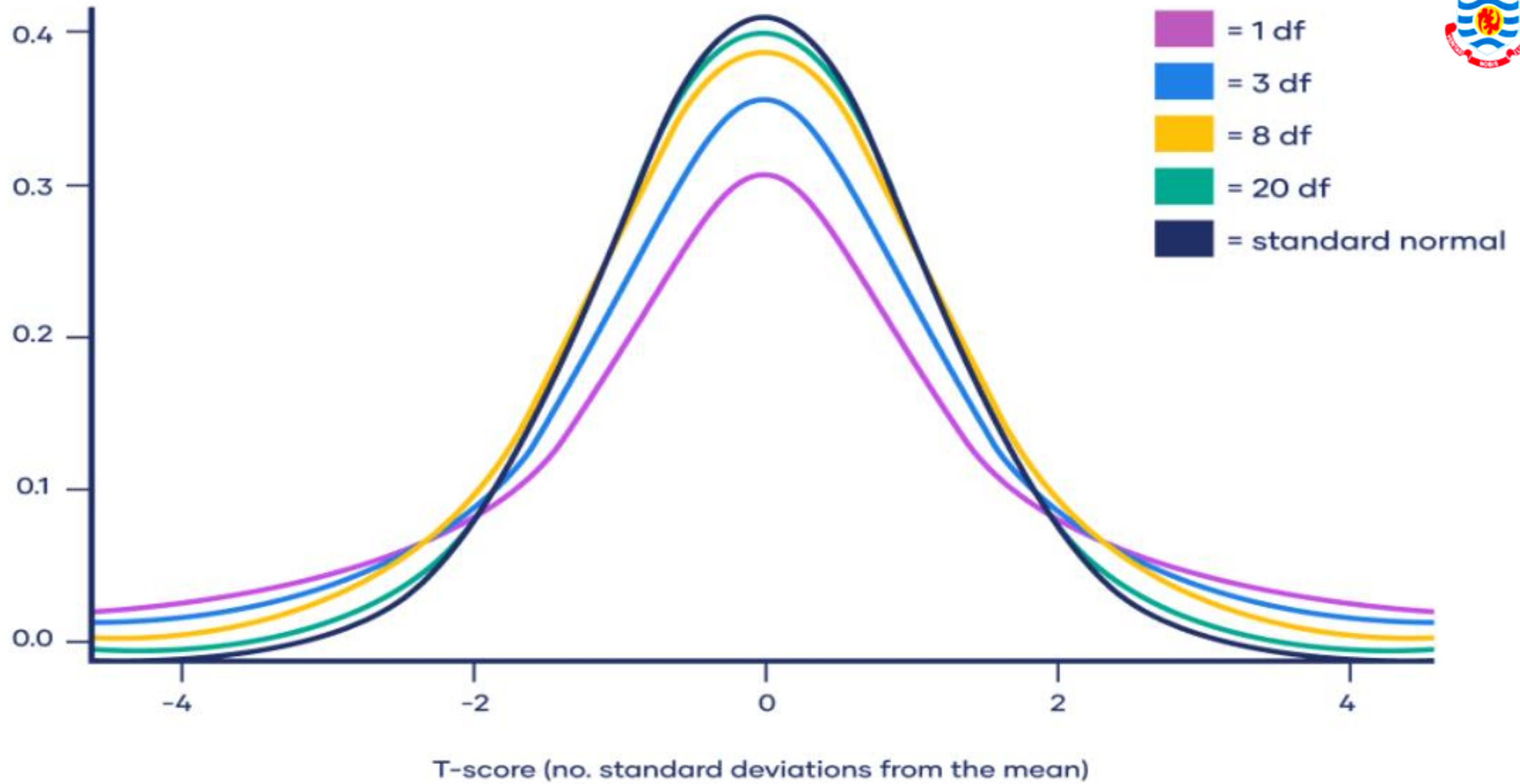


➤ The z-distribution is preferred over T-distribution when we are to make statistical estimates because we know the variance

■ i.e., it is more precise to estimate with the z-distribution instead of the T-distribution, when the variance is approximated using the degrees of freedom of the data



Probability



Features of Z- and T-scores



- For z-scores, 0 is the mean score, positive scores are scores above the mean, and negative scores are scores below the mean
- An individual's performance can be described as far above average, above average, just above average, just below average and far below average
- In case of T-scores, 50 is the mean score, and scores greater than 50 are above average, and scores less than 50 are below average
- Z-scores ranges between -4 and +4 while T-scores are between 10 and 90

Uses of Standard Scores



- 1. It helps the teacher to know an individual's position in relation to the rest of the class
- e.g., a student with a Z-score of 3.2 is performing far above average
- 2. It enables the teacher to compare student's performances in different subjects to know individual strengths and weaknesses

Paired-samples t-test



- ❖ The paired-samples t-test is used when we want to compare the mean scores for the same group of persons on two different occasions, or when we have matched pairs
 - e.g., test scores of a single group of 100 students in quizzes 1 and 2 on Nursing Perspectives
- ❖ The paired-samples t-test is, also, used when you are interested in changes in scores for participants tested at Time 1, and then again at Time 2
 - e.g., pre-intervention test and post-intervention test of a single group of 200 students experiencing a 21st century skills intervention lesson on Principle and Practice of Health Assessment
- ❖ The samples are **related** because they are the same people tested each time

Independent-samples t-test



- The independent-samples t-test is used when you want to compare the mean scores of two different (independent) groups of people or conditions
 - e.g., male and female nutritionists; endowed and less-endowed hospitals; professional and non-professional health-care workers, experienced and novice midwives, certificate and diploma nurses
- In this case, you collect information on only one occasion but from two different sets of people
- e.g., comparing certificate and diploma midwives attitude towards pregnant women due for delivery



