STATISTICS IN NURSING RESEARCH

Dr. Hoda Ibrahim Ibrahim Rizk

Assistant professor of Public Health & Community Medicine

Dr. Marwa Rashad Salem

Lecturer of Public Health & Community Medicine

Faculty of Medicine, Cairo University

STATISTICS IN NURSING RESEARCH

1. Definition

Statistics is the study of how to collect, organizes, analyze, and Interpret data.

2. Importance

- Statistics plays an important role in the research:

• Helps to answer important research questions and field in study.
• Important to understand what tools are suitable for a particular research study

- Statistics plays an important role in the decision making process.

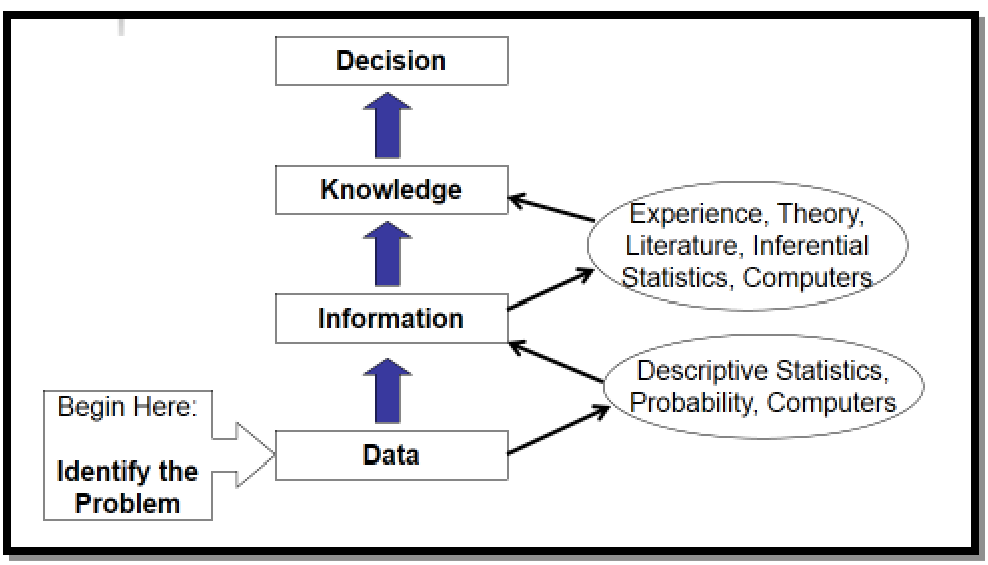


Fig.1The Decision Making Process

3. **Types of statistics**

There are two approaches to the statistical analysis of data

1. **Descriptive Statistics**: Descriptive statistics are techniques which help the investigator to organize, summarize and describe measures of a sample.

2. **Inferential statistics**: to make statements about a population by examining sample results.

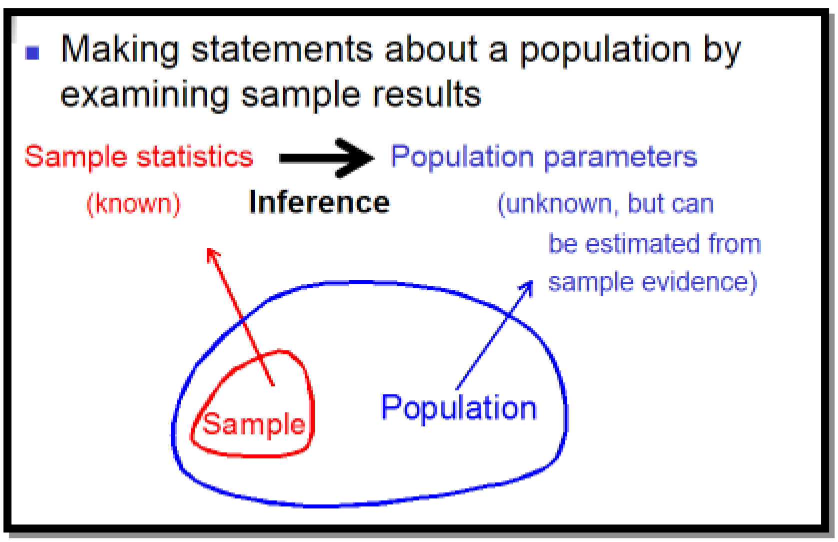


Fig. 2 Inferential Statistics

Variables and data

➢ Definition

A variable is something whose value can vary. For example, age, sex and blood type are variables

Data are the values you get when you measure a variable. For example, 32 years (for the variable age), or female (for the variable sex) as shown in Table (1).

Information is translation of the data to a meaningful knowledge

Example: Your blood pressure is 190/95 ……….You are Hypertensive

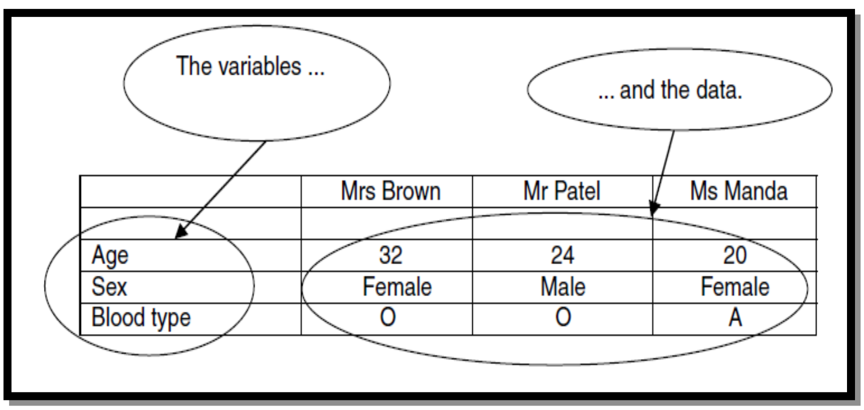


Fig. 3 Variables versus Data

➢ Types of variables

There are two major types of variable – *categorical (Qualitative) variables* and *metric* (*Quantitative)* variables. Each of these can be further divided into two sub-types, as shown in Figure 1

