2a. BIOSTATISTICS IN CLINICAL CHEMISTRY; THE SIGNIFICANCE OF THE NORMAL.

Statistics: science of **data**

Biostatistics: data from: medicine biological sciences

Statistics: is a field of study concerned with

1. Collection, organization, summarization and analysis of data
2. Drawing of inferences about a body of data when only a part of the data is

When data are derived from the biological science and medicine we use the term biostatistics.

Sources of data:

1. **Routinely kept records**
   E.g. hospital records contain immense amounts of information on patients
2. **External sources**
   Data needed to answer a question may already exist in the form of published reports, commercially available data banks or the research literature i.e. someone else has already asked the same question.
3. **Surveys**
   The source may be a survey, if the data needed is about answering certain questions e.g. if the administrator of a clinic wishes to obtain
4. **Experiments**
   Frequently the data needed to answer a question are available only as the result of an experiment

A VARIABLE

It is a characteristic that takes on different values in different persons place or things

1) Quantitative variable it can be measured in the usual sense
2) Qualitative variable many xtics are not capable of being measured. Some of them can be ordered or ranked
3) Discrete variable xtized by gaps or interruptions in the values that it can assume
4) Continuous variable can assume any value within a specified relevant interval of values assumed by the variable. E.g. height, weight

A POPULATION

It's the largest collection of values of a random variable for which we have an interest at a particular time.

A SAMPLE

It's a part of a population.

Descriptive statistics

- Grouped data the frequency distribution
- Measures of central tendency- it's a measure which indicates where the middle of the data is (mean, mode, median)
- Measures of dispersion (dispersion, Variation, spread, scatter)

Arithmetic Mean: Center of a particular set of numbers obtained by adding the quantities dividing by the total number of observations (n).

$$\overline{X} = \frac{\sum Xi}{n}$$

This is the representation of the population if the histogram is symmetrical. The assumption is that data obtained in the field will give you a symmetrical histogram. If the entire population is used to calculate the arithmetic mean say of blood pressure, what will be calculated here is the true mean. It is not represented by the X but by the sign μ

If the population is not symmetrical, then the mean will not give the measure of central tendency; the **mode** may become the central tendency. Another term used is **median** which will give the central tendency

Properties of mean

- Uniqueness only one in a set of data
- Simplicity
- Affected by extreme values

Mode: This is the most common value when somebody is looking at the mean of a population and recurs many times.

Median: The value at which half the values above and below are equal in number. The digits are arranged in either ascending or descending order.

Properties of median

- Uniqueness only one
- Simplicity- it's easy to calculate
- It's not affected by extreme values

An example of asymmetrical distribution is called **bimodal distribution** meaning two distributions in one. Each distribution will have its own mean and its own distribution. This is an example of non-symmetrical distribution but is very uncommon.

From the mean position, the value can used to know by how much degree the values vary from the mean.

Data Variation (measures of dispersion)

Indicators of the variations of measurements or the spread of distributions are usually given by the following;

- Range
- Variance
- Standard deviation
- Coefficient of variation

Range: This is the difference between the lowest and highest values in a data set.

- This is useful in indicating data spread where the number of population is small.
- Range also makes no assumptions about the shape of the distribution; this helps in determining whether the data is symmetrical or asymmetrical

Variance: Obtained by adding the squares or the differences between the individual values and the mean divided by the number of values minus 1 for big values

$$v = \frac{\sum(\bar{x} - x)^2}{N - 1}$$

Standard Deviation: This is the square root of variance

$$s = \sqrt{\frac{\sum fx^2}{N - 1}}$$

If the standard deviation is calculated from the entire population, then the standard

deviation is not represented by S but by (σ-sigma). The symbol S is used only when calculating a sample population

Coefficient of Variation

This is the standard deviation divided by the mean then multiplied by 100 to give a percentage value. It is simply 100% i.e. the quotient of standard deviation divided by the mean.

$$C_oV = \frac{S}{\bar{X}} \times 100$$

When multiplied by 100, it becomes the relative dispersion

This is usually given as a percentage and usually does not have units. It used to express random variation of analytical methods in units independent of analytical methodology. It is assumed that coefficient of variation of analytical method is independent of concentration.

