Data and information:

Data consists of discrete observation of variable carry little or no meaning when considered alone.

Information : variables that have meanings

Statistics : is the science of collecting ,summarizing ,presenting and interpreting data

Types of data

1-constant :-person have 2 eyes

2-variables :-

Types of Variables

In order to collect data about an event, one should identify the items to be collected. Every item is called a variable i.c. age is a variable, sex is another variable, weight is a variable etc.....

Variables are of 2 type

A) **Qualitative variables;** which are items that can be described. they are further divided into 2 subtypes:

.**Nominal variables**: They are descriptive variables that identify subjects regarding acharacteristic e.g. gender (male, female) - blood groups – pregnancy(+ve, ve)etc.

.**Ordinal variables**: these are descriptive variables that rank subjects according to a characteristic e.g. mild, moderate, severe malnutrition - patients with different grades of malignancy Grade 0, I, II, III..etc

B) Quantitative variables: these are items that have Quantitatities. they are further divided into 2 subtype

.**Continuous variable**: these are variables that can be measured in fractions .for example :weight 8.5 kg .

.**Discrete variable**: these variables take only integer values i.e. they don't have fractions e.g. no of fingers ,heart bets ,family members

Mathematical presentation of data

Measures of central tendency

These are measures for central location i.e. they show how much the observations are aggregated around a central point.
They include:
1) Arithmetic mean
2) Mid-range
3) Median
4) Mode.

1) Arithmetic mean

It is equal to the sum of the values (observations) in the group divided by the number of values (observations) comprising that group.

$$\bar{x}=\frac{\Sigma x}{n}$$

$\bar{X}$=mean
X=value (observation)
N=number of values (no, of observations)
Σ=sum
Example

X1=11.85 X2=11.75

X3-11.80 X4-13.15

X5=12.45

Thus

$$\bar{x}=\frac{\Sigma x}{n}=\frac{x1+x2+x3+x4+x5}{5}=\frac{61}{5}=\mathbf{12.2}$$

Properties of the arithmetic mean:

1. Every item is included in the computation.

2. An extreme value can exert an influence on the arithmetic mean.

3. The arithmetic mean may take a value that is not present among the values of the variable from which it was computed.

4. If the value of arithmetic mean is multiplied by the number of items involved in the computation, the product equals the sum of Xi values.

5.The sum of the deviations of the values of individual items from the arithmetic mean is equal to zero.

2) Median

It is defined as the value of the middle item in a distribution when the items have been arranged in an ascending or descending manner according to their magnitude.

When the no. of observations is odd e.g.

2, 5, 12, 17, 19, 25, 40

The position occupied by the median item = $\frac{n+1}{2}=\frac{7+1}{2} = \frac{8}{2} = 4th \backslash position$

So, the rank of the median is the 4 position and its value is 17.

When the no. of observations is even e.g. 3, 7, 15, 30, 40, 45, 50, 51
The position of the median ;

$\frac{n}{2}$, $\frac{n}{2}+1$ $\quad i.e \quad$ $\frac{8}{2}$, $\frac{8}{2}+1$

So, the rank of the median is 4^{th} & 5^{th} ., and the value of the median equals $\frac{30+40}{2} = 35$

Properties of the median:

1. Not affected by extreme values.

2. The value obtained for the median may be non- representative if the individual items have great variability (don't cluster at the centre of the distribution).

3) Mode

It is defined as the most commonly occurring value ie, the value of the variable that occurs with the greatest frequency. The following data represent the ages of a group of children; Sometimes, the distribution may be bimodal or have more than 2 modes.

2, 3, 5, 8, 10, 5, 9, 5, 12

the mode of the age is 5

Measures of dispersion

Measures of dispersion provide information with respect to the extent of scattering in a set of data. They are useful in evaluating the representativeness of a measure of central tendency ie more information is gained when a measure of scattering is put beside a measure of central tendency for a set of data.