❖ Level of measurement

- the dependent variable should be measured at the interval or ratio level, being **continuous**
- the independent variable should be categorical at two levels

❖ Random sampling

- Scores should be obtained from random sampling from the population

Assumptions for t-test



❖ Independence of observations

- The observations that make up our data must be independent of one another
 - i.e., each observation or measurement must not be influenced by any other
 - violation of this assumption is very serious

❖ Normal distribution

- It is assumed that the populations from which the samples are taken are normally distributed
 - e.g., in a lot of research scores on the dependent variable are not nicely normally distributed
 - With large enough sample sizes (e.g., 30+), the violation of this assumption should not cause any major problems

Hypothesis Testing



- A **statistical hypothesis** is a formal claim state of nature of structured within the framework of a statistical model
- ✓ i.e., a statistical hypothesis is about the nature of a population, often stated as a population parameter
- ✓ In terms of numerical statement about an unknown parameter
- Parameters are summary values describing a population

Hypothesis testing



- These summary values (parameters) include;
 - population mean (μ)
 - population variance (σ^2)
 - population standard deviation (σ)
- These summary values (parameters) are difficult to obtain, hence numerical guesses are made about what the likely values would be
 - NB: These guesses are **statistical hypotheses**

Hypothesis testing



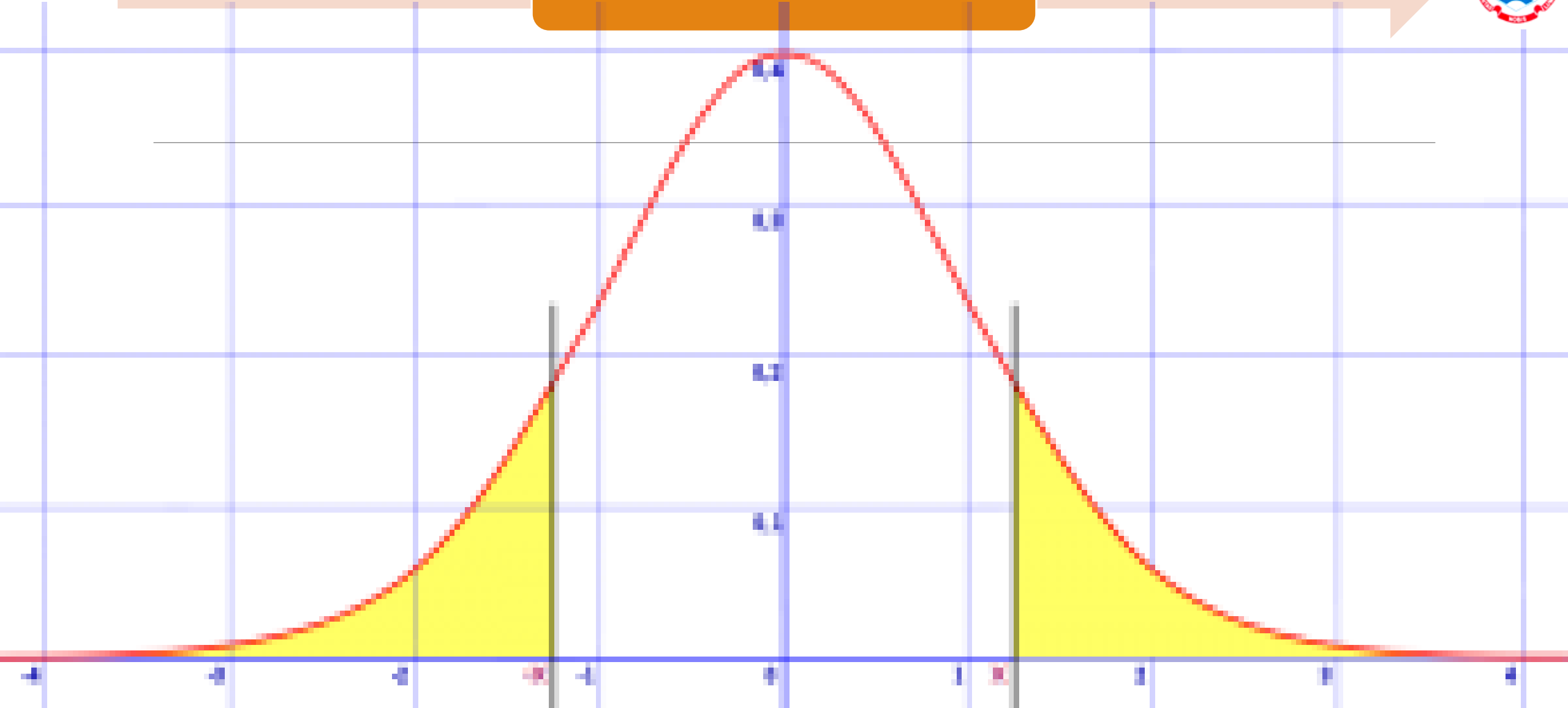
- The guesses are tested under various conditions to find out how close they are to the **actual values**
- The guessed values, which form the hypothesis, are stated in a form called **null hypothesis** (H_0)
- A **null hypothesis** is a statistical hypothesis to be tested
- An **alternative hypothesis** (H_1) is also provided, in case the **guess** is not correct

Hypothesis testing



- ❖ Statistical hypothesis may be one-tailed or two-tailed
 - i.e., a two-tailed test tells us that we are finding the area in the middle of a distribution, and it is the place where the researcher will reject the stated null hypothesis
- This gives an indication that the hypothesis is non-directional
- Phrases like *is equal to (say 25%) or is (25%)* suggest that it is either less than 25% or greater than 25% (as seen in the next figure)
- Hence, the **one-tailed test** is directional suggesting say *less than 25% or greater than 25%*

