Lecture 1

DEFINITIONS;

We are living in the information revolution that about 0.5 million new articles published only in the medical field.

Health statistics; it is the subject or science of data collection, summarization, analysis, presentation of information at valid conclusion for decision making.

Data; is a discrete observation about attribute, phenomena, facts, events or individuals made on populations or samples (it's the raw material).

Value; it's the numerical representative of the measurement of the variable.

Variable; is a symbol of person, object or phenomena that to be study & show the difference between the members of the sample or population & can assume any value with a standard scale.
variable could be quantitative or qualitative;
Quantitative variables; are numerical data arising from count or measurement When the value is a **whole number & not decimal** (books, tables, newborn), the variable / data is called **Quantitative discrete variable** While **if the value can full in continuum & decimal number it will be called Quantitative continuous variable** (weight, height, hemoglobin).
Qualitative variables; arise when objects, persons, or phenomena full in categories& have no numerical relation (gender, color).
Variables either: -Independent:- affects, causes or associated with the problem as
Smoking **Independent variable** **(** Smoking + Lung cancer)

Dependent variable; that measures the problem under study
smoking غير مستقل ويعتمد على وجود ← **dependent variable**← Lung cancer

- **Information;** is the end result of statistical manipulation of data / variable

- **Population;** is a group of individuals, objects, subjects, share same characteristics.

- **Sample**; is a subset of population & should represent it .

- **Measurement;** is the term defined as the procedure of applying a standard scale to a variable or a set of values to evaluate,, or qualify them.

Measurement Scales; There are five scales**;**

1. **Dichotomous**; when the scale contains only (2) mutually exclusive categories e.g; gender (male & female).

2. **Nominal;** when the scale contains more than (2) unordered mutually exclusive categories e.g; race & religion.

3. **Ordinal;** when the scale contains (2) or more ordered mutually exclusive categories e.g; social class (1, 2, 3, 4) malnutrition (mild, moderate, sever

4. **Interval scales**; has equal distance with any 2 points with intervals.

5. **numerical scales;** deal with numbers & could be Quantitative discrete data no. of pregnancies) or Quantitative continuous (level of Hb)

Measurement Criteria;

Accuracy; conformity to a standard value

Precision; the quality of being sharply defined through exact details.

Reliability; having the same results when repeated under the same conditions (measurement of blood pressure).

Validity; highly sensitive true (+ve) & specific true (-ve)

DALYS; ((Disability-adjusted life years lost)) – an international of burden of that expresses both time lost through premature death & time lived with disability. **Meta-analysis**; a methodology to a critically review research studies & statistically combine their data to help answer questions (collection of studies among one disease all over the world).

Multi variant analysis (Anova); assessment of multiple independent variables on dependent variable.

Proposal; A document written for the purpose of obtaining funding for a research study project.

Protocol; the detailed written plan of the study & any research study should have a protocol.

Pilot study; a preliminary study to test the feasibility of the protocol before implementing the study proper– also called pre test.

Questionnaire; means collecting data from people where they provide written or responses to a set of questions, as in their own wards (open ended questions) by selecting form among pre-defined answers (closed response question).

Estimations; Statistical M & **Parameter M;** mean of sample estimates mean of population & SD of sample estimates of parameters that called point estimation (limited)

Uses of Statistics:

1. To measure the health status of the community.

2. To compare the health status of the community with other.

3. For planning of health services.

4. For evaluation of health services.

5. For estimation of future needs.

6. Apply study results to patient care.

7. Interpreting the vital statistics, information about drugs &Equipments, using diagnostic measures.

8. Understanding the epidemiological problems. 9. Appraising guidelines.

10. Evaluation study protocols & articles & participate in research projects.

Lecture 2

Health statistics it is the science of data collection, summarization, analysis, presentation of information at valid conclusion for decision making.

* Collecting data
* Summarizing data
* Interpreting data
* Presenting data
* Using of test hypothesis

Data Collection:- many mechanisms are used for collection of data which form base of health status , that should be standardized , nationally , internationally .

Census of population:-provide base for calculation rate , variance .

- it count of people & record age , sex ,social status ,distribution &other information.
Population pyramid :

- age, sex , distribution of rural & urban .

- put them in histograms with 5years intervals.

- in developing c→ birth rate < Death of infant , with less Reaching elderly.

- in developed c→ birth rate > death of infants , more reaching elderly .

Local census:- do census in local areas for accurate epidemiology

Registrations of birth & death:

- should be compulsory specially in developing countries .
- even its compulsory in developing countries, but some of these countries not apply it

- birth & death records are very important to know the health status of community & should be collected by will trained persons , in addition to recording fact of death , causes of death start from 1st level →diagnosis till reach health axillaries & more specific diagnosis with highly trained persons) & post mortem diagnosis & examinations.

Methods improving registrations :

- registration centers :

Should have registration centers even in districts & adapt to social structure using persons as village heads , head of compounds , religious , & built institution in far areas .

- Rewards &penalties :

Rewards → as free primary school may be available for children whose birth have registered.

Penalties → no school admission for whom not have birth registered.

Education : → should reach proper information to the general population about the program.

Notification of diseases:

National notification :

- In every country should have a list for certain diseases, cases of which must be reported to the appropriate health authority.

- It includes of communicable diseases & industrial diseases.

- For acute epidemic diseases → the notification should be designed to provide the health authorities with information at nearly stage to take urgent action to control out break.

- Notification of chronic diseases & non epidemic ? provides information that can be used for long term planning , health services , monitoring of control programs.

Information can be collected from:

1. Census of population: it counts of people & record age, sex, social status, distribution &other information.

2**.** Registrations of birth & death records are very important to know the health status of community.

3. Notification of diseases.

4. Data from medical institutions as outpatient, inpatient

5. House to house visits

6. Mailed questionnaire

Methods of Data collection

- Sampling methods.
- Designing questionnaire.

Designing questionnaire

- It 's a printed paper or form that contains questions, used to collect information in a survey that should be: short, clear with simple language could be open ended, close ended.

Lecture 3

Summarizing Data

Is the organization of data in a way for easy understanding the first step of data interpretation (analysis).

Consists of the following steps:

- Data entering.
- Ordered array.
- Summarization

Data entering:

Either manual or using of computers for data entry.
Nowadays, many software are developed for data entering, presentation and data analysis a s : MS Excel, Epi-Info, SPSS, Stata.

Ordered array

It is the first step in the process of data organization after data entering.

An ordered array is a listing of values from the smallest value to the largest value that enables to determine quickly the value of the smallest measurement& the value of the largest measurement that helps to determine roughly, the proportion of people lying below or above certain value.

Summarizing data

- Summarizing qualitative data (Frequencies).
- Summarizing Quantitative data (Frequency distribution).

Summarizing qualitative data (Frequencies).

It is counting the number of observations in each category. These counts are called frequencies, they are often also presented as relative percentages of the total numbers.

Summarizing quantitative data frequencies):

(Frequency distribution):

- One of the ways of summarizing data.

-It is a table, showing the number of observations at different values or within Certain range.

Steps of constructing frequency distribution:

- Count the number of observations.
- Identify the lowest & highest values.
- Group the data, by selecting a set of class intervals to find the number of Class Intervals
- Determine width of class intervals

Summarizing Quantitative data:

- Frequency distribution (frequency tables).

- Measures of Variability (Range, Variance, Standard Deviation,

Standard Error, Coefficient of Variation)

- Central Tendency (Mean, Median, Mode), gives an idea of what is

- Common value for a given variables

Class intervals: There must be no overlapping between these intervals, like

0-5, 5-10, 10-15, 20-25, 25-30, 35-40 ---------

few & many intervals are undesirable, because few intervals is losing information while many intervals is cancelling the objective of summarization.

The appropriate No. of class intervals is **6-15**.

Number of class intervals
Sturge's rule:

Sturge's rule (formula): **K= 1+3.322 (log n)**

K = No. of class intervals. n = sample size.

Note: can increase or decrease the No. of class intervals for convenience and clear presentation.

Width of class intervals; Width (W) of class interval, in general, is equal, but some times this is not possible.

W = R/K

K = No. of class intervals

R = Range (difference between smallest and largest observation).

For convenience, a width of **5** units or **10** units is used.

Data ; may be grouped or Ungrouped

Ungrouped Data:

3 , 5 ,7 , 2 , 3 , 7 ,10 , 4 , 3 , 5 ,7 , 4 , 2 , 2 , 5 ,4 , 10 , 8 , 5, 18 , 10 , 20 ,10 , 15 , 11 , 15 , 10 , 18 , 10 , 20 ,10 , 15 , 10 , 20 ,10 , 15 , 11 , 15 , 10 , 18 ,11 , 8

1. Ordered array; ordered data quickly arrange data from smallest values to biggest values:

2 , 2 , 2 , 3 , 3 , 3 , 4 , 4 , 4 ,5 , 5, 5 , 5, 7 , 7 , 7 , 8 , 8 , 10 , 10 , 10 , 10 , 10 , 10 , 10 , 10 , 10 ,11 , 11 , 11 , 15 , 15 , 15 , 15 , 15 , 18 , 18 , 18 , 20 , 20 , 20

2. Frequency distribution;

2	**-**	**3**
3	**-**	**3**
4	**-**	**3**
5	**-**	**4**
7	**-**	**3**
8	**-**	**2**
10	**-**	**9**
11	**-**	**3**
15	**-**	**5**
18	**-**	**3**
20	**-**	**3**

3. Number of class intervals

Sturge’s rule (formula):

K= 1+3.322 (log n)

K= 1+3.322 (log 41) = 1+3.322 × 1.623 = 1+5.391 = 6.391˜ 7

4. Width of class intervals

W = R / K

W = 18/7= 2.571 ˜ 3

5. Frequency Distribution:

Age class	Frequency	Relative Frequency		Cumulative Relative Frequency
			%	
50-59	8	0.123	12.3	12.3
60- 69	10	0.154	15.4	27.7
70- 79	16	0.246	24.6	52.3
80-89	14	0.215	21.5	73.8
90-99	10	0.154	15.4	89.2
100-109	5	0.077	7.7	96.9
110-119	2	0.031	3.1	100
	65			

Lecture 4

Central tendency

The main purpose of computing measures of central tendency is to give an idea of what is a typical or common value for a given variable. Central tendency is including: Mean, Mode and Median

1. Mean (Average):

Mean is the most common statistic used to measure the center, or middle, of a numerical data set. Denoted as $\overline{X}$

Mean is calculated by "the sum of all the numbers divided by the total Number of numbers (sample size)".

Mean $\overline{X}$ = ***X*/** N f**or ungrouped data** ; sum of scores/number of scores

Example; 4,3,2,6,5 mean=2,3,4,5,6/5=6

For grouped data; a- Find mid Interval class (X)

b- Find frequency of each class (F)

c- Find sum of frequency sum F

d- Find sum of FX

e- Use equation **′x = Σ F X / Σ F**

Example for grouped data;

Kg	Mid class (X)	Frequency (F)	F X
60-62	61	5	305
63-65	64	18	1152
66-68	67	42	2814

Mean = Σ F X/Σ F

2814/ 42 = 67

Advantages of Mean: It´s always present & unique

- It is simple to compute.

- All values are included.

- It´s enable to test of statistical significance

Disadvantage: The main disadvantage of mean is the presence of extreme values, i.e. (very high or very low values).

2. Median (50th percentile) (ungrouped data):

The median of a data set is the value that lies exactly in the middle.

The position of the median depends on the number of observations

□ For odd number of observations (n+1 /2)

or mid point of scores

□ For even number of observations (n/2) & (n/2 +1)

Or Σ of 2 middle scores / 2

Grouped data:

$$\text{median} = L + \frac{w}{f_{\text{med}}}\left(0.5n - \sum f_b\right)$$

Step 1: Construct the cumulative frequency distribution.

