Module plan

Topic :	Biostatistics
Subject:	Public Health Dentistry
Target Group:	Undergraduate Dentistry
Mode:	Powerpoint
Platform:	Institutional LMS
Presenter:	Dr Sandesh N

Biostatistics

- Dr Sandesh N

Dept of community dentistry

Introduction

Normal BP 120/80 mm Hg

Europeans are taller than Asians

Average male adult weighs 70kgs

Drug A is better than drug B

- Endless

Cannot be arrived by just Raw data

Numbers tell tales Speak the language of STATISTICS Adds meaning to data helps to interpret data

Thus lending significance to the study

Descriptive statistics

Statistic

means a measured or counted fact or a piece of information stated as a figure

Data

Can be defined as a set of values recorded on one or more individuals or observational units

VARIABLE

A general term for any feature of the unit which is observed or measured.

STATISTICS

Is the science of compiling, classifying & tabulating numerical data and expressing the results in a mathematical or graphical form.

OR

Statistics is the study of methods & procedures for collecting, classifying, summarizing & analyzing data & for making scientific inferences from such data.

- Prof P.V.Sukhatme

BIOSTATISTICS

Is the branch of statistics applied to biological or medical sciences (biometry).

OR

- *Is that branch of statistics concerned with mathematical facts and data relating to biological events.*

Basic principles of biostatistics

Collection of data

Presentation of data

Summarization of data

Analysis of data

Interpretation of data

Collection of data

Data

1. *Qualitative*
    1. *No notion of magnitude or size of the characteristics*
    2. *Calculated by counting the individuals and not by measurements*
2. *Quantitative*
    1. *Have an magnitude*
    2. *Measured either in interval or ratio scale*
    3. *Observation ascends or descends from 0 or any starting point*
    4. *Measurable in whole or in fractions*

Data

1. Primary data
2. Secondary data

Collection of primary data

1. *Observation*
2. *Interview*
   1. *Telephonic interview / Personal Interview*
      Direct/ indirect
   2. *Structured / Unstructured*
3. *Questionnaire*
   1. *MCQ*
   2. *Open End Questions*
   3. *Closed End Questions*
4. *Schedule*
5. *Clinical examination*

Collection of Secondary data

Published

Articles, conference reports, newspapers

Unpublished

Dairies, letters, Biographies

Sampling

Target population

Is the group of individuals to whom the investigator wants the conclusion of his study to apply

Sample

Is a part or subset of the target population that takes part in the investigation

Sampling frame

A list containing all sampling units is called sampling frame

Sampling design / sampling technique

Sampling is a definite plan for obtaining sample from the sampling frame or population

1. *Probability sampling*
2. *Non Probability sampling*

Probability sampling designs

1. Simple random sampling
2. Stratified random sampling
3. Multistage sampling
4. Systematic sampling
5. Cluster sampling
6. Multiphase sampling

Simple random sampling

1. *Lottery method*
2. *Table of random numbers*

Applicable only when population is small, homogenous & the readily available

Stratified random sampling

Followed when population is not homogenous

First divide into homogenous groups or classes = strata

Sample is drawn from each strata by random method

Gives more representation sample & gives greater accuracy

Multistage sampling

Systematic sampling

Cluster sampling

Multiphase sampling

Non-probability sampling designs

- Convinience sampling design
- Judgement sampling
- Quota sampling
- Snowball sampling
- Network sampling

Presentation of data

Advantages

Becomes concise without losing the details

Arouse interest in readers

Become simple & meaningful

Need few words to explain

Become helpful for further analysis

1. Tabulation
2. Drawing

Tabulation

Are devices for presenting data from a mass of statistical data

1. Simple tabulation
2. Complex tabulation

Drawings (Graphs / diagrams)

Quantitative

1. *Histogram*
2. *Frequency Polygon*
3. *Frequency curve*
4. *Line Chart*
5. *Cumulative frequency diagram or Ogive curve*
6. *Scatter or Dot diagram*