A two-tailed test



A two-tailed test



- NB:** normally a two-tailed test is used when the researcher has some doubts (or is uncertain) and the one-tailed is used when the researcher is sure (or certain)
- The significance level of 5% will be divided into two as 2.5% for both sides as seen in the figure above
- e.g., a two-tailed test is ideas for situation as: *is the mean greater 10? Is the mean less than 10?*
- The significance level of 5% remains the same for a one-tailed test

Hypothesis testing



- Also, statistical hypothesis could also be directional or non-directional
- For a **directional hypothesis**, the alternative hypothesis contains the *less than or the greater than expression*
 - Meaning we are testing whether or not there is a positive or negative effect
- For **non-directional hypothesis**, the alternative hypothesis contains the *not equal expression*
 - Meaning we are testing whether or not there is some effect, without specifying the direction of the effect
- **NB:** Null hypothesis is used for **decision making** while alternative hypothesis is used to **show direction**



$$H_1: \mu = 22$$

$$H_0: \mu \neq 22$$

■ Is this a directional or non-directional hypothesis

■ **Solution**

✓ This is non-directional because the alternative could be greater than 22 or less than 22

✓ Non-directional hypothesis are two-tailed because they show two directions for the alternative

Analyse of a question



$$H_1: \mu = 25$$

$$H_0: \mu < 25$$

- Is it directional or non-directional?

- Is it one tailed or two-tailed?

- **Solution**

- ✓ This is one-tailed. Mean age could be less than 25

- ✓ This is directional because the alternative is stated as less than 25

- ✓ Directional hypothesis are one-tailed because they only show **one direction** for the alternative

Examples of directional hypothesis



- As sleep deprivation increases, cognitive performance decreases
 - As carbon dioxide levels increase, global temperatures also increase
 - Diabetic patients are more prone to sexual weaknesses than non-diabetic individuals
 - Experienced midwives have positive attitude towards pregnant women than novice midwives
- **See if you can convert any of these to a non-directional hypothesis**



- To perform the hypothesis test, a sample is taken and the sample values (called statistics) are computed and used
- The sample values help in taking one of two decisions based on the null hypothesis
- The null hypothesis may be **rejected** or one may **fail to reject it**

Decision Making



- Where the null hypothesis is **rejected**, the difference between the population parameter and the sample value (sample statistic) is considered **significant or real**
- If the null hypothesis is **not rejected**, the difference between the population parameter and the sample value (sample statistic) is considered **not significant or due to chance**

Level of Significance, α



- **NB:** In hypothesis testing, decisions made could be wrong; **why?**
- The degree of risk taken in making a wrong decision is the **level of significance**
- The level of significance is the degree of risk involved in rejecting a true null hypothesis
- The level of significance provides an estimate of confidence in taking the risk
- ❖ Thus, the level of significance is the measurement of statistical significance where the researcher is convinced that the result is **not by chance**

Level of Significance, α



- **Critical values** are the sample values beyond which the null hypothesis is rejected
- The level of significance is expressed as $(1 - \alpha)$
- ✓ The commonest significance levels for biological or medical sciences are **0.05** or **0.01**
- e.g., if the α value equals 0.05, it indicates that there are just 5% chances of getting a difference larger than that in our research, giving that the null hypothesis exist



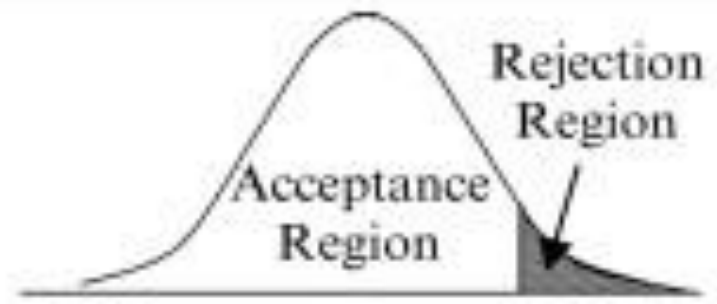
Level of Significance, α



- ✓ The level of significance helps in the decision on the critical region, the region of rejection
- ✓ A region of rejection is a set of possible values of the test statistic that causes the null hypothesis to be rejected
- ✓ Hence the level of significance, α is the measure of the strength of the evidence that must be present in our sample before we will reject the null hypothesis and conclude that the effect is statistically significant

Illustration of level of significance, α



| One-Tailed Test (Left Tail) | Two-Tailed Test | One-Tailed Test (Right Tail) |
|---|---|--|
| $H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$ | $H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$ | $H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$ |
|  |  |  |

Level of Significance and Decision Making



- When we obtained significance value which is **greater than 0.05 or 0.01** as the set level of significance, **the null hypothesis is not rejected but accepted**
 - We then say; we failed to reject the null hypothesis
- However, if the obtained significance value is **less than 0.05 or 0.01**, **the null hypothesis is rejected** in favor of the alternative hypothesis
- The result is said to be **statistically significant**



- Alternatively, with the critical values determined, the observed test statistic is compared with the critical values
- If the observed test statistic value falls in the region of rejection, the null hypothesis is rejected in favor of the alternative hypothesis
- The result is said to be statistically significant
- If the test statistic value does not fall in the critical region, one fails to reject the null hypothesis

Level of Significance and Decision Making



■ Using critical z-values

| <i>Level of Significance</i> | <i>Two-Tailed Test</i> | <i>One-Tailed Test</i> |
|------------------------------|------------------------|------------------------|
| 0.10 = 10% | ± 1.65 | +1.28 or -1.28 |
| 0.05 = 5% | ± 1.96 | +1.65 or -1.65 |
| 0.01 = 1% | ± 2.58 | +2.33 or -2.33 |