Step 2: Decide the class that contain the median.

That the first class with the value of cumulative frequency equal at least N/2

(N = the total frequency/2)

fb = (the frequency before class median)

L = (the lower boundary of the class median)

Or Median = L+ R/F × W

Where L: lower limit of class interval containing median.

R= (N/2) - previous cumulative frequency

W= width of C I.

Or Median = $L + J / F \times (U - L)$

Where $J = n/2 – (fm - 1)$

Advantages of median: - It is unaffected by extreme values.

- Can be used of width of C I

Disadvantages of median:

□ It provides no information about all values (observations).

□ It's less enable than the mean to test of statistical significance.

3. Mode:

It is the value that is observed most frequently in a given data set may be one mode, multiple modes or no mode.

Advantage of Mode:

- Sometimes gives a clue about the etiology of the disease.
- Unlike other measures can be used for qualitative data

Disadvantages of Mode:

□With small number of observations, there may be no mode.

□It is less enable to test of statistical significance.

Example: Data a range---- 3, 5 , 5 , 9 , 8 FIND ; (ungrouped data)

Mean = (3+5+5+8+9) / 5 = 6 Median = 5 Mode = 5

Measurement of Variability:

Measures of dispersion: Dispersion refers to how variable "spread out" data values are: for this reason measures of dispersions are sometimes called "measures of variability" or "measures of spread." Knowing the dispersion of data can be as important as knowing its central tendency. They are measure the distribution or dispersion of the data.

1. Range:

The range is the largest value in the data set minus the smallest value in the data set. Range (R) = Largest value - Smallest value

2. Variance:

Variance is the “sum of the squared deviation of the values from the mean divided by sample size minus one”.

Variance is calculated by the following equations:

$$V= \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1}$$

3.The standard deviation (ungrouped data):

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}}. \qquad S = \sqrt{V}$$

S.D = √ Σ (X - X¯) ² / √ (N-1)

X = sample score X¯ = mean of sample N = number of sample size

Example; scores (3+5+5+9+8) find; S.D

mean (X¯) = sum of scores / number of scores= 6

N (x)	(X - X¯)	(X - X¯) ²
3	3 - 6 = - 3	9
5	5 – 6 = - 1	1
5	5 – 6 = - 1	1
8	8 – 6 = 2	4
9	9 – 6 = 3	9
		24

S.D=√Σ(X - X¯) ² /√ (N-1) S.D = √ (24/4) = 4.9/2 = 2.4

Grouped data:

$SD = \surd\ (\Sigma F\ (X)^2 - (\Sigma FX)^2 / F) / \surd\ (F-1)$

Importance of standard deviation

The most common measures of dispersion for continuous data are the Variance and standard deviation, both describe how much the individual values in a data set vary from the mean or average value.

- The variance and standard deviation, usually accompanied by the mean, help to know how a set of data values distributes around its mean.
- A small standard deviation means that the values in the data set are close to the middle of the data set, on average while a large standard deviation means that the values in the data set are farther away from the middle, on average.

4. Coefficient of Variation (CV): The coefficient of variation (CV) allows us to compare the variation of two same variables (or more) different variables or with different means.

$$CV = \frac{s}{\bar{X}} X\ 100$$

Example: weight measures in kg for both 2 variables, or one measures by kg & other by bound

Advantage of Coefficient of Variation:

When two data set distributions have means of different magnitude, a comparison of the C.V. is therefore much more meaningful than a comparison of their respective standard deviation.

5. Standard Error of the Sample mean (SE):

- the sample mean is unlikely to be exactly equal to the population mean.
- the standard error measures the **variability of the mean of the sample** as

an **estimate of the true value of the mean for the population from which the sample was drown**

$$SE = \frac{sd}{\sqrt{n}}$$

Q: The Following data shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic .

7	**12**	**8**	**3**	**5**
12	**3**	**1**	**13**	**4**
4	**5**	**7**	**3**	**3**
8	**1**	**17**	**4**	**5**
3	**17**	**4**	**7**	**8**
1	**0**	**4**	**7**	**8**
1	**1**	**8**	**1**	**10**
5	**8**	**7**	**2**	
13	**7**	**3**	**5**	
1	**10**	**7**	**11**	

From these data construct:

1. **A frequency distribution**
2. **Cumulative distribution**
3. **A relative frequency distribution& cumulative relative distribution**
4. **Find central tendency & Find measurement of variability**

Hours	Freq	Hours	Frequency
1	4	8	6
2	1	10	3
3	6	11	2
4	5	12	2
5	5	13	2
7	7	17	2

1**. Sturge's rule (formula):K= 1+3.322(log n)**

K = 1+ 3.322× 1.653 = 6.49 ------------ 6

2. Width of class intervals

W= R/K = 16/6 = 2.67 ------- 3

Frequency distribution :

Class	Frequency	Cumulative Frequency	Relative Frequency %	Cumulative Relative Frequency
1-3	11 Fb	11	24.44	24.44
4-6	10 Fb	21	22.22	46.66
7-9	13 Fm	34	28.89	75.55
10-12	7	41	15.55	91.1
13-15	2	43	4.44	95.54
16-18	2	45	4.44	99.98

4. Central tendency; Mean:

Class	M. C. I (X)	Frequency	FX
1-3	2	11	22
4-6	5	10	50
7-9	8	13	104
10-12	11	7	77
13-15	14	2	28
16-18	17	2	34
		45	315

<u>Mean</u> :Σ F X / Σ F = 315 / 45 = 7

$$\text{median} = L + \frac{w}{f_{\text{med}}}\left(0.5n - \sum f_b\right)$$

Step 1: Construct the cumulative frequency distribution.

Step 2: Decide the class that contain the median.

That the first class with the value of cumulative frequency equal at least n/2 = 45 / 2 = 22.5 (n = the total frequency)

So 34 is the cumulative frequency for fm = 13 (the frequency of the class median)

fb = 10 + 11 = 21 (the frequency before class median)

L = 7 (the lower boundary of the class median)

W = the class width

$$\text{median} = L + \frac{w}{f_{\text{med}}}\left(0.5n - \sum f_b\right)$$

Or Median= L+r/f ×W Where L lower part f C I containing median r=(n/2- the previous cumulative frequency)

W=width of the C I

Or Median = L + J / F × (U - L) للاطلاع

Where J= n/2 – (fm - 1)

Median = 7 + (3 / 13)X (22.5 - 21) = 7.23X 1.5 = 10.84

Mode= 13 for class interval 7 - 9

Measurement of Variability :

$SD = \sqrt{\ }\ \Sigma F (X)^2 - (\Sigma FX)^2 / F / \sqrt{\ }(F-1)$

Class	M.C.I (X)	F	FX	X^2	$F (X^2)$
1-3	2	11	22	4	44
4-6	5	10	50	25	250
7-9	8	13	104	64	832
10-12	11	7	77	121	847
13-15	14	2	28	196	392
16-18	17	2	34	289	578
Total		45		315	2943

$$SD= \frac{\sqrt{\ }\ 2943 - ((315)^2/45}{\sqrt{45 - 1}} = \frac{\sqrt{\ }\ 2943 - 2205}{\sqrt{44}} = \sqrt{738} / \sqrt{44} = 27.166 /6.63$$

= 4.097 **Range = 13 – 2 = 11**

C.V = SD/ Mean X 100 **4.097 / 7 X 100 = 58.53 %**

$V = S^2 = 16.79$ $SE = SD /\sqrt{N} = 0.61$

Lecture 5

Presenting Data:-

Statement	Quantitative Variable	Qualitative Variable
Definition	Counted measured	Categorical nominal
Method of collection	Interviews , observations	Interviews
Tools of collection	Questionnaires , check list	Questionnaires, check list
Summarization	Central tendency , S.D Relative frequency	Ratio , rate , % proportion Relative frequency
Methods analysis **presentation**	T – student test correlation & regression table , curve , histogram	x^2 test, diff 2 proportion, tables , pie, Bar Chart

Types of data presentation:

1. Texture presentation

2. Tabular presentation

3. Diagramatic presentation

1. <u>Texture presentation</u> :

Example: a health survey was conducted with population of 5000 which Females 2400, males 2600

Child bearing age females 1000 --------

2.Tabular presentation :

During the year 500 cases of diarrhea with 200 of males & 300 0f female , 600 cases of T. B with 300 of males & 300 0f female .

Diseases	Male	Female	Total
Diarrhea	200	300	500
T. B	300	300	600
Total	500	600	1100

Characteristics of a statistical table

1. Serial number

2. Title: should have a precise title , simple ,self explanatory ,refers to place time, person

3. Left column has different items on which the information have been collected .

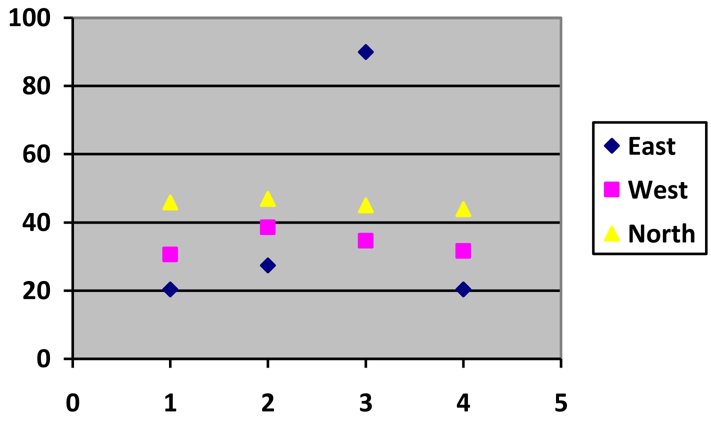
4. Caption: the heading column is indicting different categories, different periods --- that called Caption

5. Box head: put the title puts in upper middle part of table.

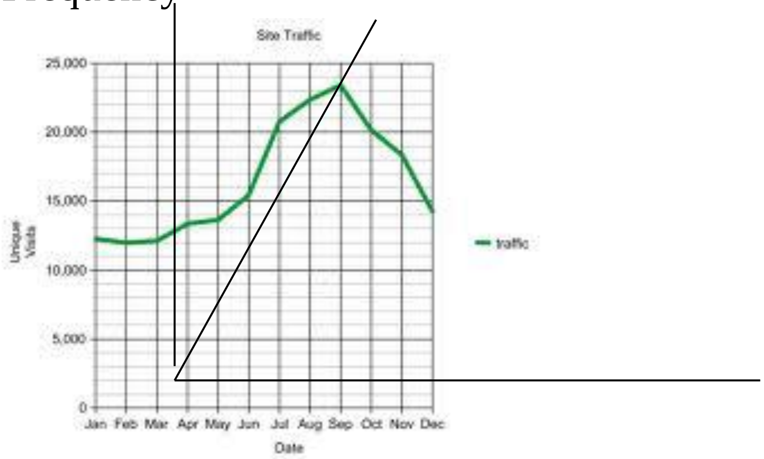
6. Body of the table: all numerical data put in the body of the table

7. Footnote : below the table that indicates the source of the data , any remarks or abbreviations should put in footnote