It is important to know that the more the measure of dispersion, the more is the scattering among the group of observations. And as the measure of dispersion for the group decreases, the observations among the group become more homogenous

Measures of Dispersion include:

1. Range

2. Standard deviation

3. Variance

4. SE. i.e, standard error of the mean.

1- Range

It refers to the difference between the largest and smallest observations in the data set

Ex.: The Weight of 7 children is as follows: 18-10-15-30-25-17-21

The Range= 30-10=20

The major disadvantage associated with the range is that it is based on the dispersion of the two extreme values i,e. affected only by extreme values.

2- Standard Deviation (S.D.)

It is the most frequently used measure of scattering as it is used in most equations and all observations of the group enter in its calculation. The computation process consists of determining the deviations of each of the individual items from the mean, squaring the deviations (negative signs become positive), then the squared deviations are summed and divided by the number of observations to get Variance, which is the average squared deviation from the mean

S.D=√ variance

Then, taking the square root of the variance to get Standard deviation.

$$S.D=\sqrt{\frac{\Sigma(xi-\bar{x})^2}{n}}$$

3-Variance

It is the square of the S.D. It is calculated exactly as the S.D. but without square rooting the summation of squared deviations divided by the no. of observations.

Normal Distribution Curve

Many biological characteristics e.g. H.R., B.P., Hb., Cholesterol etc...when plotted follow the normal distribution curve i.e. have a curve with certain characteristics common for all normal distributions.

Properties of Normal distribution curve:

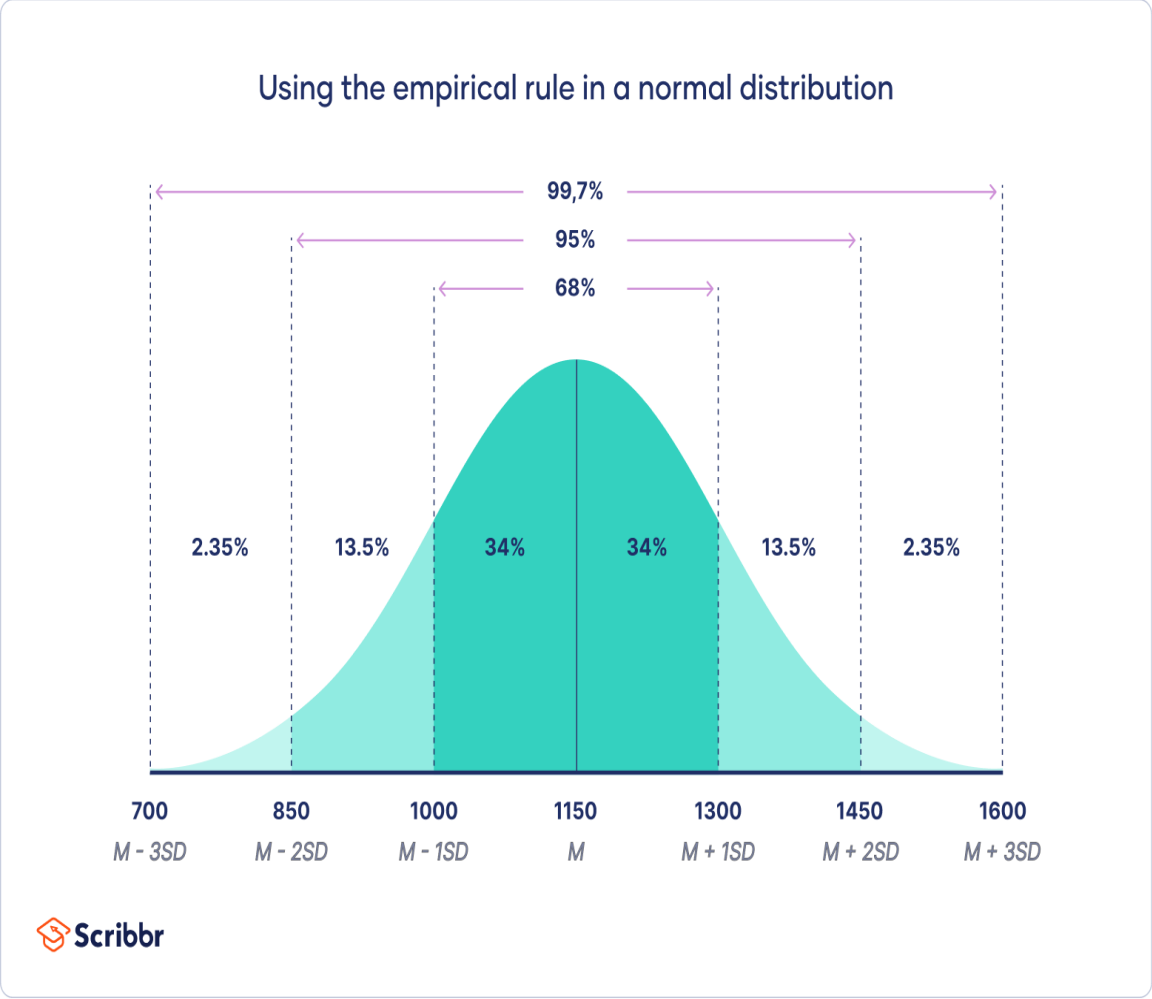
1. It is bell-shaped, bilateral and symmetrical, with the peak at the mean which is located at the mid point of the base.