❖ Type I Error

- Type I error is a statistical concept, referring to rejection of an accurate null hypothesis
- Simply, type 1 error is to reject a true null hypothesis
- Type 1 error is usually seen as a false positive conclusion
- It is considered a serious type of error

Errors in Decision Making



- The probability of making a type I error is α (the level of significance) set by researcher in advance
- e.g., In medical testing, a type I error would cause the appearance that a treatment for a disease has the effect of reducing the severity of the disease when, in fact, it does not
- When a new medicine is being tested, the null hypothesis will be that the medicine does not affect the progression of the disease. Let us say a lab is researching a new cancer drug. Our null hypothesis might be that the drug does not affect the growth rate of cancer cells. After applying the drug to the cancer cells, the cancer cells stop growing. This would cause us to reject their null hypothesis that the drug would have no effect. If the drug caused the growth stoppage, the conclusion to reject the null, in this case, would be correct. However, if something else during the test caused the growth stoppage instead of the administered drug, this would be an example of an incorrect rejection of the null hypothesis



❖ Type II Error

- Type II error is a false negative conclusion, involving failure to reject false null hypothesis
- The probability of making a type II error is β
- This risk can be minimized through careful planning in our study design
- e.g., when the test results says you do not have coronavirus, but you actually do

Errors in Decision Making



Possible Hypothesis Test Outcomes

| Decision | Accept H_0 | Reject H_0 |
|----------------|-----------------------------|-----------------------------|
| H_0 is true | Correct Decision (No error) | Type I Error |
| | Probability = $1 - \alpha$ | Probability = α |
| H_0 is false | Type II Error | Correct Decision (No error) |
| | Probability = β | Probability = $1 - \beta$ |



- The **power** of an experiment that you are about to carry out quantifies the chance that you will correctly reject the null hypothesis if some alternative hypothesis is really true
- i.e., the statistical power (or **sensitivity**) is the probability of a significance test detecting an effect when there actually is one
 - A true effect is a real, non-zero relationship between variables in a population
 - An effect is usually indicated by a real difference between groups or a correlation between variables

Power ($1 - \beta$)



- High power in a research indicates a large chance of a test detecting a true effect
- Low power means that our test only has a small chance of detecting a true effect
- or that the results are likely to be distorted by random and systematic error
- Power is mainly influenced by
 - sample size
 - effect size
 - significance level



- Having enough statistical power is necessary to draw accurate conclusions about a population using sample data
- In hypothesis testing, we start with null and alternative hypotheses
 - a null hypothesis of no effect and an alternative hypothesis of a true effect (our actual research prediction)
 - Our aim is to collect enough data from a sample to statistically test whether we can reasonably reject the null hypothesis in favor of the alternative hypothesis



➤ If we have a research question concerning whether spending time outside in nature can curb stress in B.Sc. (Nursing) graduates, we can then craft a null and alternative hypotheses as;

- **Null hypothesis:** Spending 10 minutes daily outdoors in a natural environment has no effect on stress in recent B.Sc. (Nursing) graduates
- **Alternative hypothesis:** Spending 10 minutes daily outdoors in a natural environment will reduce symptoms of stress in recent B.Sc. (Nursing) graduates

Power ($1 - \beta$)



- Power is the probability of avoiding a Type II error
 - i.e., the higher the statistical power of a test, the lower the risk of making a Type II error
- Which is when we conclude that spending 10 minutes in nature daily does not affect stress when it actually does
- Power is usually set at 80%
 - i.e., if there are true effects to be found in 100 different studies with 80% power, only 80 out of 100 statistical tests will actually detect them
- If we do not ensure sufficient power, our research may not be able to detect a true effect at all

Power ($1 - \beta$)



- too much power means our tests are highly sensitive to true effects
- e.g., very small ones
- Resulting we finding statistically significant results with very little usefulness in the real world
- To balance these pros and cons of low versus high statistical power, one should use a power analysis to set an appropriate level



- A **power analysis** is a calculation that helps in determining a minimum sample size for a research
- A power analysis is made up of four main components
- If we have estimates for any three of these, we can calculate the fourth component
- **1. Statistical power:** the probability that a test will detect an effect of a certain size if there is one, usually set at 80% or higher
- **2. Sample size:** the minimum number of observations needed to observe an effect of a certain size with a given power level



- 3. **Significance level (alpha):** the maximum risk of rejecting a true null hypothesis that we are willing to take
-i.e., usually set at 5%
- 4. **Expected effect size:** a standardised way of expressing the magnitude of the expected result of our study
-i.e., usually based on similar researches or a pilot research

Power ($1 - \beta$)



➤ Generally, a researcher wishes to increase the power of the test

- 1. For a given value of the parameter being tested, the power of the test of H_0 increases as the sample size, n , increases

- e.g. from 100 to 200

- 2. For a given value of the parameter being tested, the power of the test of H_0 increases as α increases from 0.05 to 0.10

- but this also increases the risk of making a wrong decision

Steps in hypothesis testing



- Identify the form of the population distribution before taking the steps
 - This could be normal, F, χ^2 or t
- 1. State the statistical hypothesis as null or alternative.
- 2. Specify the degree of risk, α , the level of significance
- 3. Determine the critical region or values depending on the probability distribution given
- 4. Compute the test statistic or the significance value
- 5. Make a decision to reject or fail to reject the null hypothesis, H_0
- 6. State the conclusion and relate this to the hypothesis
 - Indicate whether the result is statistically significant or not