3.Diagramatic presentation : - line Diagram



Frequency

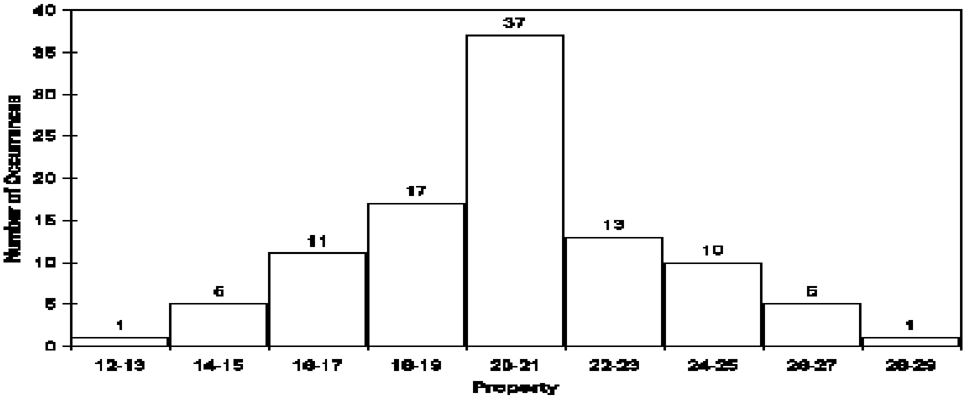


Categories

Histogram: The frequency histogram is a very effective graphical and easily interpreted method for summarizing data provides information about:

- the average (mean) of the data
- the variation present in the data
- the pattern of variation
- whether the process is within specifications

frequency



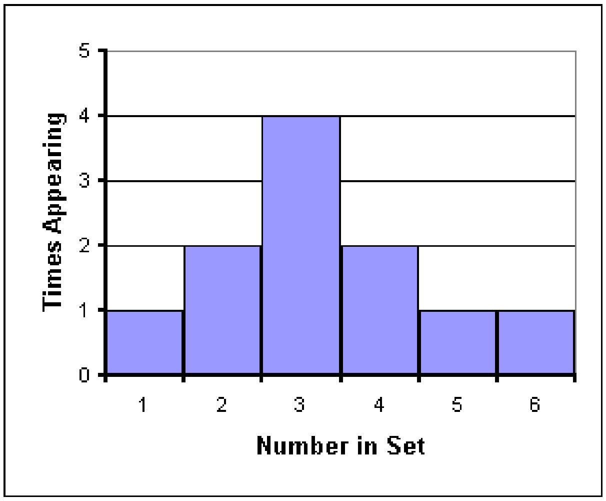
histogram:

Drawing Frequency Histograms

In drawing frequency histograms, put in mind the following rules:
- Intervals should be equally spaced
- Select intervals to have convenient values
- Number of intervals is usually between 6 to 15 or 20
- Small amounts of data require fewer intervals
- 10 intervals is sufficient for 50 to 200 readings

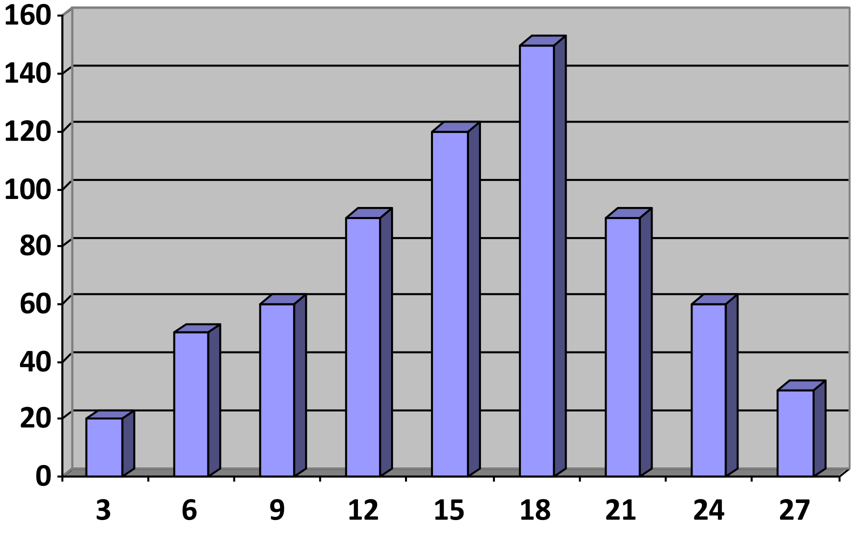
What is a Histogram?

A histogram is "a representation of a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies." groups of numbers according to how often they appear. Thus if we have the set {1,2,2,3,3,3,3,4,4,5,6}, we can graph them like this:



Part of the power of histograms is that they allow us to analyze extremely large data sets by reducing them to a single graph that can show primary, secondary and tertiary peaks in data as well as give a visual representation of the statistical significance of those peaks. To get an idea, look at these three histograms:

Frequency distribution

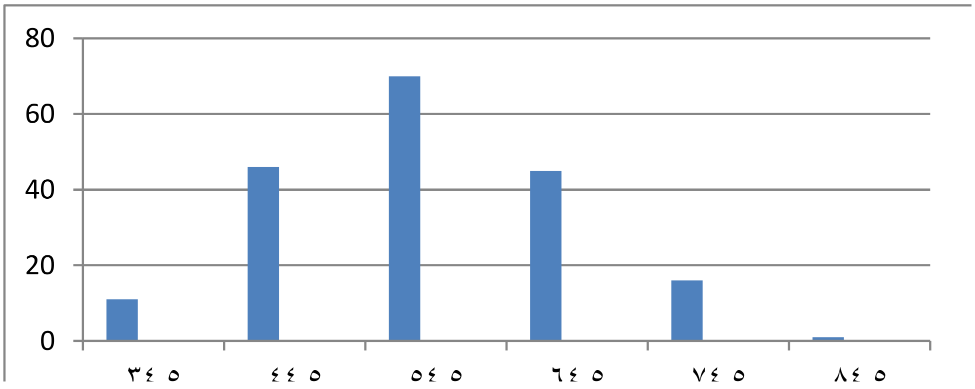


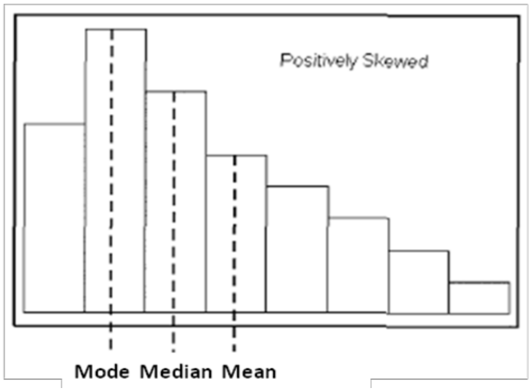
Children Weight

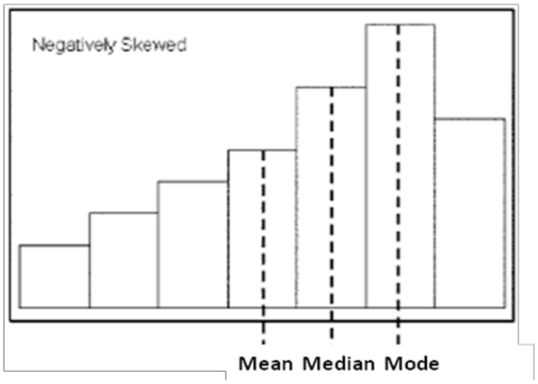
Example; age distribution as the following; Draw a histogram & frequency polygon:

C.I	Frequency
29-39	11
39-49	46
49-59	70
59-69	45
69-79	16
79-89	1

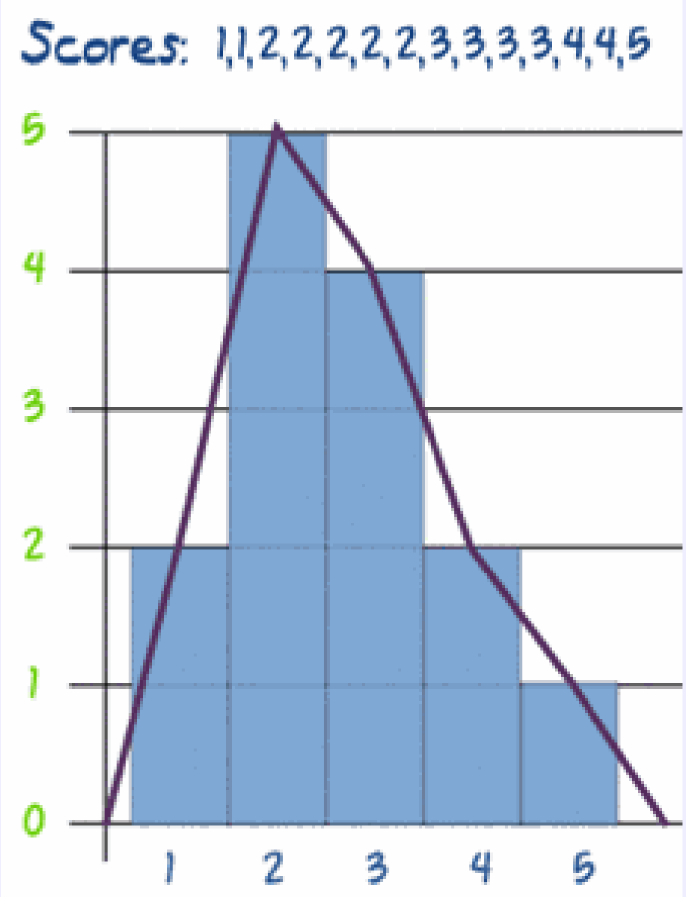
1. Find mid class interval
2. Draw histogram



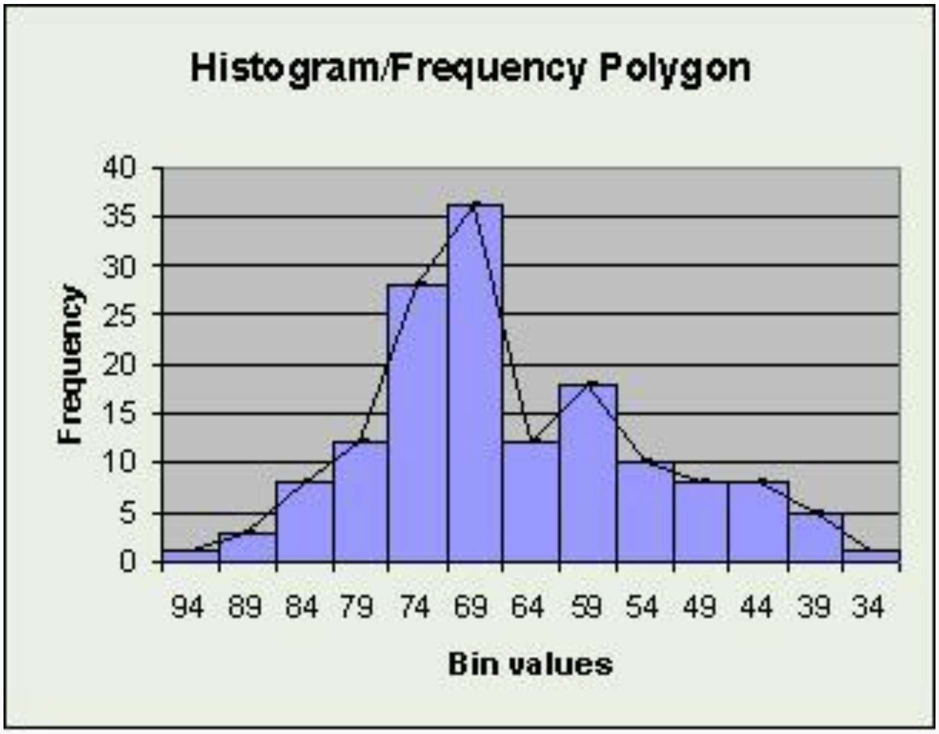




Frequency Polygon



A graph made by joining the middle-top points of the columns of a frequency histogram



This is a histogram with an overlaid frequency polygon.