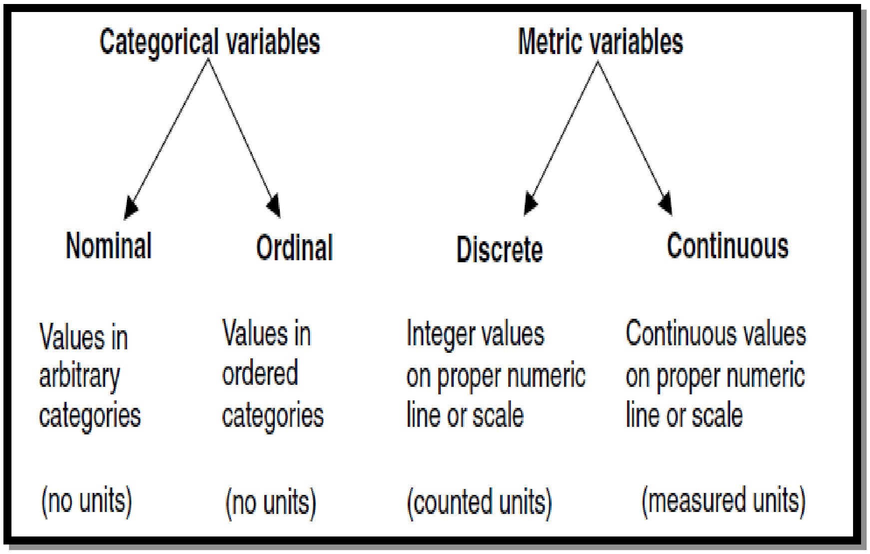


Fig.4.a Classification of variables

- **An algorithm to help to identify variable type**

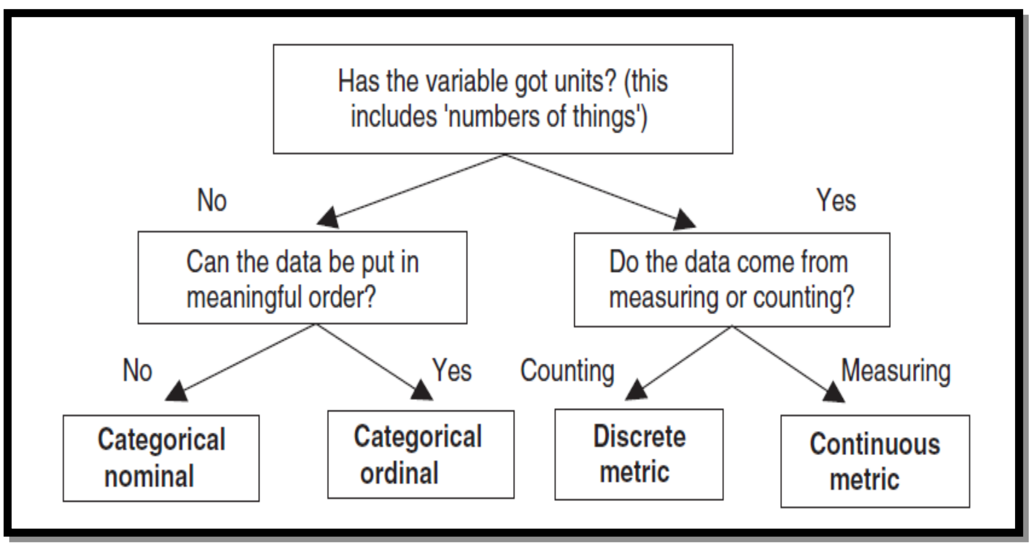


Fig.4.b Classification of variables

Categorical variables can either be ordinal or nominal.

Ordinal variables: These are grouped variables that are ordered or ranked in increasing or decreasing order: e.g. severity of a disease: severe, moderate, mild

Nominal variables: The groups in these variables do not have an order or ranking in them e.g. Sex: Male and Female

Metric (quantitative) variables: can either be continuous or discrete

Discrete variables*:* the measurements of the different values of the variable are separated. Examples: number of live births per mother, number of cigarettes per day.

Continuous variables*:* the measurements of the different values take on any numerical value within the variable's range. Examples: age, blood pressure, blood sugar.

Example

Number	**SEX**	**AGE**	**Academic Performance**	**Pulse**
1	Male	23	Good	70
2	Female	26	V. Good	68
3	Male	31	Good	78
4	Male	28	Good	72
5	Female	27	Excellent	66

- Sex is a qualitative nominal variable
- Age is a metric (quantitative) continuous variable
- Academic Performance is a qualitative ordinal variable
- Pulse is a metric (quantitative) discrete variable

Presentation of Data

After data collection, computer data entry, and analysis, the data have to be presented in an easy to understand way. The purposes of data presentation are:

- Find out the common finding,
- Find out group variations

<u>Proper data presentation</u>

Tables: When details of data are needed.

Graphs: When only impressions are needed.

Parameters: Precise mathematical summary, useful for comparison.

I. <u>Tables</u>

1. Simple tables showing single variable

a. Tables with data on qualitative variables (nominal) (e.g. percent distribution of the enrolled nursing students by sex) as shown in Table 1

b. Table with data on quantitative variable (continuous) (e.g. percent distribution of the enrolled nursing students by age)

Table 1: Percent distribution of the enrolled nursing students by sex

Sex	Number	Percent
Males Females	11 7	61.1 38.9
Total	18	100.0

2. Contingency tables or cross tabulation of two variables.

In such tables two variables are used. In the following example two variables are presented: education and receiving antenatal care (ANC)

Table.2 Percent distribution of mothers according to education and receiving ANC for live births born during the last 5 years

Education	Received ANC		Did not receive ANC		Total	
	No.	%	No.	%	No.	%
Non-educated	64	40.0	96	60.0	160	100.0
Educated	216	90.0	24	10.0	240	100.0
Total →	280	70.0	120	30.0	400	100.0

The table should fulfill the following characteristics

1- No. of the table

2- Title of the table describing its contents

3- Suitable number of rows (4-12).

4- Title for each column and each row

5- Totals

6- Meaningful percent from row or column

7- Make sure that the tables and text refer to each other (through the table number)

8- Not everything displayed in the table needs to be mentioned in the text

II. Graphical representation

1. **Pie chart**

- A pie chart is a circular chart (pie-shaped); it is split into segments to show percentages or the relative contributions of categories of data.

- A pie chart gives an immediate visual idea of the relative sizes of the shares of a whole.

The following is an example of a pie chart of the Family planning methods use in the studied community

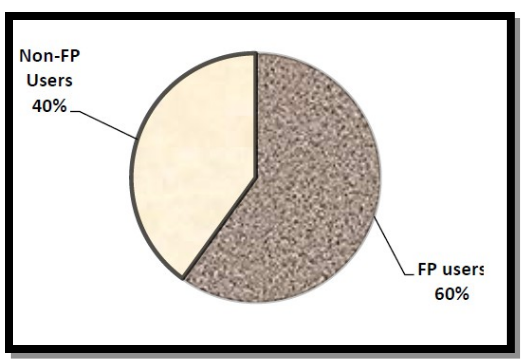


Fig.5 Contraceptive prevalence in the studied community

2. Bar Charts

A bar chart consists of parallel, usually vertical bars with their lengths corresponding to the frequency or percentage of each value. The bars are separated from each other by a space as to reflect on the categorical aspect of the variable.