- **Confidence Interval** is a range of results from a poll, experiment, or survey that would be expected to contain the population parameter of interest
- i.e., confidence interval is a range of values that is likely to contain a population parameter with a certain level of confidence
- Confidence intervals are constructed using significance levels or confidence levels
- However, confidence level is the probability that if a poll, experiment, or survey was repeated over and over again, the results obtained would be the same



- When a confidence interval and confidence level are put together, the result is a statistically sound **spread of data**
 - e.g., a result might be reported as $50\% \pm 6\%$, with a 95% confidence
- This is interpreted as
 - 1. The confidence interval = $50\% \pm 6\% = 44\%$ *to* 56%
 - The confidence level = 95%

Confidence Interval Estimates



- Confidence intervals are intrinsically connected to confidence levels
- Confidence levels are expressed as a percentage
- e.g., a 90% confidence level
- Should we repeat survey of breast cancer awareness with a 90% confidence level, we would expect that 90% of the time our results will match results we should get from a population
- Confidence intervals are a range of results where you would expect the true value to appear

Degrees of Freedom



- **Degree of Freedom** is the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample
-
- i.e., degrees of freedom are the number of independent variables that can be estimated in a statistical analysis and tell us how many items can be randomly selected before *constraints* must be put in place
 - e.g., within a data set, some initial numbers can be chosen at random
 - However, if the data set must add up to a specific sum or mean, e.g., the number in the data set is constrained to evaluate the values of all other values in a data set, then meet the set requirement
 - Degrees of freedom (**df**) are normally reported in brackets beside the test statistic, alongside the results of the statistical test

Degrees of Freedom



- Consider a data sample consisting of five test scores of students in Medical Microbiology and Parasitology. The values of the five test scores must have an average of six. If four items within the data set are {3, 8, 5, and 4}, the 5th score must be 10. Because the first four numbers can be chosen at random, the degree of freedom is four
- Consider a data sample consisting of five test scores. The values could be any score with no known relationship between them. Because all five can be chosen at random with no limitations, the degree of freedom is four

Degrees of Freedom



- Suppose we randomly sample 10 Covid-19 patients and measure their daily calcium intake. We use a one-sample t -test to determine whether the mean daily intake of Covid-19 patients is equal to the recommended amount of 1000 mg .
- The test statistic, t , has 9 degrees of freedom:

$$df = n - 1$$

$$df = 10 - 1$$

$$df = 9$$

Degrees of Freedom



- We calculate a t value of 1.41 for the sample, which corresponds to a p value of .19. We report our results as:
- The Covid-19 patients' mean daily calcium intake did not differ from the recommended amount of 1000 mg, $t(9) = 1.41, p = 0.19$.

Degrees of Freedom



- When the sample size is **small**, there are only a few independent pieces of information, and therefore only a few degrees of freedom
- When the sample size is **large**, there are many independent pieces of information, and therefore many degrees of freedom

Analysis of Variance (ANOVA)



➤ ANOVA is an analytical tool used in statistics that splits an observed aggregate variability found inside a data set into two parts;

- 1. systematic factors have a statistical influence on the given data set
- 2. random factors do not have influence on the given set of data
- The systematic factors, while the random factors do not
- ANOVA test is used to determine the influence that independent variables have on the dependent variable in a regression study

ANOVA

- ANOVA test will help us to compare data with multiple means across different (independent) groups, and allows us to see patterns and trends within complex and varied data
 - i.e., in using ANOVA, we are interested in comparing the mean scores of more than two groups
 - Where one independent variable (referred to as a factor) has a number of different levels
 - Hence, ANOVA is a statistical method that separates observed variance data into different components to use for additional tests

ANOVA

- Examples of scientific research requiring the use of ANOVA are;
- As an expert you may recommended that a group of psychiatric patients try three different therapies: counseling, medication, and biofeedback. After some period, you want to see if one therapy is better than the others
- A researcher may interested in teacher's pedagogy and students learning. Some Lecturers facilitating the course, Clinical Microbiology were asked to used different participatory learning approaches in facilitating the course for a semester
- Students from different nursing and midwifery colleges take the same end of semester examination. We may want to see if one college outperforms the other

ANOVA

- In the three scenarios created earlier, the investigator controls one or more factors of interest
- Each factor contains **three** or more levels
 - Levels can be numerical or categorical
 - Different levels produce different groups
 - Think of the groups as populations
- Observe effects on the dependent variable
 - Are the groups the same?
 - Experimental design: the plan used to collect the data

ANOVA

- **NB:** The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method
- ANOVA is also called the Fisher analysis of variance
- It is the extension of the t- and z-tests
- The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers"
- It was employed in experimental psychology and later expanded to subjects that were more complex

ANOVA

➤ Mathematically, ANOVA has the formula;

$$F = \frac{MST}{MSE}$$

- Where F is ANOVA coefficient
- MST is mean sum of squares due to treatment
- MSE is mean sum of squares due to error

ANOVA

- ANOVA test offers a lot of help to the researcher;
- 1. The ANOVA test is the initial step in analyzing factors that affect a given data set
- Once the test is finished, a researcher performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency
- 2. The researcher utilises the ANOVA test results in an F-test to generate additional data that aligns with the proposed regression models

ANOVA

- 3. The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them
- i.e., the result of the ANOVA formula, the F statistic allows for the analysis of multiple groups of data to determine the variability between samples and within samples
- If no real difference exists between the tested groups, the result of the ANOVA's F-statistic will be close to 1
- The distribution of all possible values of the F statistic is the F-distribution

❖ Assumptions of ANOVA

- 1. **Normality**; each sample is drawn from a normally distributed population
- 2. **Equal Variances**; the variances of the populations that the same come from are equal
- 3. **Independence**; the observations in each group are independent of each other and the observations within groups were obtained by a random sample

❖ Others are:

- 4. the dependent variable should be continuous (being interval or ratio measurement)
- 5. the independent variable (of a single factor should be of three levels), two independent variables (should be two or three levels)

❖ Forms of ANOVA

- 1. between-subjects ANOVA applied when examining for differences between independent groups on a continuous level variable
 - e.g., one-way ANOVA and factorial ANOVA
- 2. Within-Subjects ANOVA, applied when examining for differences in a continuous level variable over time
 - e.g., when examining for differences over two or more time periods
- 3. Mixed-Model ANOVA, applied when examining for differences in a continuous level variable by group and time
 - e.g., applicable if the purpose of the research is to examine for potential differences in a continuous level variable between a treatment and control group, and over time (pretest and posttest)

One-Way ANOVA



- One-way ANOVA is used to assess/determine differences in one continuous (dependent) variable between **one** grouping or categorical (independent) variable
- i.e., when we want to examine the difference among the means of three levels of categorical groups
- ✓ e.g., What does effective use of instructional resources in teaching postpartum depression in nursing and midwifery colleges meant to nursing officers, senior nursing officers, and principal nursing officers?
- ✓ What is the differences in job satisfaction levels of health care workers in Cape Coast Teaching Hospital

➤ Think through the two research questions (or scenarios) above to deduce the

- Dependent variable
- Independent variable
- Nature of dependent variable
- Levels of independent variable



■ Other scenarios

- We have a group of individuals randomly split into smaller and completing tasks. That is, we want to study the effects of tea on weight loss and we decided to form groups where individuals drink green tea, black tea, and no tea
- In a similar situation, we are interested in studying individuals based on leg strength according to their weight.

❖ Limitations of One-Way ANOVA

- One-way ANOVA will tell us that at least two groups were different from each other, but it will not tell which groups were different
- Hence, we will need to run a post hoc test
- The ad hoc test will tell the difference is

ANOVA



- i.e., for ANOVA F statistic is used to determine if the groups have significantly different means
- If the probability associated with the F statistics is **0.05** or less then we can assert that there is a difference in the means
- If there is a difference, then there is the need to perform a **post hoc analysis** to show where the difference(s) is/are among the groups

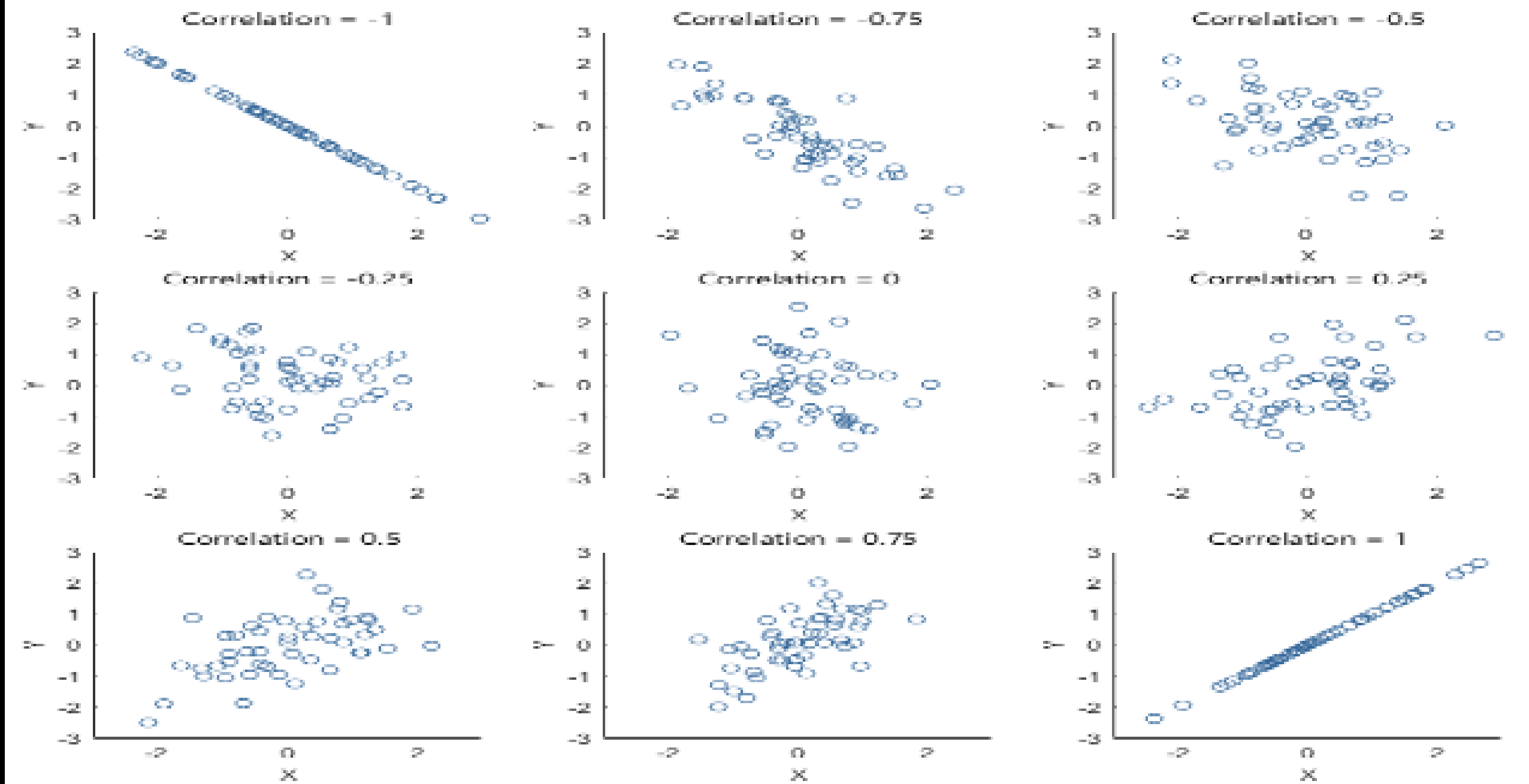
Linear Correlation and Prediction



➤ Linear correlation is a statistical tool employed to measure the dependence between two random variables