Mid points of the interval of corresponding rectangle in a histogram are joined together by straight lines. It gives a polygon i.e. a figure with many angles. it is **used when two or more sets of data are to be illustrated on the same diagram such as death rates in smokers and non smokers, birth and death rates of a population. To form a frequency polygon is to connect the midpoints at the top of the bars of a histogram with line segments (or a smooth curve).** Sometimes it is beneficial to show the histogram and frequency polygon together.

Unlike histograms, frequency polygons can be superimposed so as to compare several frequency distributions.

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.
To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. that should include one class interval below the lowest value in data and one above the highest value. The graph will then touch the X-axis on both sides.

.

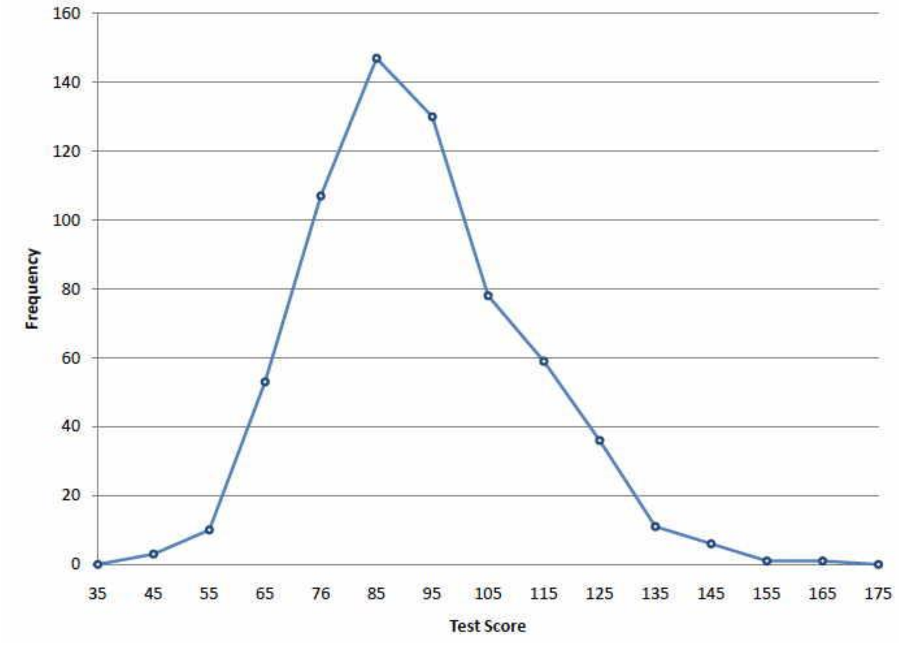


Figure 1. Frequency polygon for the psychology test scores.

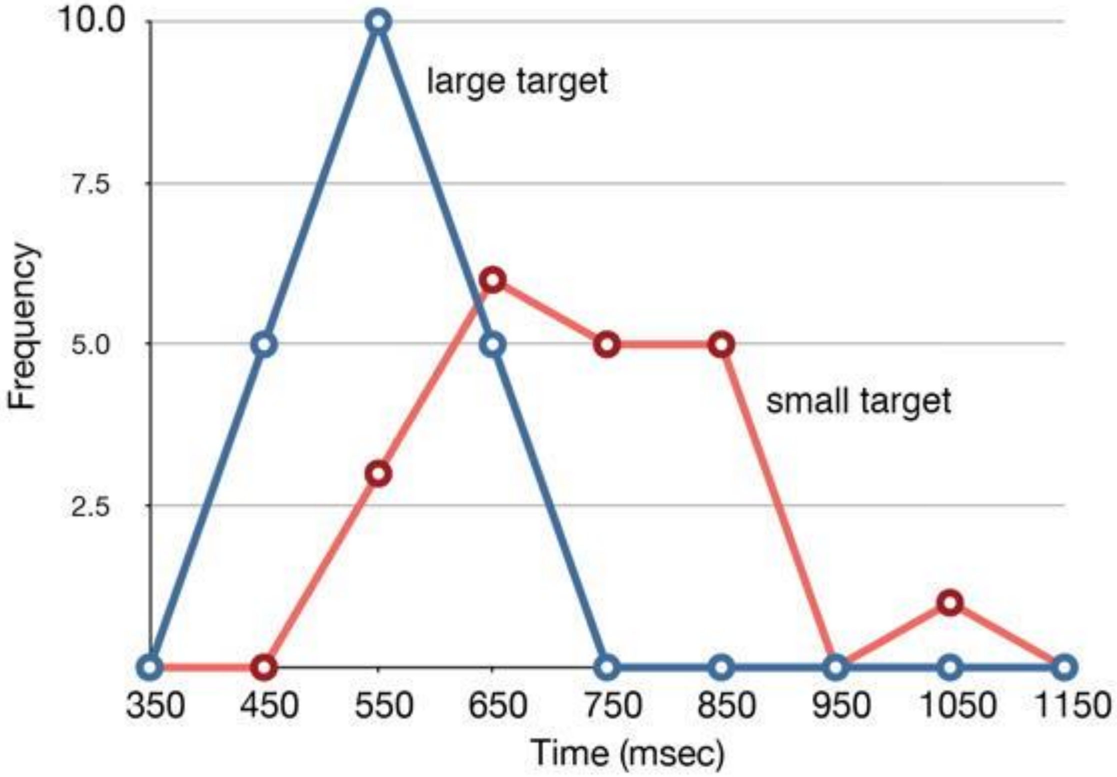
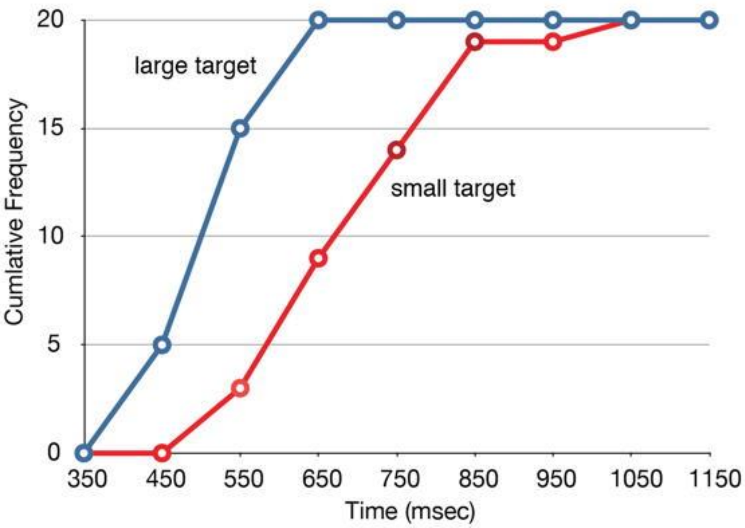
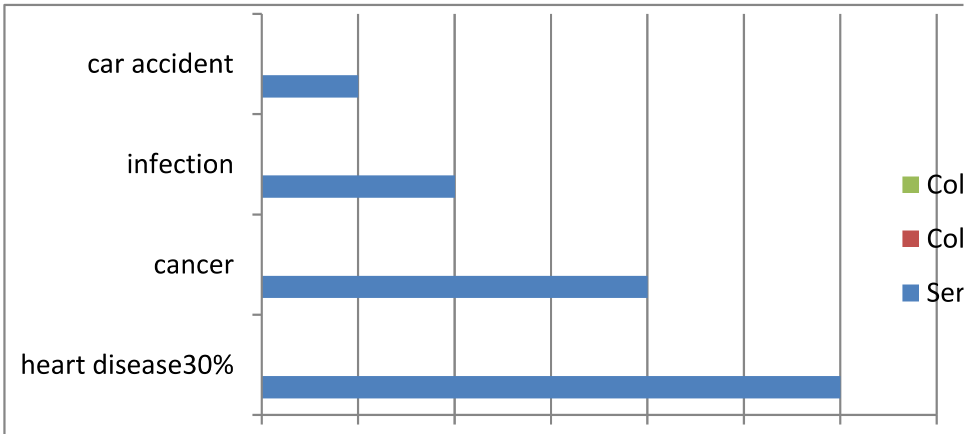


Figure 2. Overlaid frequency polygons.

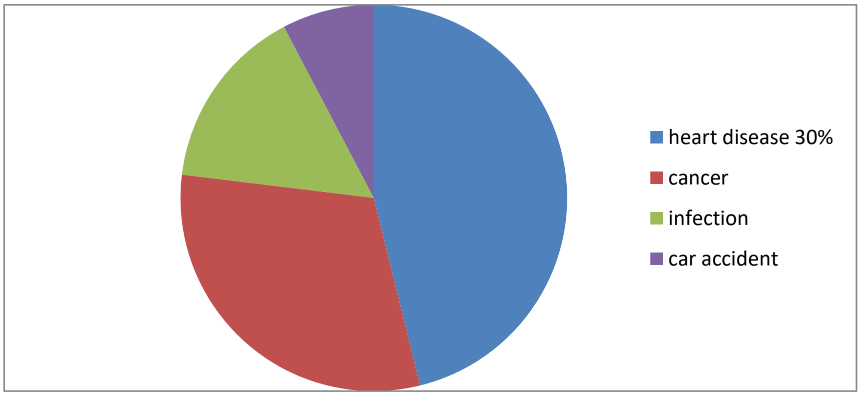
It is also possible to plot two cumulative frequency distributions in the same distributions for the two targets is again evident.



cumulative frequency polygons.



Bar chart



Pie chart

Probability; Health & Medicine are inexact rather than exact sciences &Described as being probabilistic. **-** It is the relative frequency of an incidence of an event in relation to the total events

Types of probability;

1.Continuous Probability

2.Discrete Probability

Example ; Continuous Probability; Incidence Rate of Tuberculosis in Baghdad is 18%,what is the probability of individuals fall in a frequency

Discrete Probability;

Example;

Pregnant women (boy or girl) Probability = one/event = 1/ 2 = 50%

Multiplication Rule; have 2 independent events (A,B) & the event (A)has no effect on B event , the probability is------ **Multiplication**

Example; T.B meningitis, had a case fatality 20%, the probability of 2 randomly selected patients is to be find

P(A&B)= P(A) × P(B) = (20% / 100) × (20%/100)=0 .04

Additional Rule ; when (A) event occurs or (B) event occur P(A)or(B) =P(A)+P(B)

Example; T.B meningitis , had a case fatality 20%,the probability of one of 2 randomly selected patients is to be find

P(A or B) = p(A)+ p(B) = (20 % / 100) + (2o %/100) = 0.4

Mutually exclusive; P of (A) or (B) or both ---------- P(A)+P(B) - P(A&B)

Example; T.B meningitis, case fatality 20%, what is the probability that this disease be fatal at least one of (2) or both randomly patients?

P=P(A)+P(B)- P(A)&P(B)= 0.2+0.2-(0.2X0.2)= O.36

Lecture 6

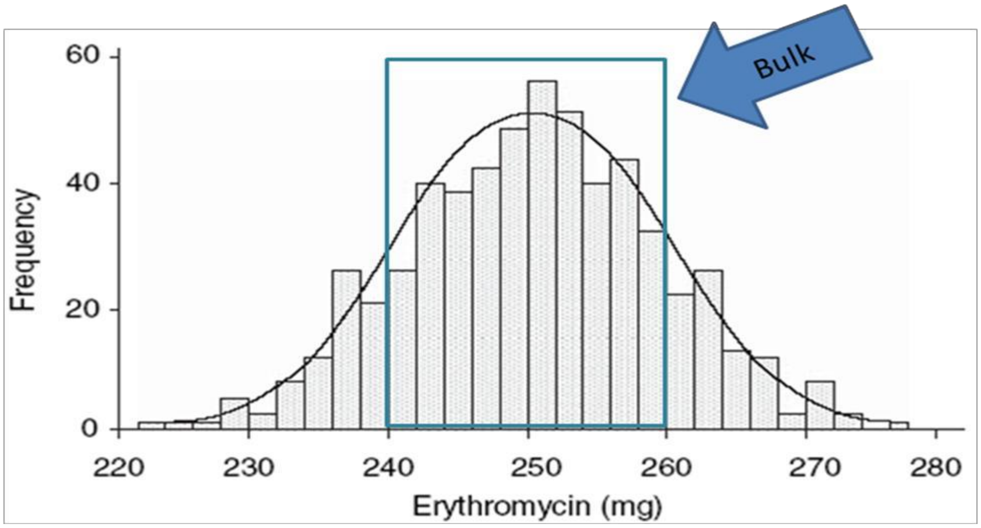
Normal distribution:

The standard deviation is a way statisticians use to measure the amount of variability (or spread) among the numbers in a data set.