Example:

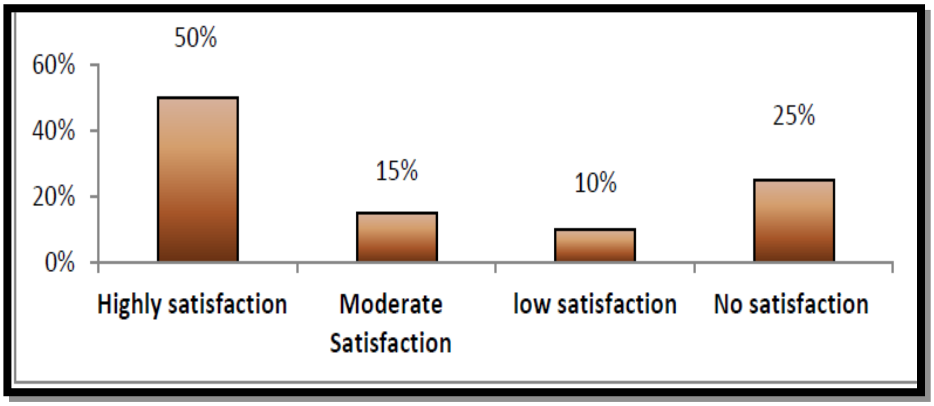


Fig.6 Percent distribution of the interviewed PHC centers clients according to the level of satisfaction from the service

N.B Bar charts and Pie charts are often used for qualitative (category) data

2. Map presentation of Data:

Map presentation of data can be applied at different levels, i.e. within different geographical zones or governorate of one country, within region between countries in same region or global presentation between different worldwide countries.

Example

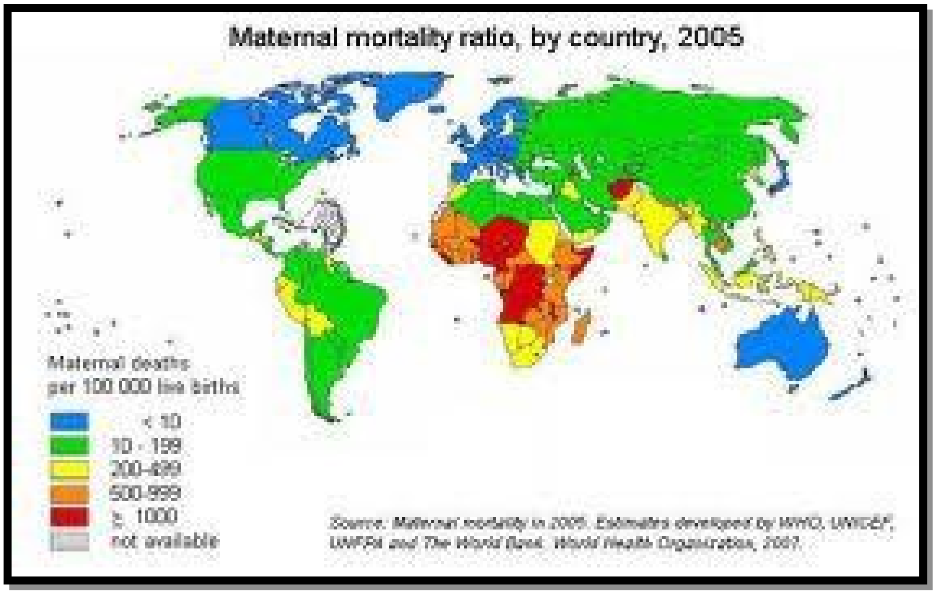


Fig.7 Maternal mortality ratio, by the country, 2005

4. Line charts

- Show values of one variable vs. time
- Time is traditionally shown on the horizontal axis

Example

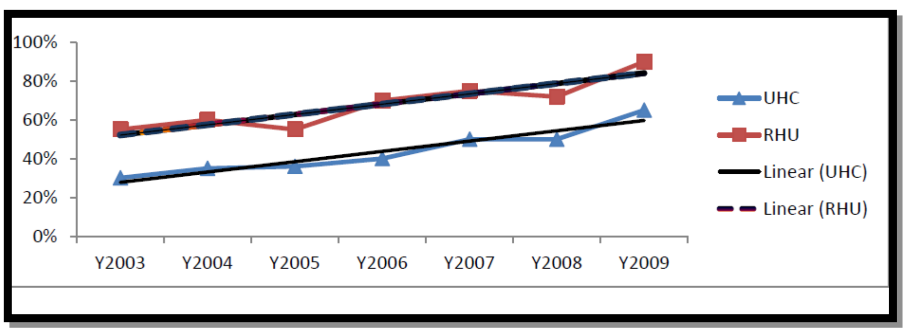
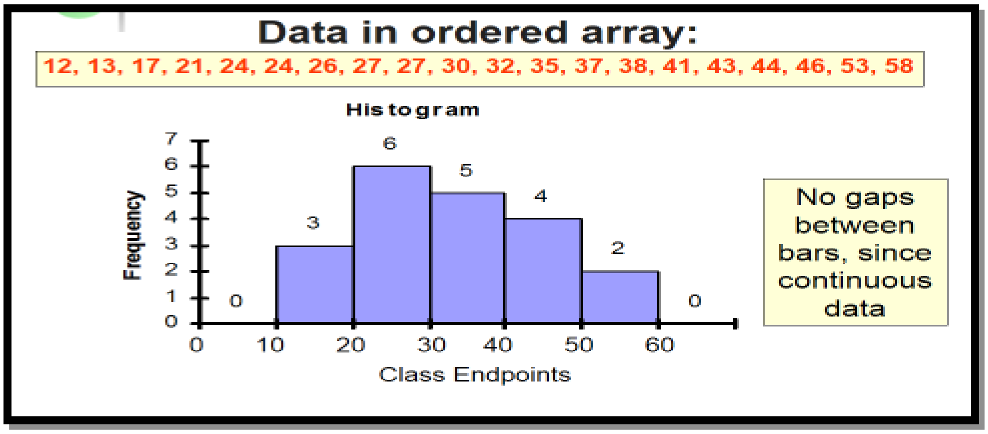


Fig. 8 Trend in antenatal care coverage in Giza for urban health care centers (UHC) and Rural Health care centers (RHC) (2003-2009)

5. Histogram

- It is appropriate for continuous variables (interval).
- It is similar to bar chart, but in histogram the bars are placed side by side.
- The bar length represents the percent (frequency) falling within each interval.
- Each histogram has a total area of 100%.