2. The mean, median and mode coincide together.

3. The curve shades off to small values above and below the mean.

4. The curve has a point of inflexion on both sides of the centre where the curvature changes from convex upwards to concave upwards. The point of inflexion is located at one standard deviation above the center and one standard deviation below the center. Between these 2 points, we find 68.2% of the area under the curve and 68.2% of the total frequency of the population.

5. The percentage of the area included within other multiples of the S.D. above and below the mean is always fixed, as shown in figure (14), Between 2 S.D. (above or below the mean) 95.4% of the area is included i.e. 95% of the population are located in this area. Between 3 S.D. -99.7% of the area is included i.e. 99% of the population are located in this area.
For simplicity, the above figures are rounded to be; -
Between 1 S.D. Below and above the mean, 68% of population are located.
Between 2 S.D. Below and above the mean, 95% of the populations are located.
2.5% of the population in the upper tail i.e. above 2 S.D.
2.5% of the population in the lower tail i.e. below 2 S.D.



Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness:

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

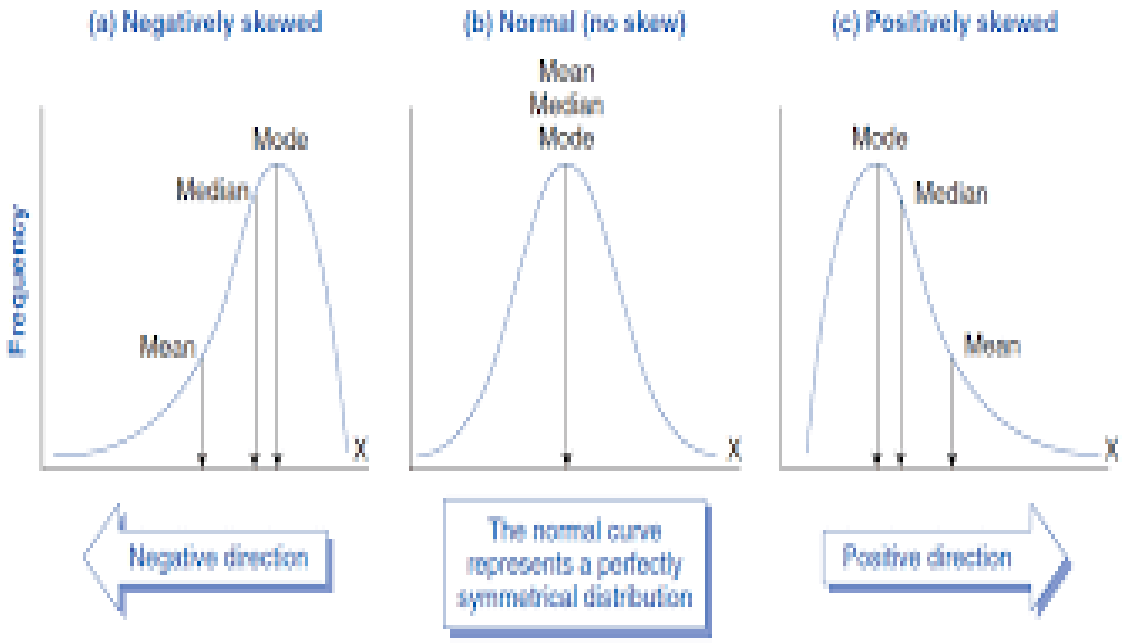
The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. In other words;