Confidence Intervals

Are normally distributed population whether height, weight, blood pressure but the population will be described by two statistics;

- Arithmetic mean
- Standard deviation

The two will completely describe data sets. Confidence of believe in the set increases as the population and subjects increase and vice versa. The more the observable sets, the more the level of confidence.

Parametric Statistics: Goshen statistics are utilized under parametric. Only 50% of biological data will give you symmetrical population. The rest of the population will be skewed (leaning towards one side) or kurtotic and is positive if it leans towards the right and is negative when a lot of data is at the left side at the base

Normal Distribution

Symmetrical data show the mean, mode and median lying at the center. If 1 standard deviation is moved from the mean on both sides, 68% of the data will lie in the region. If standard deviations of 2 are deducted on each side, 95.5% of the data will lie in the region. Further movement will give 99.9% of the data.

These intervals having data are called **confidence intervals**. These confidence intervals are the basis of statistical quality control rules for acceptance and rejection decisions concerning analytical data

Mean

- Is more accurate than a single measure if two or more values are taken
- If different means are obtained from different sources, the individual means are distributed along a grand mean i.e. getting a mean from means
- The random variation in the population of mean is described not by standard deviation but by the standard error of the mean which is the standard deviation divided by the square root of observation number

$$s\tan\ darderror = \frac{S}{\sqrt{N-1}}$$

2b. BIOCHEMICAL REFERENCE RANGES

Reference Population: This is the population under which studies are based on.

Exclusion criterion is used on certain risk factors. Risk factors include diseases like obesity, hypertension, occupational or environmental risks and genetically determined risks.

Others in the exclusion criterion include drug abuse, tobacco smoking and miraa chewing. Others include excessive exercise and situations like pregnancy, under such, laboratory tests are excluded because the body's metabolism is different from the normal hence falls under exclusion criteria.

Types of References Ranges

1. Health Based References

These are references that are considered to be healthy e.g. medical practitioners are considered healthy and the reference values are based on them.

2. Hospital Based References

These are references that are considered off the ranges under a normal distribution and fall outside the 95% confidence interval and are rejected values. Hence the level of acceptance helps in forming reference ranges such that those under either hyper of hypo situations are considered to be under the hospital based reference

3. Pathology Based References

If you have say diabetes, the clinics and health care centers provide the information of the types of diabetes and the numbers of affected persons under each type of diabetes

and reference ranges are drawn

4. Individual Based References

This is done for an individual and sets of measurements are taken, a mean obtained and these forms basis of a reference set of values for the individual, knowing the upper and lower limits and say this person's blood pressure for instance varies from this value to this value.

In order to get any reference range, a sample size is obtained in order to get the data. The most common formula is;

$$\frac{Z^2 \ p(1 - p)}{d^2} = N$$

P = Preference in the population (Ranges from 0.05 – 0.5)

N = Number of observations

d = Precision of determining preference

Preference ranges can be obtained via urine tests or the type of survey one is doing. Others may include amniotic fluid, CSF etc but the most common is blood samples

Example of Glucose samples, one must know the outliers, they lay way out of the expected values. The best way to know the difference between outliers and non outliers is by doing a cumulative frequency distribution. If the curve is symmetrical then parametric statistics is done and the outliers are obtained (both ends of the tail)

In skewed distribution, parametric statistics cannot be used. Parametric statistics is done in chi squares and reaching at the index levels based on the symmetrical nature of the data.

If the distribution is not symmetrical, a Kolomogotor-Sminorv (KS index) is used, if the index is low, the data is skewed or Gaussian. If the data is not symmetrical, one can also use the square root or the logarithm. The square root is more powerful than the logarithm.

There are two commonly used non-parametric statistics;

- Kruskal-Wallis test
- Wilicoxons ranked sum test

- ➢ Mann Whitney U test (is a substitute for Wilicoxons)

These are simple to use because of computer software but can be destructive due to loss of data. A lot of observations are required for the data to be analyzed

Rule of the Thumb

- If the same data was obtained in another survey; regardless of where the survey was done before, one is allowed to use up to 40 observations or at times up to 120 observations if the data had no prior reference ranges to make new reference ranges