❖ Characteristics

- it ranges between -1 and 1
 - it is proportional to covariance
 - its interpretation is very similar to that of covariance
- *Think through the following;*





- Let us have our usual x and y to be the two random variables
- The linear correlation coefficient between x and y is
- $$\text{Corr}[x, y] = \frac{\text{Cov}[x, y]}{\text{Std}[x]\text{Std}[y]}$$
- Where $\text{Cov}[x, y]$ is the covariance between x and y
- $\text{Std.}[x]$ and $\text{Std.}[y]$ are the standard deviations of x and y respectively
- The linear correlation coefficient is well-defined only as long as $\text{Cov}[x, y]$, $\text{Std.}[x]$ and $\text{Std.}[y]$ exist and are well defined



- The interpretation is similar to the interpretation of covariance
- The correlation between x and y provides a measure of how similar their deviations from the respective means are
- Linear correlation ranges between -1 and 1 as
$$-1 \leq \text{Corr}[x, y] \leq 1$$
- If $\text{Corr}[x, y] > 0$, then x and y are said to be positively linearly correlated (or positively correlated)



- If $\text{Corr}[x, y] < 0$, then x and y are said to be negatively linearly correlated (negatively correlated)
- If $\text{Corr}[x, y] \neq 0$, then x and y are said to be linearly correlated (or simply correlated)
- If $\text{Corr}[x, y] = 0$, then x and y are said to be uncorrelated



- The **correlation coefficient**, r measures the direction and strength of a linear relationship
- Calculating r is pretty complex as seen earlier
- so we usually rely on technology for the computations
- We focus on understanding what r says about a scatterplot

Linear Correlation and Prediction



➤ The **Pearson correlation coefficient** (r) is the most common way of measuring a linear correlation

- The Pearson correlation coefficient a value between -1 and 1 that measures the strength and direction of the relationship between two variables
- The Pearson correlation coefficient has many names as
 - Pearson's r
 - Bivariate correlation
 - Pearson product-moment correlation coefficient
 - The correlation coefficient

Linear Correlation and Prediction



| Pearson's r | Strength | Direction |
|---------------------------|----------|-----------|
| > 0.5 | Strong | Positive |
| Between 0.3 and 0.5 | Moderate | Positive |
| Between 0 and 0.3 | Weak | Positive |
| 0 | None | None |
| Between 0 and -0.3 | Weak | Negative |
| Between -0.3 and -0.5 | Moderate | Negative |
| > -0.5 | Strong | negative |

Linear Correlation and Prediction



- The Pearson correlation coefficient is a **descriptive statistic**
 - i.e., the Pearson correlation coefficient summarises the characteristics of a dataset
 - e.g., up till a certain age (in most cases), a child's height will keep increasing as his/her age increase
 - specifically, it describes the strength and direction of the linear relationship between two quantitative variables
 - Although interpretations of the relationship strength (i.e., **effect size**) vary between disciplines
- **Make references to the figures observed earlier**



➤ The Pearson correlation coefficient is a good choice for analysing data when **all** of the following are true;

- 1. Both variables are quantitative
- 2. The variables are normally distributed
- 3. The data have no outliers, i.e., observations that do not follow the same patterns as the rest of the data
- A scatterplot is one way to check for outliers
- 4. The relationship is linear
- i.e., we can once again use a scatterplot to check whether the relationship between two variables is linear



➤ The Pearson correlation coefficient is also an **inferential statistic**

- i.e., it can be used to *test statistical hypotheses*
- Specifically, we can test whether there is a significant relationship between two variables



➤ **Spearman's rho**, or **Spearman's rank correlation coefficient**, is the most common alternative to Pearson's r

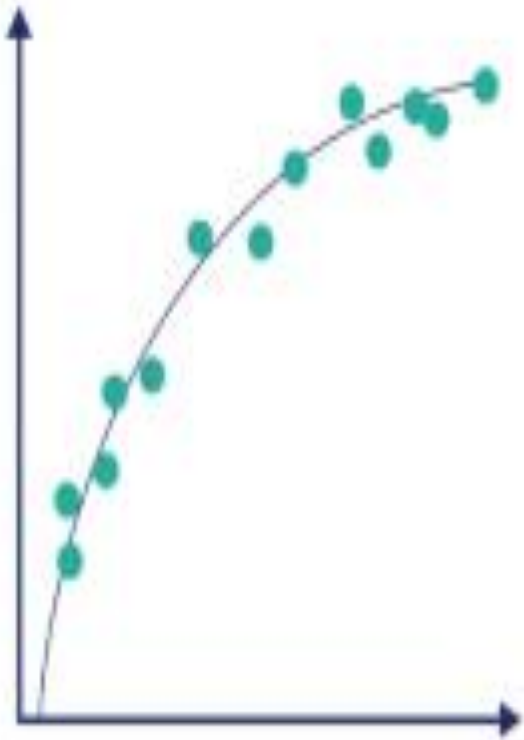
- **Spearman's rho** is a rank correlation coefficient because it uses the rankings of data from each variable
 - e.g., from lowest to highest, rather than the raw data itself
- We will use Spearman's rho when your data fail to meet the assumptions of Pearson's r
- This happens when at least one of your variables is on an ordinal level of measurement
- or when the data from one or both variables do not follow normal distributions