It is a standard (or typical) amount of deviation (or distance) from the average (or mean, as statisticians like to call it).

It is also used to describe where most of the data should fall, in a relative sense, compared to the average.

The following shows a histogram of erythromycin content of 500 tablets from an Alpha tablet machine



1. Unimodal data

2. symmetrical distribution

3. frequencies distribute & move toward higher or lower values

Normal (Gaussian) distribution

□ **The most common type of data distribution is called the *bell-shaped curve, in which most of* the data are centered around the average**

Properties of normal distribution

1. The shape of the curve is symmetric.

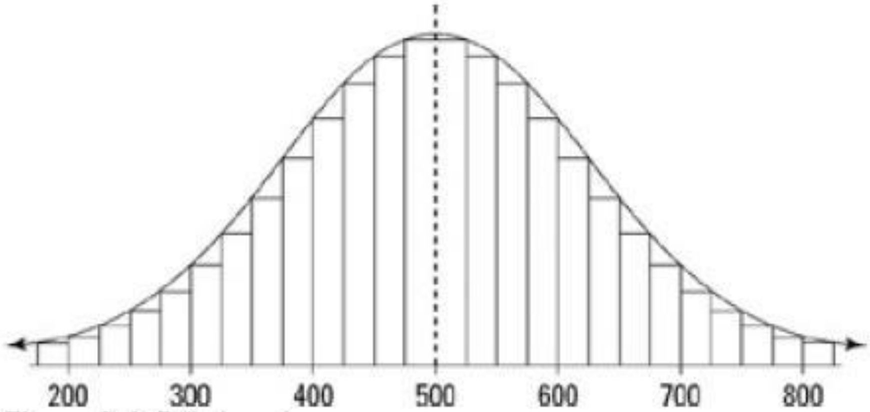
2. It has a bump (swelling) in the middle, with tails going off to the left and right.

3. The mean is directly in the middle of the distribution. The mean of the population is designated by the Greek letter μ (Mu).

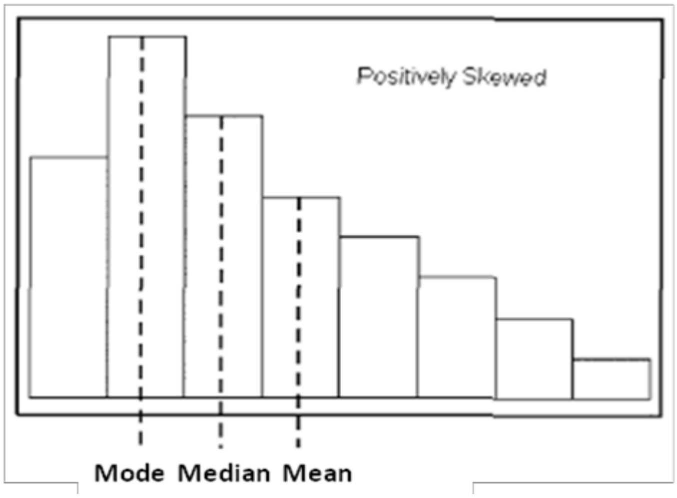
4. The mean and the median are the same value, due to symmetry.

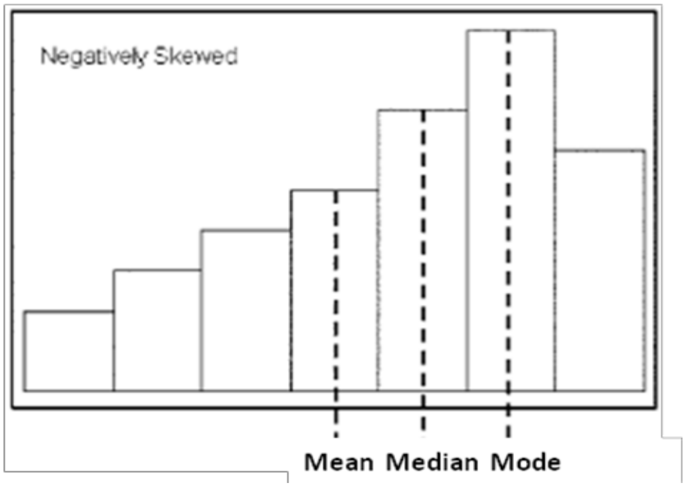
5. The standard deviation represents a typical (almost average) distance between the mean and all of the data.

6. The standard deviation of the population is designated by the Greek letter σ (sigma).
Examples: Blood pressure, body temperature, Hb. Level, height…etc



Skewed distribution





Importance of normal distribution :

-To expect the location of most of the data in relation to others (means & standard deviations)

-Form the basis of significance testing.

Standard normal distribution;

1. The standard normal distribution is a special normal distribution with a mean equal to 0 and a standard deviation equal to 1.

2. The standard normal distribution is useful for examining the data and determining statistics like percentiles, or the percentage of the data falling between two values.

3.* Any normal random variable can be converted to a standard normal random variable by computing the corresponding *z score (standard score).

The z score is computed by the following formula:

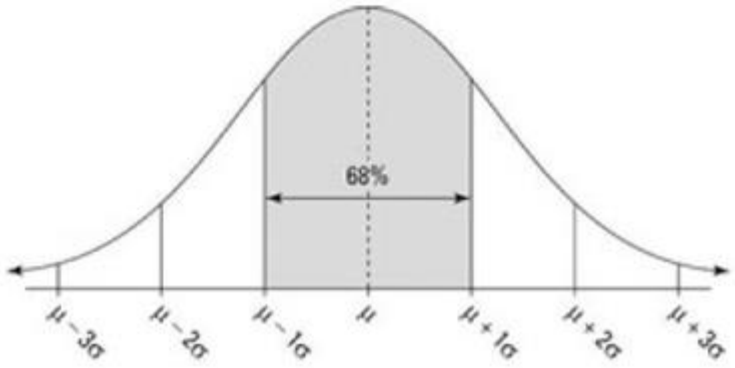
$$z = \frac{X - \mu}{\sigma}$$

z= standard score±

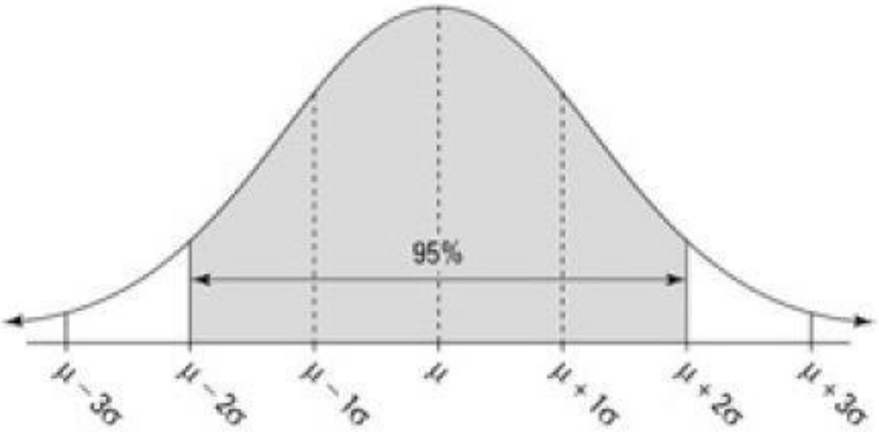
= original score μ= population mean σ= population standard deviation

Z score rule :

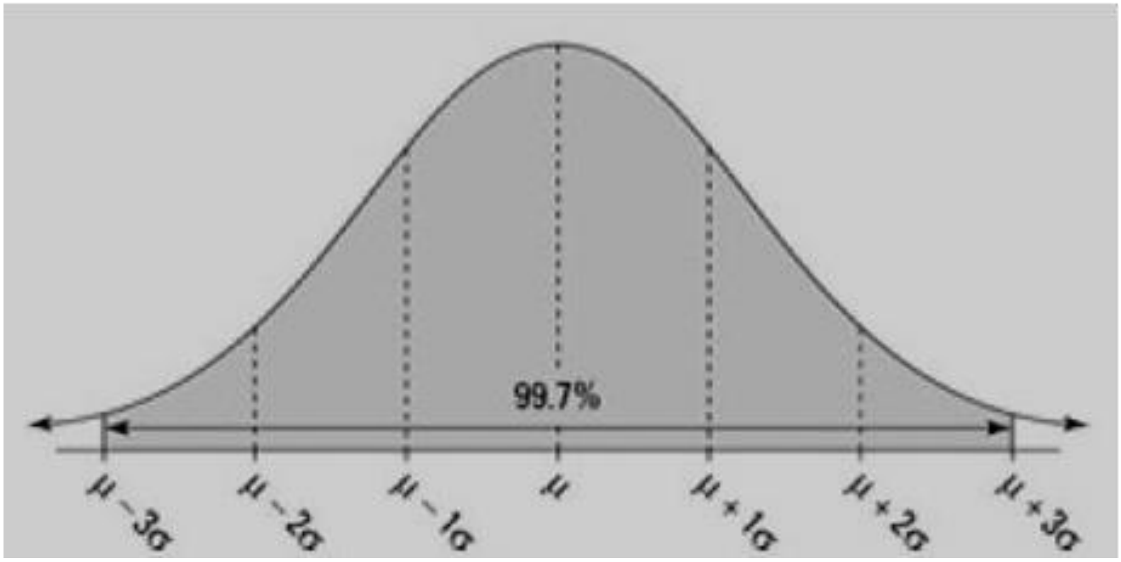
- (1) sigma rule: 68% (68.2%) of observations lie between minus & plus one time SD. (-1Z & +1Z) ±1.



- (2) sigma rule: 95% (95.4%)of observations lie between minus 2 & plus 2 SD units ±2.