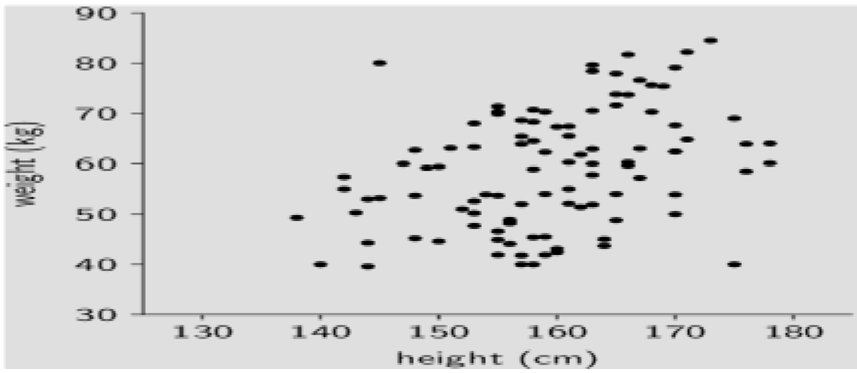
Example:



6. Scatter Diagrams

- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Example:



Scatter diagram showing the relationship between height in centimeters (cm) and weight in kilograms (Kg)

Fig.9 Scatter diagram showing the relationship between height in centimeters (cm) and weight in kilograms (kg).

III. Parameters

- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.
- For Qualitative variables: proportion and ratios are used.

$$\text{Ratio} = \frac{\text{Part a}}{\text{Part b}}$$

$$\text{Male to Female Ratio} = \frac{11}{7} \cong 1.6$$

$$\text{Proportion} = \frac{\text{Part}}{\text{Total}} \times 100$$

$$\text{Proportion of Males} = \frac{11}{18} \times 100 = 61\%$$

- For Quantitative variables: **measures of central tendency** and **measures of dispersion** (variation) are used.

A. Measures of central tendency

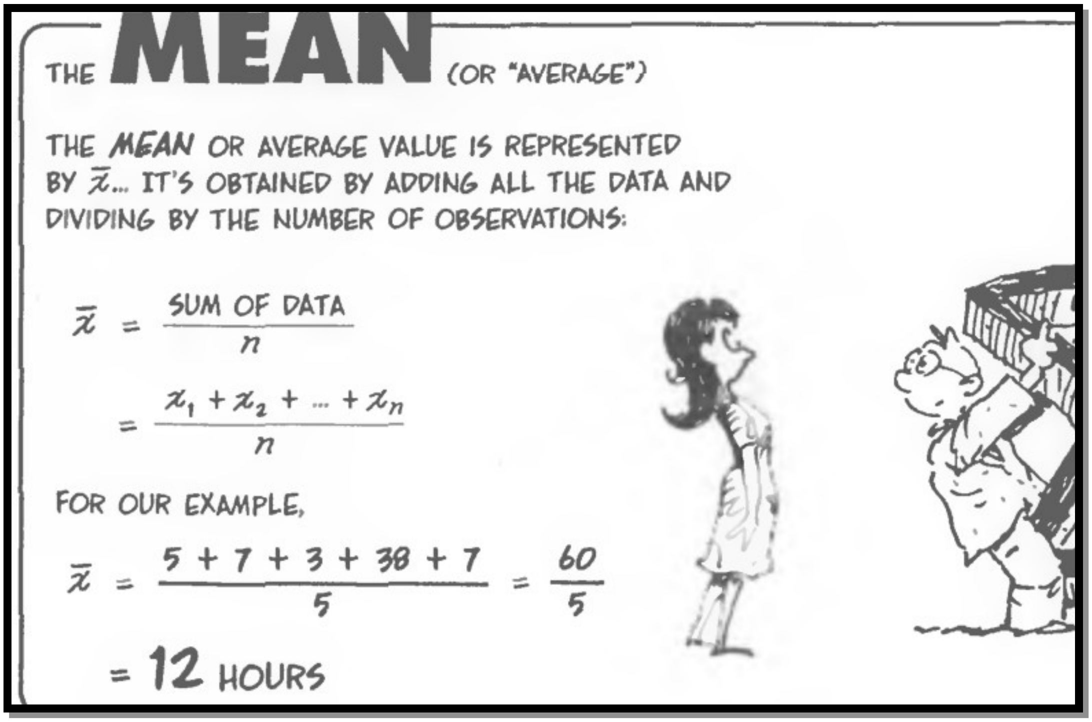
- These represent a value around most of the data cluster around.
- The mean, median, mode and midrange are statistical measures that show the average characteristics or tendency of the group.

A.1 Mean:

The mean is the sum of all the values, divided by the number of values.

Example

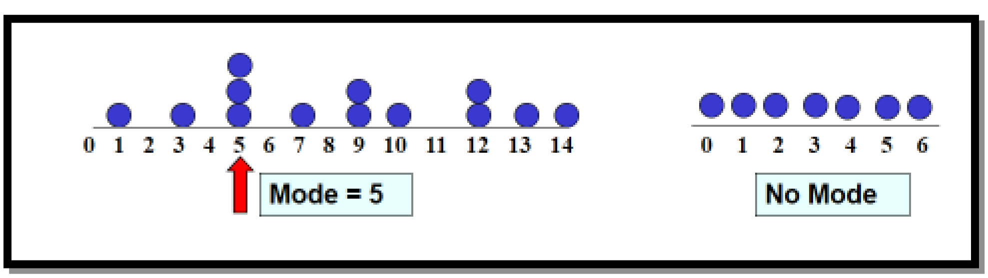
Five women in a study on lipid-lowering agents are aged 52, 55, 56, 58 and 59 years.
Add these ages together:
52 + 55 + 56 + 58 + 59 = 280
Now divide by the number of women: 56
So the mean age is 56 years.



A. 2 Mode

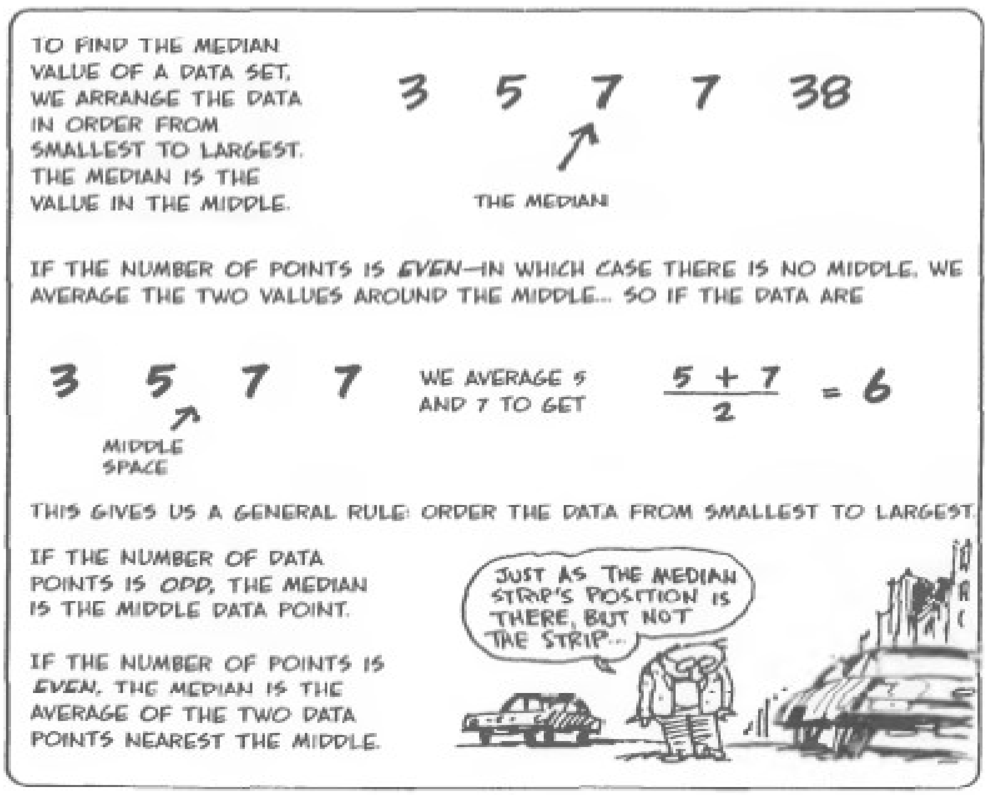
- The value that occurs most often
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes

Example



A.3 Median:

- **Median** is the middle number, or the number that divides the data in two halves. To calculate the median, the values of the continuous variables have to be sorted in an ascending or descending order, after which the middle value is chosen.
- If the number of patients is even, then the median will be the average of the middle two values.
- One of the advantages of a median is that it is not sensitive to extreme values, but the disadvantage would be that all the patients will be ignored except of the middle one.



Example:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
25	28	30	30	32	33	34	35	36	37	40	40	41	42	44	45	45	55

$$\text{order of the middle individual} = \frac{\text{number of individuals}}{2} = \frac{18}{2} = 9$$

$$\text{Median age} = \frac{36+37}{2} = \frac{73}{2} = \mathbf{36.5}\,years$$

The value belonging to the person in middle of group

B. Measures of Dispersion

They are used to measures the extent of variations between observations,

- Range
- Standard deviation
- Percentiles
- Quartiles and inter-quartile range.

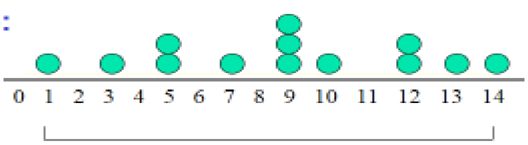
B.1 Range

- Simplest measure of variation
- Difference between the largest and the smallest observation

Example:

$$\text{Range} = x_{maximum} - x_{minimum}$$

Example:



Range = 14 - 1 = 13

B.2 STANDARD DEVIATION

Standard deviation (SD) is used for data which are “normally distributed” to provide information on how much the data vary around their mean.

$$SD = \sqrt{\frac{\Sigma\left(x - \overline{X}\right)^2}{n-1}}$$

Good news – it is not necessary to know how to calculate the SD.

SD indicates how much a set of values is spread around the average/Mean.

A range of one SD above and below the mean abbreviated to

Mean+/- 1 SD includes 68.2% of the data.

Mean+/-2 SD includes 95.4% of the data.

Mean +/-3 SD includes 99.7% of the data

EXAMPLE

Let us say that a group of patients enrolling for a trial had a normal distribution for weight. The mean weight of the patients was 80 kg. For this group, the SD was calculated to be 5 kg.

1 SD below the average is 80 – 5 = 75 kg.

1 SD above the average is 80 + 5 = 85 kg.

Mean+/- 1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 75 and 85 kg.

95.4% will weigh between 70 and 90 kg (Mean+/-2 SD).

99.7% of patients will weigh between 65 and 95 kg (Mean +/- SD).

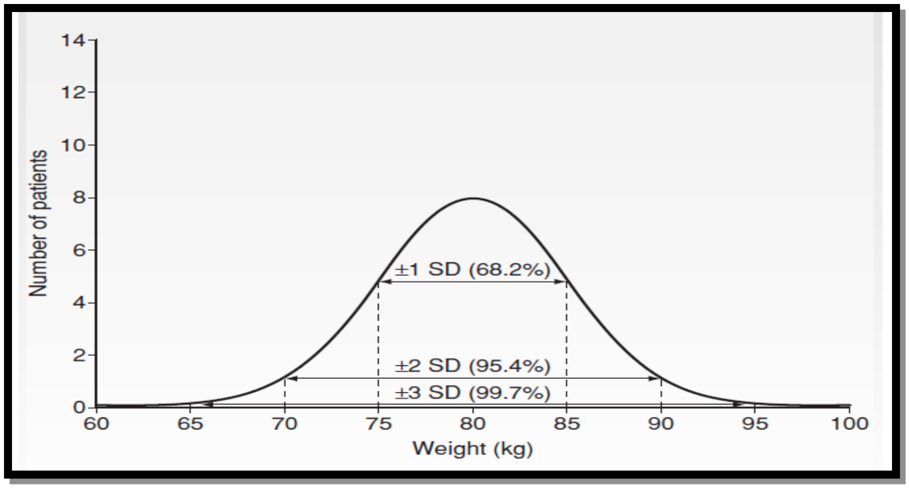


Fig. 10 Normal distribution of weights of patients in a trial with mean 80kg and SD 5kg

If we have two sets of data with the same mean but different SDs, then the data set with the larger SD has a wider spread than the data set with the smaller SD.
For example, if another group of patients enrolling for the trial has the same mean weight of 80 kg but an SD of only 3, ±1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 77 and 83 kg. Compare this with the example above.

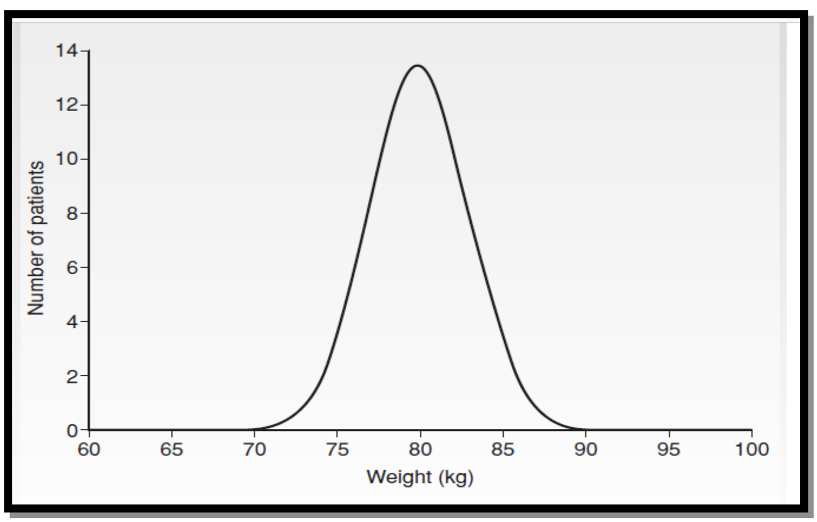


Fig. 11 Normal distribution of weights of patients in a trial with mean 80kg and SD 3kg

SD should only be used when the data have a normal distribution. However, means and SDs are often wrongly used for data which are not normally distributed.