Linear Correlation and Prediction

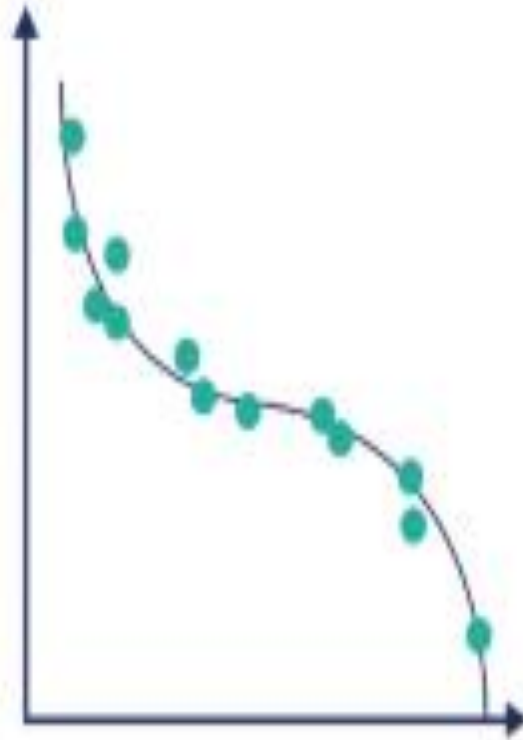


- While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships
- i.e., in a linear relationship, each variable changes in one direction at the same rate throughout the data range
- In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate
- Positive monotonic: when one variable increases, the other also increases
- Negative monotonic: when one variable increases, the other decreases

**Positive monotonic
relationship**



**Negative monotonic
relationship**



**Non-monotonic
relationship**





❖ Spearman's rank correlation coefficient formula

➤ The symbols for Spearman's rho are ρ for the population coefficient

■ r_s for the sample coefficient

■ The formula calculates the Pearson's r correlation coefficient between the rankings of the variable data

➤
$$r_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

r_s = strength of the rank correlation between variables

d_i = the difference between the x-variable rank and the y-variable rank for each pair of data

Linear Correlation and Prediction



$\sum d_i^2$ = sum of the squared differences between x- and y-variable ranks

n = sample size

- In using this formula, we will first rank the data from each variable separately from low to high
- every data point gets a rank from first, second, or third, etc.

Linear Correlation and Prediction



- If we have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair
- If we have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable
- If we have a correlation coefficient near zero, then there is no monotonic relationship between the variable rankings

Chi Square Test



- A chi-square (χ^2) statistic is a test that measures how a model compares to actual observed data
- The data used in determining a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample
- For example, the results of individuals who had a job in the peak period of Covid-19 in Ghana

Chi Square Test



- Chi-square statistics is often used to test hypotheses
- The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, with respect to the size of the sample and the number of variables in the relationship
- For a chi-square statistic, **degrees of freedom** are used to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment
- As with any statistic, the larger the sample size, the more reliable the results



❖ When to Use Chi-Square Statistics

- A chi-square statistic is a measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables
- Chi-square is useful for analysing differences in categorical variables, especially those nominal in nature
- Chi-square depends on the size of the difference between actual and observed values, the degrees of freedom, and the sample size
- Chi-square can be used to test whether two variables are related or independent from each other
- Chi-square can also be used to test the goodness of fit between an observed distribution and a theoretical distribution of frequencies



❖ The formula for chi-square is

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Where C is the degrees of freedom
- O is the observed value(s)
- E is the expected value(s)



❖ Forms of Chi-Square

- 1. the **test of independence**, which asks a question of relationship
 - e.g., Is there a relationship between student gender and selection of midwifery as a programme in Nursing and Midwifery Colleges?
- 2. **Goodness-of-fit test**, which asks something like
 - How well is my DNA matching that of my father's children?



❖ Test of Independence

- When considering student gender and programme choice, a *chi-square* statistic for independence could be used
- i.e., we can collect data on the two chosen variables (gender and programme choice) and then compare the frequencies at which male and female students select among the offered classes using the formula and a χ^2 statistical table

Chi Square Test



- If there is no relationship between gender and programme choice (that is if they are independent), then the actual frequencies at which male and female students select each offered programme should be expected to be approximately equal
- or conversely, the proportion of male and female students in any selected programme should be approximately equal to the proportion of male and female students in the sample
- A *chi-square* statistic for independence can tell us how likely it is that random chance can explain any observed difference between the actual frequencies in the data and these theoretical expectations



❖ Goodness-of-Fit Test

- Goodness-of-fit test is a chi-square statistic provides a way to test how well a sample of data matches the (known or assumed) characteristics of the larger population that the sample is intended to represent
- If the sample data do not fit the expected properties of the population that we are interested in, then we would not want to use this sample to draw conclusions about the larger population



❖ How to Perform a Chi-Square Test

- Create a table of the observed and expected frequencies
- Use the formula to calculate the chi-square value
- Find the critical chi-square value using a chi-square value table or statistical software
- Determine whether the chi-square value or the critical value is the larger of the two
- Reject or accept the null hypothesis



❖ Limitations of the Chi-Square Statistic

- The chi-square statistic is sensitive to sample size
- Relationships may appear to be significant when they are not simply because a very large sample is used
- The chi-square statistic cannot establish whether one variable has a causal relationship with another
- i.e., chi-square can only establish whether two variables are related

Chi Square Test



- Pearson's chi-square (X^2) tests, often referred to simply as chi-square tests, are among the most common **nonparametric tests**
- If we want to test a hypothesis about the distribution of a **categorical variable** we will need to use a chi-square test or another nonparametric test
- The categorical variables can be nominal or ordinal and represent groupings such as species or nationalities
- Because the categorical variables can only have a few specific values
- The categorical variables cannot be a normal distribution

Introduction to Linear Regression