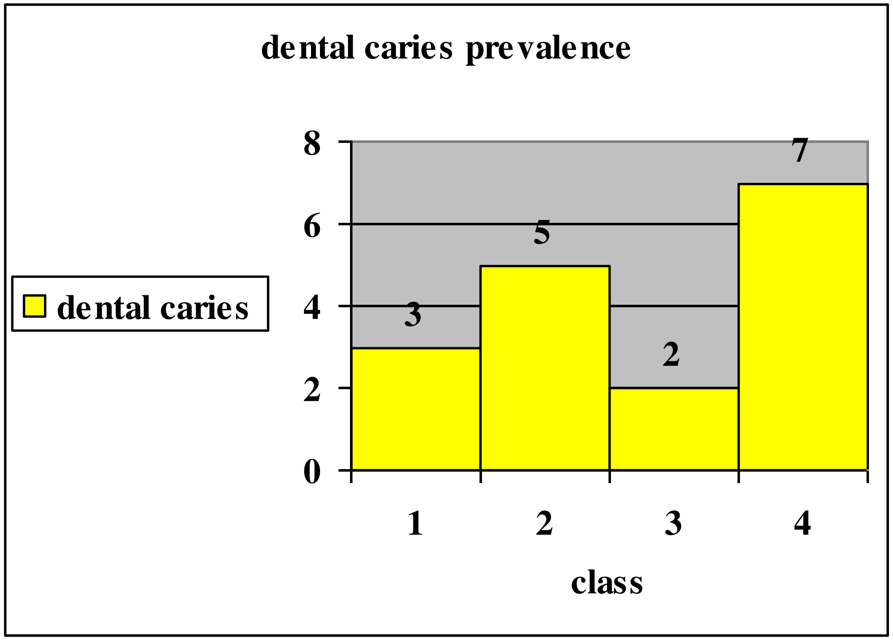
Qualitative

1. *Bar diagram*
2. *Pie diagram*
3. *Pictogram*
4. *Spot map*

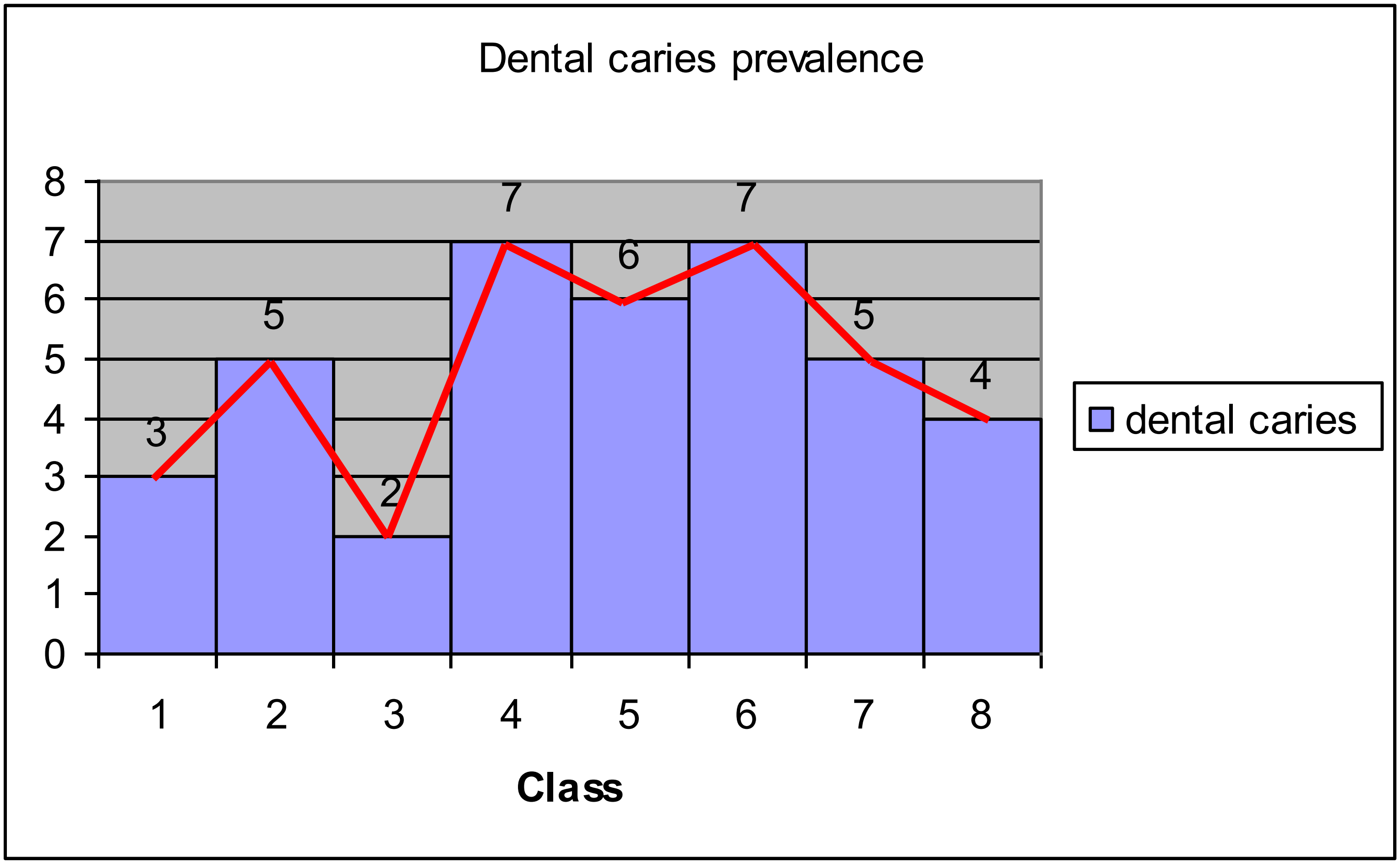
Histogram

Variable on the x axis (abscissa)