A simple check for a normal distribution is to see if 2 SDs away from the mean are still within the possible range for the variable. For example, if we have some length of hospital stay data with a mean stay of 10 days and a SD of 8 days then:

Mean – 2 × SD = 10 – 2 × 8 = 10 – 16 = -6 days.

This is clearly an impossible value for length of stay, so the data cannot be normally distributed. The mean and SDs are therefore not appropriate measures to use.

Examiners may ask what percentages of subjects are included in 1, 2 or 3 SDs from the mean.

B.3 Percentiles

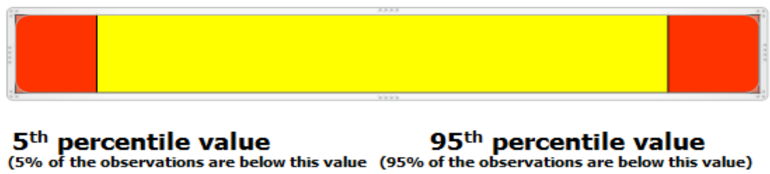
- Percentile is a score value above which and below which a certain percentage of values in the distribution fall
- Percentiles are points that divide all the measurements into 100 equal parts.
- The observations should be first arranged from the lowest to the highest values, just like when finding the median, which is the 50th percentile.
- Example: A sample of 100 newborn children was weighted, and the data were arranged from the lowest to the highest.

  The value (score) of the 95th percentile was found to be 3.8 Kg. Such findings indicates that 95% of the children had body weight less than 3.8 Kg and 5% have body weight more than 3.8 kg.

  The value (score) of the 5th percentile was found to be 2.5 Kg.

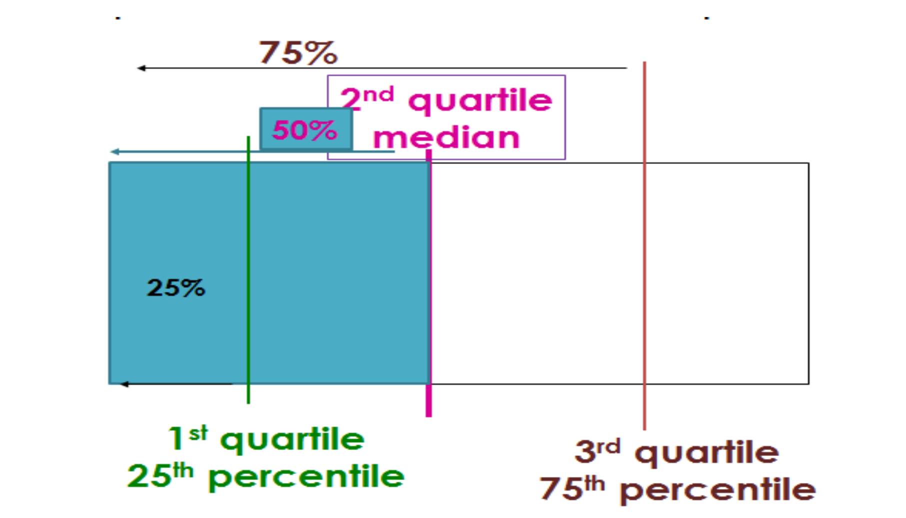
  Such findings indicates that 5% of children had body weight less than 2.5 kg and 95% had body weight more than 2.5 kg.

Children are considered within normal weight if they are between 2.5 Kg - 3.8 Kg or the 5th and 95th percentiles.



B.4 Quartiles

Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1 (lower quartile), Q2 (median), and Q3 (upper quartile)



- **Quartiles:**
  - ☐ **lower (1st)quartile (25%) = 25% lower & 75% greater**

- **upper (3^{rd})quartile (75%) = 75% lower & 25% greater**
- **Interquartile range** (IQR = Q3 − Q1)

  In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

Normal Distribution Curve

When the frequency curve is done for a large group of **quantitative values**, the curve takes the shape of the "Normal distribution"

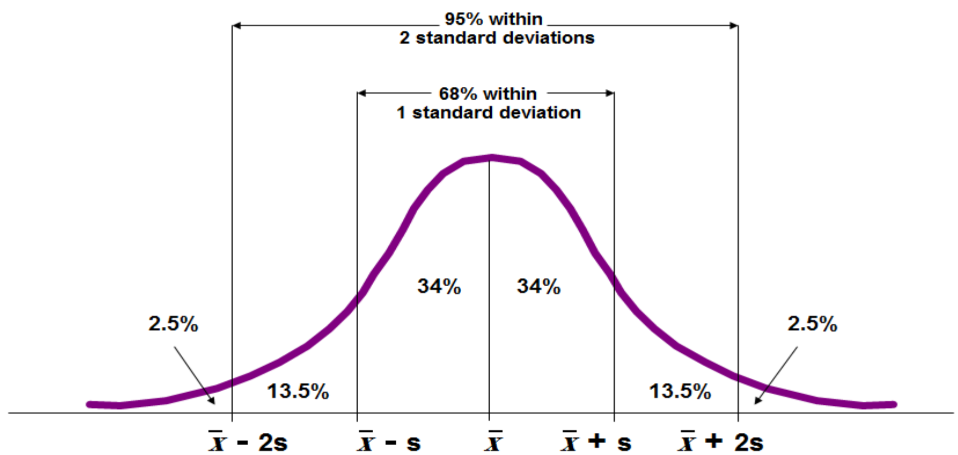


Fig. 12 Normal Distribution Curve

Characteristics of the Normal Distribution Curve

1. A bell shaped curve with most of the values clustered near the mean and a few values out near the tails.
2. The normal distribution is symmetrical around the mean.
3. The mean, the median and the mode of a normal distribution have the same value.
4. The number of individuals with certain range of values is known:
    - Mean +1 SD = 68 % of observations
    - This means that about two thirds (680) of the healthy males have systolic blood pressure values between 115 and 125 mm/Hg.
    - Mean + 2 SD = 95 % of observations
    - This means that 950 healthy males have systolic BP ranges between 110 and 130 mm/Hg.
5. The normal distribution curve does NOT mean it is the distribution of normal healthy people; The word normal refers to the distribution of values

   Example:

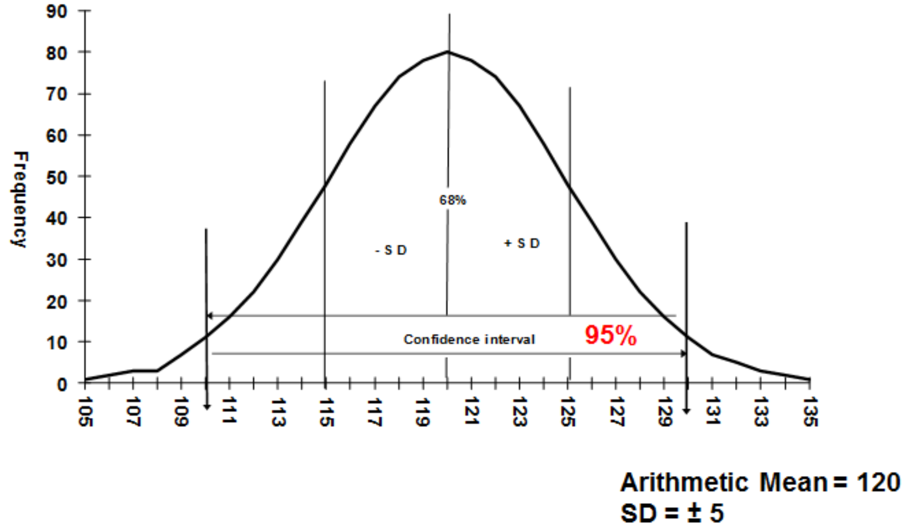


Fig. 13 Normal Distribution Curve of Systolic Blood Pressure of groups of individuals

- **Mean +/- 1 SD includes 68.2% of the dat**a; it means that 68% of the individuals have systolic blood pressure range from 115-125
- **Mean +/-2 SD includes 95.4% of the data**; it means that 95.4% of the individuals have systolic blood pressure range from 110-130
- **Mean +/-3 SD includes 99.7% of the data**; it means that 99.7% of the individuals have systolic blood pressure range from105-135

Symmetrical data (normally distributed data):

Data that follows normal distribution (mean=median=mode)

Report by mean & standard deviation & n

Skewed data (not normally distributed data):

Not normally distributed (mean≠median ≠mode)

Report by median & IQ Range

Hypothesis testing

The statement being tested in a statistical test is called the null hypothesis. The test is designed to assess the strength of the evidence against the null hypothesis.

- Null hypothesis (H0) is usually a hypothesis of "no difference" e.g. no difference between blood pressures in group A and group B.
- Alternative hypothesis (H1) is usually the hypothesis you set out to investigate. It contradicts the null hypothesis. It is accepted if the test of significance rejects the null hypothesis.

    For example, question is "is there a significant (not due to chance) difference in blood pressures between groups A and B if we give group A the test drug and group B a sugar pill?"

- If statistical analysis indicates that the difference or effect is not likely to have occurred by chance then the null hypothesis is rejected in favor of the alternative hypothesis, stating that a real effect has occurred.
- Instead, a finding is described as “not statistically significant” if the null hypothesis is accepted and “statistically significant” if the alternative hypothesis is accepted.

P Values

- The P value gives the probability of any observed difference having happened by chance.
- It is not important to know how the P value is derived – just to be able to interpret the result.

- The P (probability) value is used when we wish to see how likely it is that a hypothesis is true. The hypothesis is usually that there is no difference between two treatments, known as the “null hypothesis”.
- $P = 0.05$ means that the probability of the difference having happened by chance is 0.05 in 1, i.e. 1 in 20
- It is the figure frequently quoted as being “statistically significant”, i.e. unlikely to have happened by chance and therefore important.
- The lower the P value, the less likely it is that the difference happened by chance and so the higher the significance of the finding.
- Example:

  Out of 50 new babies on average 25 will be girls, sometimes more, sometimes less. Say there is a new fertility treatment and we want to know whether it affects the chance of having a boy or a girl.

  Therefore we set up a null hypothesis that the treatment does not alter the chance of having a girl.

  Out of the first 50 babies resulting from the treatment, 15 are girls. We then need to know the probability that this just happened by chance, i.e. did this happen by chance or has the treatment had an effect on the sex of the babies?

  The P value gives the probability that the null hypothesis is true.

  The P value in this example is 0.007.

  Do not worry about how it was calculated, concentrate on what it means. It means the result would only have happened by chance in 0.007 in 1 (or 1 in 140) times if the treatment did not actually affect the sex of the baby.

  This is highly unlikely, so we can reject our hypothesis and conclude that the treatment probably does alter the chance of having a girl.

Statistical data analysis

Statistical analysis aims at:

A. Comparisons between groups (e.g. diseased and non-diseased according to exposure) where tests of significance are used.

1. **comparing** between 2 groups regarding **qualitative** variable (e.g. smokers and non-smokers regarding chronic bronchitis) so we should use **Chi square test**

2. **comparing** between 2 groups regarding **quantitative** variable (e.g. smokers and non-smokers regarding heart rate) so we should use **independent sample t-test** if the data was normally distributed and use **Mann-Whitney test: For testing equality of two medians** if the data not normal y distributed

B. Association: Correlation and regression. (e.g. increase in the level of LDL with the increase duration of exposure to smoking)

Detecting the association (**correlation**) between **2 quantitative variables** (e.g. age and creatinine level) so we should use **Pearson correlation** if the data was normally distributed

Screening test

Tests done among apparently well people to detect early (asymptomatic) disease or precursors of disease

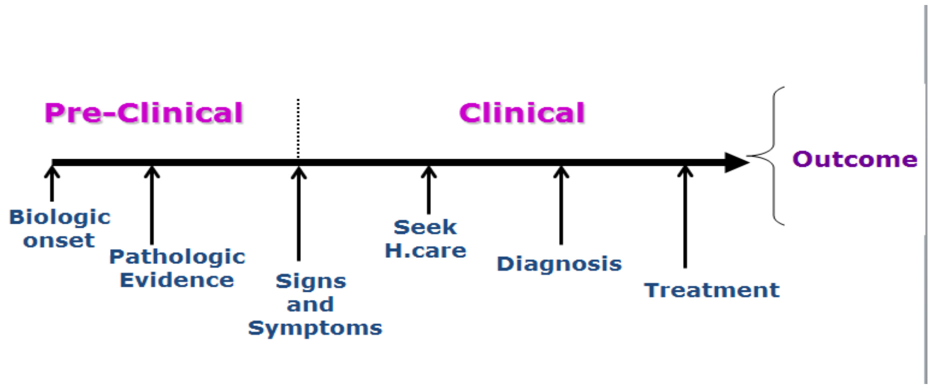


Fig. 14 Natural History of the Diseases

- The identification of an undiagnosed disease, condition or risk factor by;
  - Physical examination (blood pressure measurement)
  - Laboratory testing (e.g. cholesterol level)
  - Type II diabetes
  - Thyroxin in newborn
  - Cytological screening for cervical cancer

Screening tests are used to identify patients who should be referred for further diagnostic evaluation. To determine the quality of a screening test, it is best to have a gold standard to compare it with. The gold standard provides a definitive diagnosis of the disease. For healthy individuals, the tests, if they are numerical, there is a range called the normal range