1- If a distribution is perfectly symmetrical, the measure of skewness is equal to

zero.

2. If the distribution is asymmetrical and the tail of the distribution extends in the direction of the positive values positive skewness (shift to the right) where the mean will be to the right (greater than) median and mode.

3. If the tail extends in the direction of negative values→→→ negative skewness (shift to the left) where the mean will be to the left (less than) median and mode.



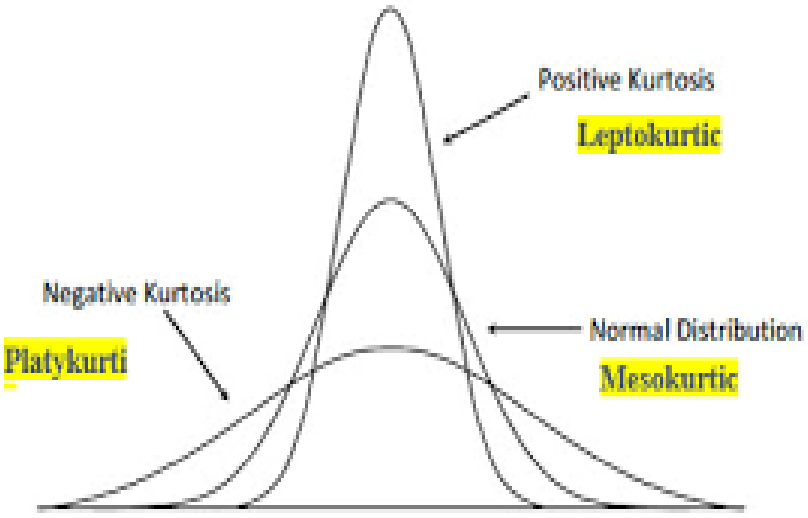
Kurtosis:

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean. A uniform distribution would be the extreme case.

Distributions with zero kurtosis are called mesokurtic, or mesokurtotic. The most prominent example of a mesokurtic distribution is the normal distribution. A distribution with positive excess kurtosis is called leptokurtic, or leptokurtotic. "Lepto-" means "slender". In terms of shape, a leptokurtic distribution has a more acute peak around the mean. A distribution with negative kurtosis is called platykurtic, or platykurtotic.

Platy" means "broad". In terms of shape, a platykurtic distribution has a lower, wider peak around the mean

The standard normal distribution has a kurtosis of zero. Positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution



Study design

Study design : is the protocol for selecting persons to study and the method in which data are collected

Types of studies

Observational studies:

1- descriptive studies

2- analytical studies:

a) prospective cohort

b) cross sectional

c) case control

Experimental studies :

-randomized clinical trials

-field trials

Descriptive studies

Usually undertaken when little is known of the epidemiology of the disease

Does not involve hypothesis testing

Concerned with observing the distribution of the disease

1- time distribution

2- place distribution

3- person distribution

Analytical studies

1.prospective cohort study

The investigator starts with a group of individuals apparently free of the disease

This group is divided into those exposed to a possible risk factor and those not exposed

Then is followed through time in order to determine the incidence rate among the exposal and the unexposed

number of new cases in a group in a specified period of time

1- incidence (risk)=population at risk during that period 1000

$$\text{Relative risk} = \frac{\text{incidence in exposed population}}{\text{incidence in non exposed population}}$$