- (3) 99 (99.8%) of observations lie between minus 3 & plus 3 SD units **±3.**

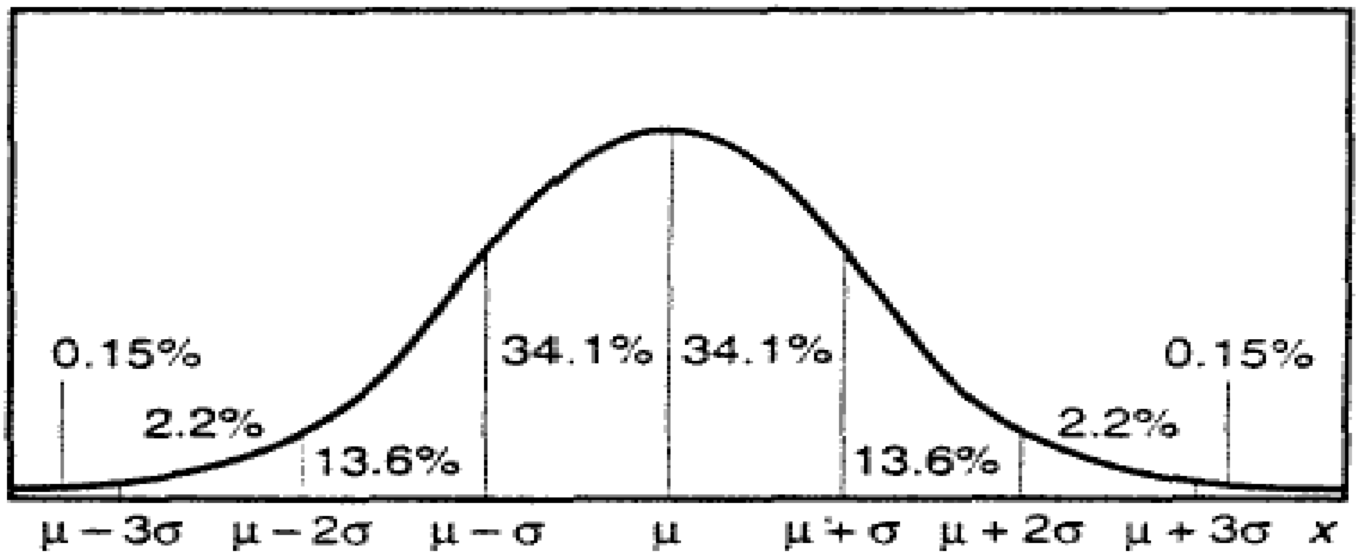


Empirical rules:

□ **(1)** sigma rule: 68% (68.2%) of observations lie between minus & plus one SD. (-1Z & +1Z).

□ **(2)** sigma rule: 95% (95.4%) of observations lie between minus 2 & plus 2 SD units

□ **(3)** sigma rule: 99 (99.8%) of observations lie between minus 3 & plus 3 SD units.



Z-score: Z score is a measure of the distance that a particular population member is from the population's mean. It is so named because it is frequently used on populations that have a normal distribution, which is also known as a Z distribution.

A z-score is also known as a standard score because it is measured in units of standard deviation, which allows observations from different distributions to be compared.

In statistics, a z-score (or standard score) is used to compare means from Different normally distributed sets of data. The actual score indicates how many standard deviations an observation is above or below the mean.

The z- score is useful in research utilizing statistical analysis because it allows for the comparison of observations from different normal distributions. In effect, when items from different data sets are transformed into z-scores, then they may then all be compared. This article will show you how to calculate a z-score (or standard score).

Example

□ Assuming systolic blood pressure (BP) in normal healthy individuals is normally distributed with **μ = 120** and **σ**= 10 mm Hg,

□ *What area of the curve is above 130 mm Hg?*

□ *What area of the curve is between 100 and 140 mm Hg?*

Answer

# z = (130 - 120) / 10 = 1.00, and the area above 1.00 is 0.159. So 15.9% of

normal healthy individuals have a systolic blood pressure above 1 standard deviation (>I30 mm Hg).

# z = (100 - 120) / 10 = -2.00. and= (140 - 120)/10 = 2.00; the area between -2 and +2 is 0.951.

So 95.4% have a systolic blood pressure between -2 and +2 -standard deviations (between 100 and 140 mm Hg).

Benefit of Sigma rule (Mean±SD):

Normal ranges are frequently based on the mean ±2SDs.

Low SD, the data are all clustered tightly around the mean and the distribution is tall and thin.

Higher SD, the data are more scattered, the distribution is low and wide.

Calculate 95% Confidence interval.

95% Confidence interval:

The mean derived from the sample is the best available estimate of the population mean and is referred to as the point estimate.

A mean derived from a sample is unlikely to be a perfect estimate of the population mean.

A range within which we are reasonably confident the true populations mean lies, it is called 95% CI.

□ Standard deviation tells us about the variability (spread) in a sample.

□ The CI tells us the range in which the true value (the mean if the sample were infinitely large) is likely to be.

□ Greater SD gives wider intervals.

□ Greater sample sizes give narrower intervals.

Table of probabilities related to multiples of SDs

Number of SD	**Probability of observation showing at least as large a deviation from the population mean**
0.674	0.50
1.0	0.317
1.645	0.10
1.960	0.05
2.0	0.046
2.576	0.01

Confidence limits:

□ 68% of observations lie between minus & plus one SD. (-1Z & +1Z).

□ 95% of observations lie between minus 1.96 & plus 1.96 SD units.

□ 99% of observations lie between minus 2.58 & plus 2.58 SD units.

□ 99.7% of observations lie between minus 3 & plus 3 SD units.

Lecture 7

Sampling

Sampling is the process of selecting units (people, organizations) from a population of interest so that by studying the sample we may generalize our results back to the population from which they were chosen. The sampling process comprises several stages: defining the population of concern.

- A shortcut method for investigating a whole population
- Data is gathered on a small part of the whole parent population or sampling frame, and used to inform what the whole picture is like

Specifying a sampling frame, a set of items or events possible to measure specifying sampling method for selecting items or events from the frame determining the sample size & implementing the sampling plan and data collecting.

TYPES;

PROBABILITY SAMPLING ;

a. quantitative data

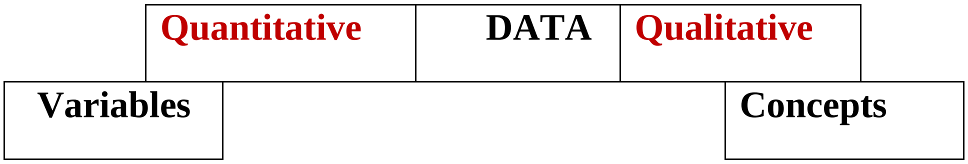
b. selection on random way & give chance to every person to participate in the study. Results can be generalized to total population

NON PROBABILITY SAMPLING;

a. qualitative data

b. selection on non random way

c. results can not be generalized to total population



Quantitative

DATA

Qualitative

Variables

Concepts

1. Simple . R .

2. Systematic = =

3. Multi stages =

4. Stratified = =

5. Cluster = =

1. Accidental (convenience)

2. Purposive S.

3. Quota S.

4. Snow ball S.

5.Volunteer s.

1. Probability sampling ; (Random sampling);

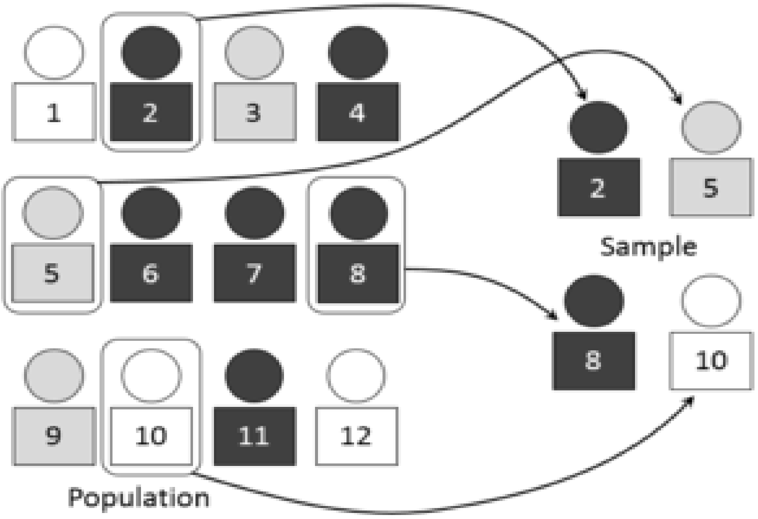
a. Simple random S. ;

n a simple random sample (SRS) of a given size, all such subsets of the frame are given an equal probability. Furthermore, any given pair of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimizes bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

- Lottery method.

- Random numbers of table.

- Computer method



- **Objective**: To select *n* units out of *N* such that each has an equal chance of being selected.

- **Procedure**: Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.

b. Systematic random S ;

from the target population the sample is selected in a systematic way by using sampling fractions; N = target population n = sample size K = sampling fraction

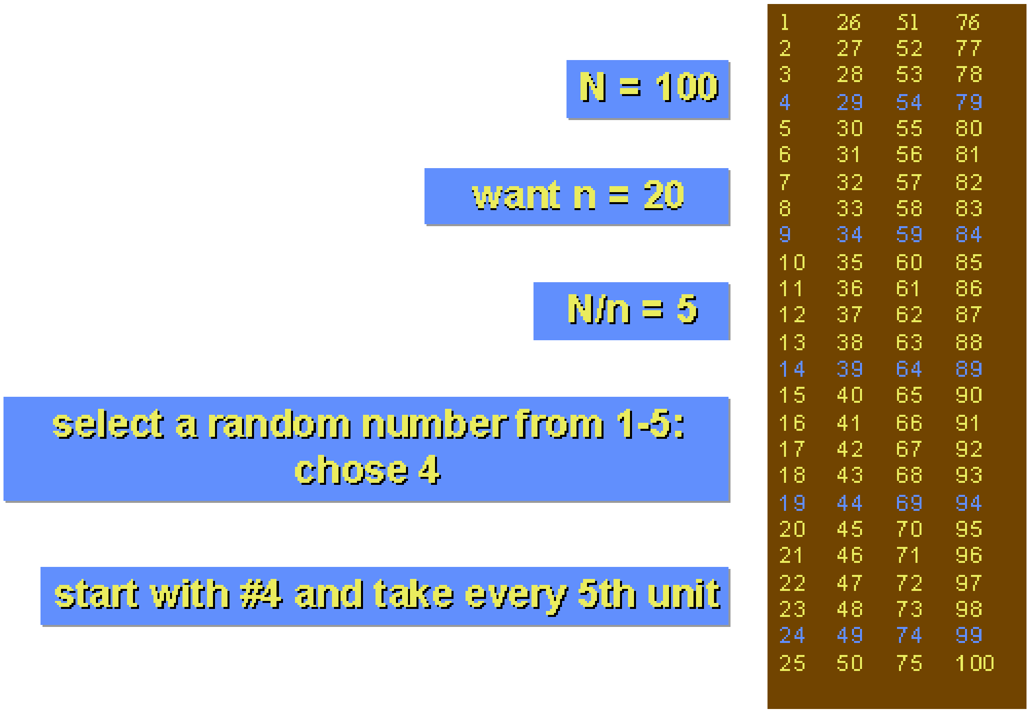
Ex; N= 54,000 (population frame) n = 6,000 (sample size)

K= N/n = 54,000/ 6,000 = 9; Can pick any person& if pick number (1) the systematic sampling will be as the following ;

1,2,3,4,5,6,7,8,(**9**),10,11,12,13,14,15,16,17,(**18**),19,20,21,22,23,24,25,26,(**27**)till reach sample size

Here are the steps need to follow in order to achieve a **systematic random sample**:

- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- k = N/n = the interval size
- randomly select an integer between 1 to k
- then take every kth unit



N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

Example: interval size, k, is equal to N/n = 100/20 = 5. Now, select a random integer from 1 to 5. In our example, imagine that you chose 4. Now, to select the sample, start with the 4th unit in the list and take every k-th unit (every 5th, because k=5). You would be sampling units 4, 9, 14, 19, and so on to 100 and you would wind up with 20 units in your sample.