Benefits of screening test

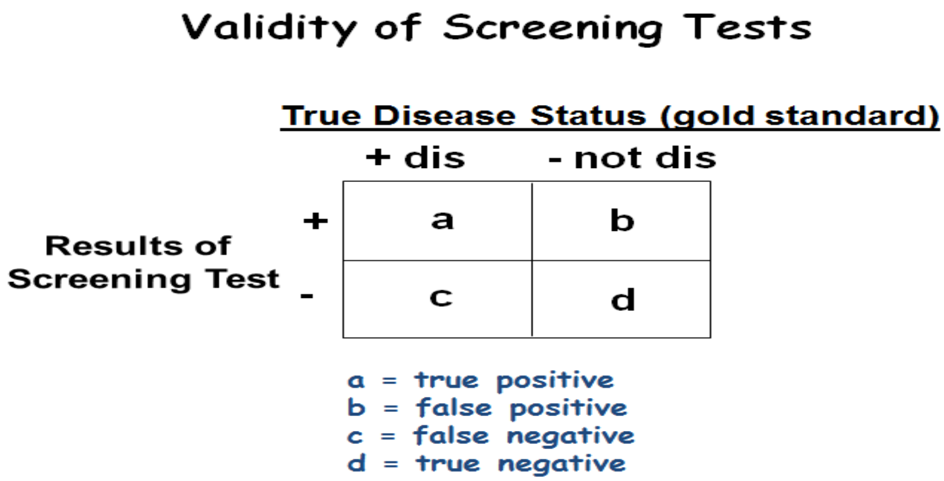
- Screening permits **early detection** of disease and medical referral
- Clinicians then make an **early diagnosis**
- prescribe **early treatment** lead to **better outcome**

Criteria for Good Screening Test

- Simple & quick
- Can be done by paramedics
- Safe
- Inexpensive
- Acceptable to population
- Reliable
- Valid and Accurate

Validity of a Screening Test

How good is the screening test compared with the confirmatory diagnostic test?



Sensitivity and Specificity

They are used to analyze the value of screening tests.

Think of any screening test for a disease. For each patient:

- the disease itself may be present or absent;
- the test result may be positive or negative

Sensitivity: If a patient has the disease, we need to know how often the test will be positive, i.e. "positive in disease".

This is calculated from: **A/ A + C**

This is the rate of pick-up of the disease in a test, and is called the *Sensitivity.*

Specificity: If the patient is in fact healthy, we want to know how often the test will be negative, i.e. "negative in health".

This is given by: **D/ D + B**

This is the rate at which a test can exclude the possibility of the disease, and is known as the S*pecificity.*

Sensitivity & Specificity

	Disease Present	Disease Absent
Test Positive	True Positive	False Positive
Test Negative	False Negative	True Negative
Total	All Diseased	Not diseased

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative+ False Positive}}$$

١٩

Sensitivity & Specificity

	Disease Present	Disease Absent
Test Positive	a	b
Test Negative	c	d
Total	a+c	b+ d

a= true positives

b=false positives

c= false negatives

d= true negatives

Sensitivity= a / a+c*100

Specificity= d / b+d*100

٢٠

Validity of Screening Tests

Physical Exam & Mammo-graphy		Breast Cancer +	Breast Cancer -
	+	132	983
	-	45	63650

Sensitivity: a / (a + c)
Sensitivity = 132 / (132 + 45) = 74.6%

Specificity: d / (b + d)
Specificity = 63650 / (983 + 63650) = 98.5%

Sensitivity: Screening by physical exam and mammography will identify 75% of all true breast cancer cases.

Specificity: Screening by physical exam and mammography will correctly classify 98.5% of all non-breast cancer patients as being disease free.

Choosing the right test

- An ideal test of 100% sensitivity and 100% specificity does not exist.
- We generally have to choose from the available range of tests with varying sensitivities and specificities.

How does one choose the right test?

- If very important not to miss a disease which is serious and potentially treatable (cancer, for example), it would be better to use a test which has greater sensitivity. One would like to pick up as many cases as possible doing this test

- On the other hand, if making a positive diagnosis would result in much worry, stigma (HIV, for example) or cost, then it would be better to use a test which has high specificity.

Group Discussion

1. **The median of the following data, is: 1,2,4,6,8,10,11,13**
   a. 6
   b. 8
   c. 7
   d. 10
   e. 9

2. **A household survey of 10 families was conducted by students of 2nd year. In the collected data, the ages of heads of families were: 32, 34, 35, 36, 36, 42, 44, 46, 48, and 52. The mean age of heads of families is**
   a. 36
   b. 38.5
   c. 40
   d. 40.5
   e. 42

3. The birth weights in a hospital are to be presented in a graph. This is best done by a:

a. Bar diagram
b. Pie chart
c. Histogram
d. Pictogram
e. Frequency chart

4. If six families were surveyed and the numbers of children per family were found to be 2, 3, 4, 4, 5, 6, find the mean number of children per family

a. 2

b. 3.5

c. 4

d. 6

e. 4.5

5. Which of the following can have more than one value?

a. The mean

b. The range

c. The mode

d. The median

e. Standard deviation.

6. A study was conducted to assess the height of students of 4th year in 10 Medical colleges the values of heights ranged between 5.5 – 5.10 feet. A histogram has been selected by the researcher to present these results as it is a:

a. Nominal data

b. Categorical data

c. Both qualitative and quantitative data

d. Continuous data

e. Discrete numerical data

7. After arranging the data is ascending or descending order of magnitude, the value of middle observation:

a. Mean

b. Mode

c. Median

d. Geometric mean

e. Mean deviation

8. A household survey of 10 families was conducted by students of 4th year MBBS. In the data they collected, the ages of heads of families were: 32, 32, 36,48,34,46,35,44,36 and 32 years. The mode in this series:

a. 32

b. 34

c. 36

d. 44

e. 46

9.The following are the weights of 7 newborn infants recorded at the primary health care center during visit to the well-baby clinic.

4.5, 4.5, 4.4, 2.7, 2.1, 3.2, 3.6

1. a Calculate the mean, the median and the mode

Mean =

Median=

Mode=

References

- **Honorary D.B.**: Medical **Statistics from Scratch**. **An Introduction for Health Professionals Second Edition**

- **Harris M. and Taylor C.: Medical statistics made easy, 2003 Martin Dunitz, an imprint of the Taylor & Francis Group, ISBN 0-203-50277-9 Master e-book ISBN**

- **Chernick M.R.: The Essentials of Biostatistics for Physicians, Nurses, and Clinicians, 2011 by John Wiley & Sons, Inc. ISBN: 978-1-118-07195-3**

- **Gerstman B. B.: Basic biostatistics, statistics for public health, jones and Bartlett publisher 2008**