RR of 1 indicates the incidence in exposed is equal to the incidence in non exposed

RR > 1 indicates that the exposed individuals are at a greater risk than non exposed

RR<1 indicates that the exposed individuals are at a lower risk than non exposed

Attributable risk: it is the proportion of disease incidence which can be attributed to specific exposure

AR=(incidence in exposed group) –(incidence in non exposed group)

The attributable risk percent =

(incidence in exposed group) –(incidence in non exposed group)

Incidence in exposed group 100

Advantages of cohort studies

1- they allow complete description of the individuals experience subsequent to exposure

2- they provide a clear temporal sequence of exposure and disease

3- they provide excellent opportunity to study rare exposures

4- they permit the assessment of multiple outcomes

5- they permit the direct estimation of the rate of health problem and the RR associated with the exposure of interest

6- less chance for bias

7- they provide more understandable information to non epidemiologists

Disadvantages of cohort studies

1- not suitable for rare diseases where large numbers of subjects are required

2- long term follow up may be necessary when the latency period for the outcome of interest is long

3- the most serious problem is attrition or loss of people from the sample during the course of the study

4- they are very time consuming and expensive

5- the exposure status may change during the conduct of the study

6- there may be attrition among the investigators

Case control studies (retrospective studies)

They are efficient and common epidemiological studies

They depend on exposure history among cases and controls

They investigate the association of a disease condition with a risk factor by contrasting the exposure of a series of cases with the exposure of selected controls

1-Selection of cases

In regard to :

Histologic type

stage of disease

Date of diagnosis
Geographic location
2-Selection of controls
To obtain estimates of the frequency of attribute or risk factor for comparison with its frequecy among cases
The comparison group may be:
1- a probability sample of a defined population
2- a sample of patients admitted to the same institution as cases
3- a sample of relatives of the cases
matching
it is the process by which we select controls in such a way that they are similar to cases with regards to certain selected pertinent variables (eg. Age) which are known to influence the outcome of disease
The odds ratio is a measure of relationship between exposure and disease
If OR=1 the exposure is not related to disease
If OR>1 the exposure is positively related to disease
If OR<1 the exposure is negatively related to disease
the odds ratio is a good estimate of the relative risk in case con

Advantages of case control studies

1- it is the most frequent undertaken type of epidemiological studies
2- they are useful for studying health problems that occur infrequently
3- they are useful for studying health problems with a long latent interval
4-less time consuming and less expensive than cohort studies
5- they are useful for studying the effects of multiple risk factors on the health problems under study
6- it requires a smaller sample than other studies
7-there is no problem of attrition
8- this is considered the earliest study provide leads to be followed up by more definitive cohort studies

Disadvantages of case control studies:

1- selection bias : because case and controls may be selected from two separete populations , it is difficult to be comparable
2-Recall bias: exposure data are collected from records or by recall after the disease has occurred . Records may be incomplete and recall of past events is subject to human error
3-temporality is a serious problem ,where it is not possible to determine whether

risk led to the disease or vice versa

4- if the health problem is relatively common in the population (>5-10%) the odds ratio is not a reliable estimate of the relative risk

5- they can not be used to determine the other possible health effects of an exposure.they are concerned with only one outcome

3.Cross-Sectional studies (Prevalence Study):

- It is an analysis of collected data on a group of individuals at a point of time or over a period of time. These are conducted to know what is occurring now.

Applications:

-Diagnosis of diseases.

-Staging of diseases.

-Screening for diseases.

-Identification of risk factors

Limitations of This type of study is:

a. Very expensive.

b. Takes long time.

c. There is high rate of drop out.

d. Its results are more conclusive than retrospective studies.

Mid term exam

الاسم:.....

.

رقم الجلوس:.....

›

› 1. 13 14 14 15 16 16 16 17 17 18 20

›

› **Calculate mean ,median , mode and S.D**