Advantages:

- It is more straight-forward than random sampling
- Sampling just has to be at uniform intervals
- A good coverage of the study area can be more easily achieved than using random sampling

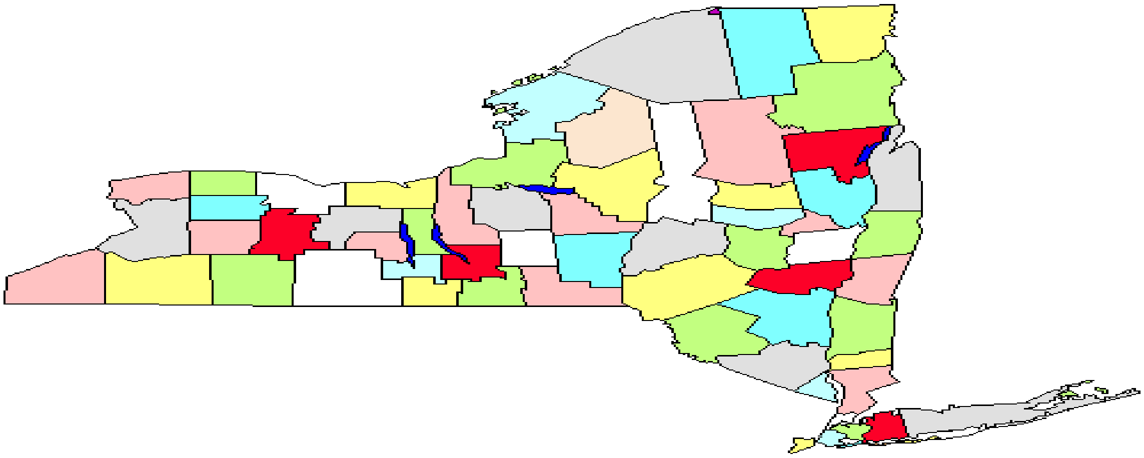
Disadvantages:

- It is more biased, as not all members or points have an equal chance of being selected
- It may therefore lead to over or under representation of a particular pattern

Cluster (Area) Random Sampling:

select by using simple random S or systematic the clusters or groups of elements such as classes , schools ,districts, streets, places ,houses and each cluster can select a subsample by using simple random S or systematic .

The problem with random sampling methods when we have to sample a population that's disbursed across a wide geographic region is that you will have to cover a lot of ground geographically in order to get to each of the units sampled.



Cluster sampling, we follow these steps:

- Divide population into clusters (usually along geographic boundaries)
- Randomly sample clusters
- Measure all units within sampled clusters.

Multi-Stage Sampling;

This method involves drawing samples . draw the sample from target population , continue drawing till reach required sample by using simple R.S or systematic R.S that get the primary sample then continue to draw another sample until get the required sample .

Stage 1 = 1000 from 10,000

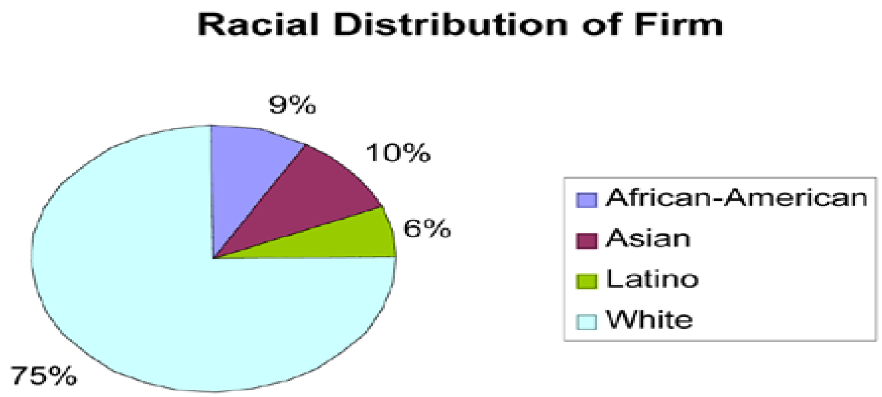
Stage 2 = 500 from 1000 **Stage 3 = 100 from 500**

The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

For example, consider the idea of sampling New York State residents for face-to-face interviews. Clearly we would want to do some type of cluster sampling as the first stage of the process. We might sample townships or census tracts throughout the state. But in cluster sampling we would then go on to measure everyone in the clusters we select. Even if we are sampling census tracts we may not be able to measure *everyone* who is in the census tract. So, we might set up a stratified sampling process within the clusters. In this case, we would have a two-stage sampling process with stratified samples within cluster samples.   Within selected districts, we might do a simple random sample of schools. Within schools, we might do a simple random sample of classes or grades. And, within classes, we might even do a simple random sample of students. In this case, we have three or four stages in the sampling process and we use both stratified and simple random sampling. By combining different sampling methods we are able to achieve a rich variety of probabilistic sampling methods that can be used in a wide range of social research contexts.

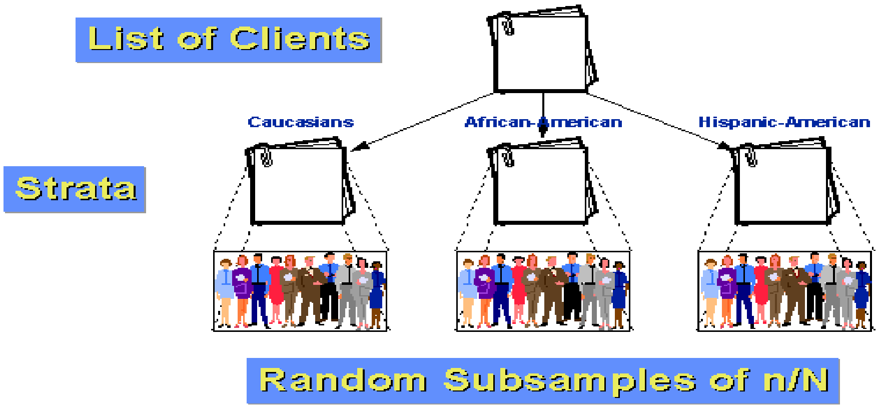
c.Stratified sampling ;

Divide the population into strata as; age, sex, social class, marital status & pick every strata by using Simple R.S or Systematic R.S, also sometimes called *proportional* or *quota* random sampling, involves dividing the population into homogeneous subgroups and then taking a simple random sample in each subgroup. In more formal terms:



The population into non-overlapping groups (i.e., *strata*) N_1, N_2, N_3, ... N_i, such that $N_1 + N_2 + N_3 + ... + N_i = N$. Then do a simple random sample of $f = n/N$ in each strata.

There are several major reasons why might prefer stratified sampling over simple random sampling. First, it assures that will be able to represent not only the overall population, but also key subgroups of the population, especially small minority groups. If want to be able to talk about subgroups, this may be the only way to effectively assure will be able to. If the subgroup is extremely small, can use different sampling fractions (f) within the different strata to randomly over-sample the small group.



Second, stratified random sampling will generally have more statistical precision than simple random sampling. This will only be true if the strata or groups are homogeneous. If they are, we expect that the variability within-groups is lower than the variability for the population as a whole. For example, let's say that the population of clients for our agency can be divided into three groups: Caucasian, African-American and Hispanic-American. Furthermore, let's assume that both the African-Americans and Hispanic-Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of n=100 with a sampling fraction of 10%, we would expect by chance alone that we would only get 10 and 5 persons from each of our two smaller groups .

Advantages:

- It can be used with random or systematic sampling, and with point, line or area techniques

-If the proportions of the sub-sets are known, it can generate results which are more representative of the whole population

-It is very flexible and applicable to many geographical enquiries

-Correlations and comparisons can be made between sub-sets

Disadvantages:

-The proportions of the sub-sets must be known and accurate if it is to work properly

- It can be hard to stratify questionnaire data collection, accurate up to date population data may not be available and it may be hard to identify people's age or social background effectively

2. <u>Non probability sampling</u>;

a. Accidental sampling (convenience) ; called haphazard , that the researcher interviews the respondents for the study who comes in contact accidentally during the research time .

b. Purposive sampling ;the researcher chooses respondents who in the opinion of the researcher thought to be relevant to the subject under the study

d. snow ball ;the researcher begins with few of respondents who are available then ask the respondent to recommend other persons who meets the criteria of the research & who are willing to participate in the study .

e. Volunteer sampling ; base on the acceptance .

Confidence Interval & Confidence Level ; The confidence interval is the number of percentage points above or below the proportion that find in the study that the true proportion should be within. For example, if confidence interval is 3.5 % and study reveals a proportion of 57 %, the true proportion is likely between 53.5 %and 60.5%

The general expression for confidence interval is:

Confidence interval = point estimate ± (confidence multiplier x SE)

For 95%, 99% confidence level the C. multiplier is 1.96, 2.58 respectively.

Confidence level: Is the probability value attached to a given confidence interval. It can be expressed as a percentage (95%, 99%).

Factors that Affect Confidence Intervals;

There are three factors that determine the size of the confidence interval for a given confidence level:

- Sample size, population size &Percentage

Sample Size ;The larger sample size, the more sure can be that their answers truly reflect the population. This indicates that for a given confidence level, the larger sample size, the smaller confidence interval.

Percentage ; Accuracy also depends on the percentage of sample that picks a particular answer. If 99% of sample said "Yes" and 1% said "No," the chances of error are remote, irrespective of sample size. However, if the percentages are 51% and 49% the chances of error are much greater.

- Population Size; how many people are there in the group sample represents? This may be the number of people in a city

- Using 10% of Prevalence of certain disease or problem.

The sample size; The appropriate sample size for a population is determined by:
(1) The estimated prevalence of the variable of interest, when prevalence is high in community can choose small sample size.
(2) The desired level of confidence and the acceptable margin of error.
(3) Design of study; for cohort & clinical trial need large sample size while in case control can choose small sample size.

$$n = \frac{\text{CF of 95\%(1.96) X P(1-P)}}{\text{m(standard value of 0.05)}^2}$$

n= required sample size
CF = critical factor at level 95% (standard value of 1.96)
P = estimated prevalence of disease in the project area
m = margin of error at 5% (standard value of 0.05)

Example: In a survey has been estimated that roughly 30% (0.3) of the children suffer from chronic malnutrition. This figure has been taken from national statistic s on malnutrition in rural areas. Use of the standard values listed above to find sample size;

$$N = \frac{(1.96) \text{ X } 0.3(1- 0.3)}{(0.05)^2}$$

Lecture 8 & 9 :

Tests of statistical significance

Statistical M & Parameter M; mean of sample estimates mean of population & SD of sample estimates of parameters that called point estimation (limited)

They are standard statistical procedures for drawing inferences from sample estimates about population parameters. Tests of significance allow us to decide whether the sample estimates or the differences between estimates are within their normal biological variation (called variability or difference due to chance).

Chance variation can give rise to differences between samples studied, and every time a difference is observed, a question arises: is it statistically significant??? Or due to chance??

Procedure (steps) for statistical tests:

1- **State the null hypothesis**:

- Hypothesis may be defined simply as a statement about one or more population.
- Statistical hypothesis is that which is stated in such a way that can be evaluated by appropriate statistical techniques or tests.
- **Null hypothesis (H_0)** is the hypothesis to be tested (null = no difference), it states that there is no statistical or real difference between the sample mean and the population mean, and if there is any difference it is due to chance

  - Alternate hypothesis (H_A): the opposite of null hypothesis.

  it states that there is statistical or real difference between the sample mean and the population mean, and it is not due to chance.

  α error: reject null hypothesis even it is true

  β error: accept null hypothesis even it is false

2**. State the level of significance** (**α)** = 0.05, or 0.01, 0.1 level.

$P< 0.05$ means if the probability of finding a difference by chance (due to sampling process) is less than 5%. i.e the difference must be significant.

The level of significant that at area of rejection will reject true null hypothesis H_0

Then we select small value of ((⟨∝ ⟩)) in order to make the probability of rejection of true H0 is small. The most counted values of⟨∝⟩ are;1% , 5 % ,10% ,20% , mostly use 5% ---- o.o5

⟨∝⟩ - level of significance	confidence level 100 - ∝	critical factor
5 %	95%	1.96
1%	99%	2.58
10%	90%	1.64

3. **Choose the test statistic**

Statistical inference and confidence intervals

To study the properties of some population, we often need to draw a sample from that population. This subgroup of individuals in the population is selected to be, to some degree, representative of that population.
For each sample, we can calculate the mean (x^-) and standard deviation (SD).

Statistic inference: Statistic inference is the process of drawing conclusions regarding a population based upon studying only a portion of the members of that population. The process of inference (generalization) is most accurate when the subpopulation being studied is homogenous, truly representative and chosen randomly to eliminate sampling error.

Thus by inference:

1. We can estimate the population mean through the sample mean

1- Test whether a sample means come from or belong to that population.

Confidence Intervals (CI):

The mean of a sample is only a **“point estimate”** of the mean for the entire population. Although this sample mean may truly reflect the population mean, there is uncertainty in this value.

Confidence intervals are constructs used to describe the range of values possible for this estimate.

The 95%confidence interval represents 95% confidence that the lower and upper limits of this interval include the true mean of the population.

Definition of confidence interval:

CI is a range of likely values defined by upper and lower end points within which the true value of an unknown population parameter is likely to fall.

The general expression for confidence levels is:

Confidence level= point estimate ± (confidence multiplier(CF)xSE)

For 90%, 95%, 99% confidence levels, the multiplier is 1.64, 1.96, 2.58 respectively.

Confidence limits:

Are the lower and upper boundaries of a confidence level.

Confidence level:

Is the probability value attached to a given confidence level. It can be expressed as a percentage (90%, 95%, 99%) or a number (0.90, 0.95, 0.99).

Standard error of the mean (SEM):

A measurement of the spread of sample data means. It is calculated as the standard deviation of a population of sample means divided by the square root of the sample size. It is one of measures of variability and it is inversely related to the sample size. The larger the sample size, the less the standard error (the variability)

Steps of hypothesis:

1. **Assumption of normal distribution, T, X^2 distribution**
2. **Data**
3. **Set the Hypothesis**
4. **Level of Significance**
5. **Formula**
6. **Conclusion**

Test statistic (tests of significance):

The test statistic is some statistic that may be computed from the data of the sample. It tests the conflict between what has been assumed (by the null hypothesis) and what is found.

There are many types of tests for different types of data:

1- Z test
2- t-test
3- Chi squared test (X^2 test)

Z test: is applied for both quantitative and qualitative normally distributed data.

1-For quantitative data:

a. To test whether a sample is drawn from/ or belong to a population, when there is a big sample size ($n > 30$)

The test statistic will be:

$$Z = \bar{X} - \mu / SE(\bar{X})$$

$$Z = \bar{X} - \mu / \sigma / \sqrt{n}$$

b . To test the difference between two means of two samples when ($n_1 + n_2 > 30$) the test statistic will be:

$$Z = \bar{X}_1 - \bar{X}_2 / \surd\, v_1/n_1 + v_2/n_2$$

2- For qualitative data:

a. **to test whether a sample is drawn from/ or belong to a population (for proportion):**

$Z = (\bar{P} - P) / \sqrt{P(1-P)} / \sqrt{n}$ **Or = $Z = (\bar{P} - p) / \frac{\sqrt{pq}}{\sqrt{n}}$**

$\bar{P}$ = proportion of sample $q = 1-P$

P = proportion of population

b. **to test whether 2 samples has different proportion**

$p = (r_1+r_2) / (n_1+n_2)$

$SE(p_1-p_2) = \sqrt{\{p(1-p) \times (1/n_1 + 1/n_2)\}} = \sqrt{\{pq \times (1/n_1 + 1/n_2)\}}$

$Z = (p_1 - p_2) / SE(p_1-p_2)$

OR :

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - (p_1 - p_2)}{\sqrt{P_1 x q_1}/\sqrt{n_1} + \sqrt{P_2 x q_2}/(\sqrt{n_2})}$$

Example 1:

The mean systolic blood pressure of 130 men of age 50-65 years whom are put on special diet for two years is 145.5 mm hg. If the mean systolic blood pressure of the normal population is 125.0 mm hg and a standard deviation is 20 mm hg, how reasonable is it to conclude that the systolic blood pressure of that sample is not different from that of the population?

Solution:

1. Assumption of normal distribution

2. Data: X” = 145.5 $\mu = 125$ $\sigma = 20$ n = 130

3. Hypothesis:
(H_0): there is a significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population mean
(H_A): there is a significant difference between the mean systolic blood pressure of 130 men and the mean systolic blood pressure of the population

4. Level of significance: 0.05

4- Statistic test: Z test = **$Z= X”- \mu / \sigma /\sqrt{n}$**

5- $Z= (145.5 – 125) / (20/\sqrt{130})$ $Z= 11.7$

6- Conclusion: The calculated Z is > 1.96 at 0.05 level So P< 0.05

Thus there is a significant difference between the mean of systolic blood pressure of 130 men and the mean systolic blood pressure of the population at 0.05 level.

We reject Null hypothesis, and accept the alternate hypothesis at 0.05 level.

- At 0.01 level, the calculated Z (11.7) is > **2.58**

Example 2:

A survey a total of 88 households used a river for water supply, 49 of them had episodes of diarrhea against 10 from 36 households using the well water. Is there a statistically significant difference in the proportions with episodes of diarrhea between the households using river and well water supply?

Solution:

1. Assumption of normal distribution

2. Data: $r_1= 49$, $n_1= 88$ $r_2= 10$, $n_2= 36$

$P_1= 49/88= 0.556$

$P_2= 10/36= 0.227$

$P= (49 +10) / (88 +36) = 0.476$

3. set the hypothesis;

H_0: There is no difference in the proportions with episodes of diarrhea

H_A: there is a real difference in the proportions with episodes of diarrhea

4. level of significance: 0.05

5. test statistic: Z test for proportion is chosen

$p= (r_1+r_2) / (n_1+n_2) = 0.476$

$SE (p_1-p_2) = \sqrt{ \{p (1-p) \times (1/n_1 + 1/n_2)\} }$

$Z= (p_1- p_2) /SE (p_1-p_2)$

$SE (p_1-p_2) = \sqrt{ 0.476 (1- 0.476) \times (1/88 + 1/36)}$

$= 0.0988$

$Z= (0.556- 0.227) / 0.0988$

$Z= 3.32$

6- Conclusion:

The calculated Z is 3.32 > 1.96 at 0.05 level
So P< 0.1

Thus there is a significant difference in the proportions with episodes of diarrhea between the households using river and well water supply, and the difference is not due chance at 0.1 level.

We reject Null hypothesis, and accept the alternate hypothesis at 0.05 level.

At 0.01 level, the calculated Z (3.32) is > **2.58** So p< 0.01, the difference is highly significant. We reject Null hypothesis, and accept the alternate hypothesis at 0.01 level.

Example; A nutritional survey found the obesity was 31% in population, what is the probability that the proportion in the sample of 150 women who are obese to be > 40% - α = 0.05
where P" = proportion of sample p =proportion of population

1. Assumption of normal distribution
2. Data: P=31%, P" > 40%

3. Null hypothesis: there is no significant difference in the proportions of women who are obese in the population & sample.

Alternate hypothesis: there is a significant difference in the proportions of women who are obese in the population & sample.

4. Level of significance: 0.05

5. Test statistic: Z test for proportion is chosen

$* \ Z = (P''- P) / \sqrt{((Pq) / n)}$

6- Conclusion:

* $Z = (P'' - P) / \sqrt{((Pq) / n)}$

Where $\sqrt{((Pq)/(n)} = \sqrt{((0.31 \times 0.69)) / (150)} = 0.001426$

$$= \frac{0.40 - 0.31}{0.001426} = 2.38$$

finding; 2.38 > 1.96 occur in rejection area that **$P < 0.04$**

Example; A report of lower backache in Egypt was 28%, other report population found 21% had Backache.

Find the probability of random sample of 100 will have value $(P_1'' - P_2'') > 10\%$

$\alpha = 0.05$

Solution;

1. Assumption of normal distribution

2. Data: - $(p_1 = 28\%,\ p_2 = 21\%)$, $(P_1'' - P_2'') > 10\%$

3. H_O: $(P_1'' - P_2'') > 10\%$ ----------- H_A: $(P_1'' - P_2'') \leq 10\%$

4. Level of significance $\alpha = 0.05$

5. Test: $$Z = \frac{(P_1'' - P_2'') \quad - \quad (p_1 - p_2)}{\sqrt{P_1 q_1} / \sqrt{n_1} \ + \ \sqrt{P_2 q_2} / (\sqrt{n})}$$

$(p_1 - p_2) = \ 0.28 - 0.21 = 0.07$ (proportion of 2 populations)

$q = (1 - p)$

$$= \frac{(\ 0.10\ ) \quad - \quad (\ 0.07\ )}{\sqrt{0.28 \times 0.72} / \sqrt{100} + \sqrt{0.21 \times 0.79} / (\sqrt{100})} = \frac{0.03}{\sqrt{0.003675}} = 0.49$$

Conclusion: 0.49 < 1.96 ------- accept H_0 $(P_1'' - P_2'') > 10\%$

THERE IS NO REAL DIFFERENCE BETWEEN 2 PROPORTIONS

CONFIDENCE LEVEL FOR (Z) TEST ;TO ESTIMATE THE POPULATION PARAMETER:

1. FOR QUANTITATIVE DATA:

A. FOR ONE MEAM :

$X'' \pm P\ (C.F)\ (SD/(\sqrt{(N)}$

THE C.L OF THE MEAN AT 95% C.LEVEL= $X'' \pm 1.96 \times (SD/ (\sqrt{(N)}$

THE C.L OF THE MEAN AT 99% C.LEVEL= $X'' \pm 2.58 \times (SD/ (\sqrt{(N)}$

B. FOR 2 MEANS :

C.L OF 2MEAN AT 95% C. LEVEL = $\bar{X}_1 - \bar{X}_2 \pm 1.96 \times \sqrt{}\ V_1/N_1 + V_2/N_2$

C.L OF 2 MEANS AT 99% C .LEVEL = $\bar{X}_1 - \bar{X}_2 \pm 2.58 \times \sqrt{}\ V_1/N_1 + V_2/N_2$

3.FOR QUALITATIVE DATA:

THE PARAMETERS HERE ARE THE SAMPLE PROPORTION (P”), THE POPULATION PROPORTION (P), AND THE STANDARD ERROR OF THE PROPORTION SE(P) .

SE (P) = $\sqrt{P''Q''}/\sqrt{N}$ Q”= 1-P

- CONFIDENCE LEVELS FOR PROPORTION: (ONE PROPORTION) :

THE PARAMETERS HERE ARE THE SAMPLE PROPORTION ($\bar{P}$), THE POPULATION PROPORTION (P), AND THE STANDARD ERROR OF THE PROPORTION SE(P) .

$\bar{P}$= THE SAMPLE PROPORTION OF SUCCESSFUL OUTCOME

$\bar{P}$= R/N R= THE NUMBER WHO HAVE THE CHARACTERISTIC OF SAMPLE

N= TOTAL NUMBER OF POPULATION (NO. POPULATION)

SE(P) = $\sqrt{P''Q''/N}$ Q”= 1-P

C .L : 95% CI FOR P = $\bar{P} \pm 1.96 \times$ SE(P)= 95% CI FOR P = $\bar{P} \pm 1.96 \times (\sqrt{P''Q''/N})$

99% CI FOR P = $\bar{P} \pm 2.58 \times (\sqrt{P''Q''/N})$

* CONFIDENCE LEVELS FOR (2) PROPORTION:

C .L = $(P_1''- P_2'') \pm P(CF) \times \sqrt{(}\ (P''_1\ Q''_1/N_1+P''_2\ Q''_{2))}\ /N_2$