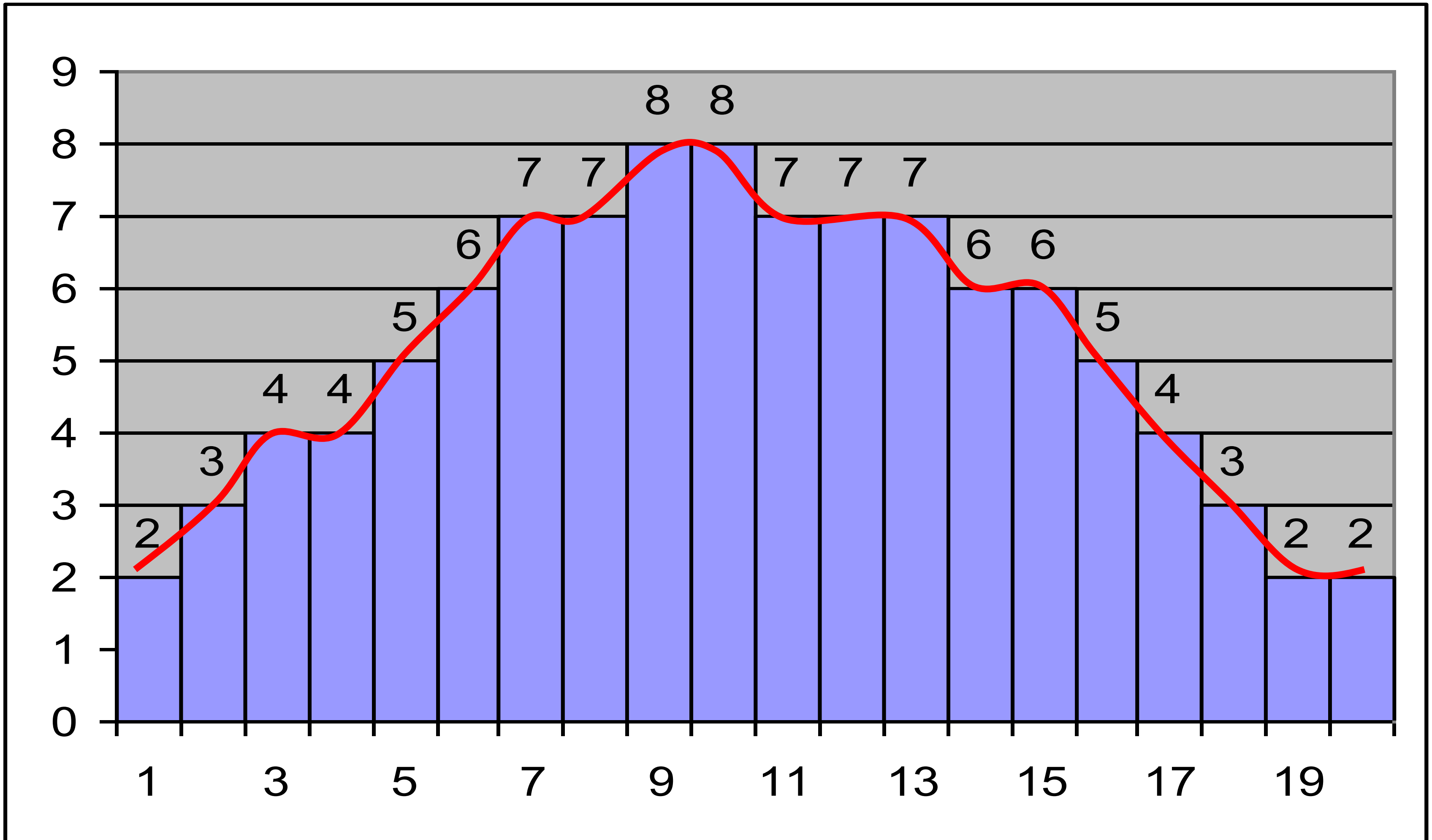
Frequency on the y axis (ordinate)



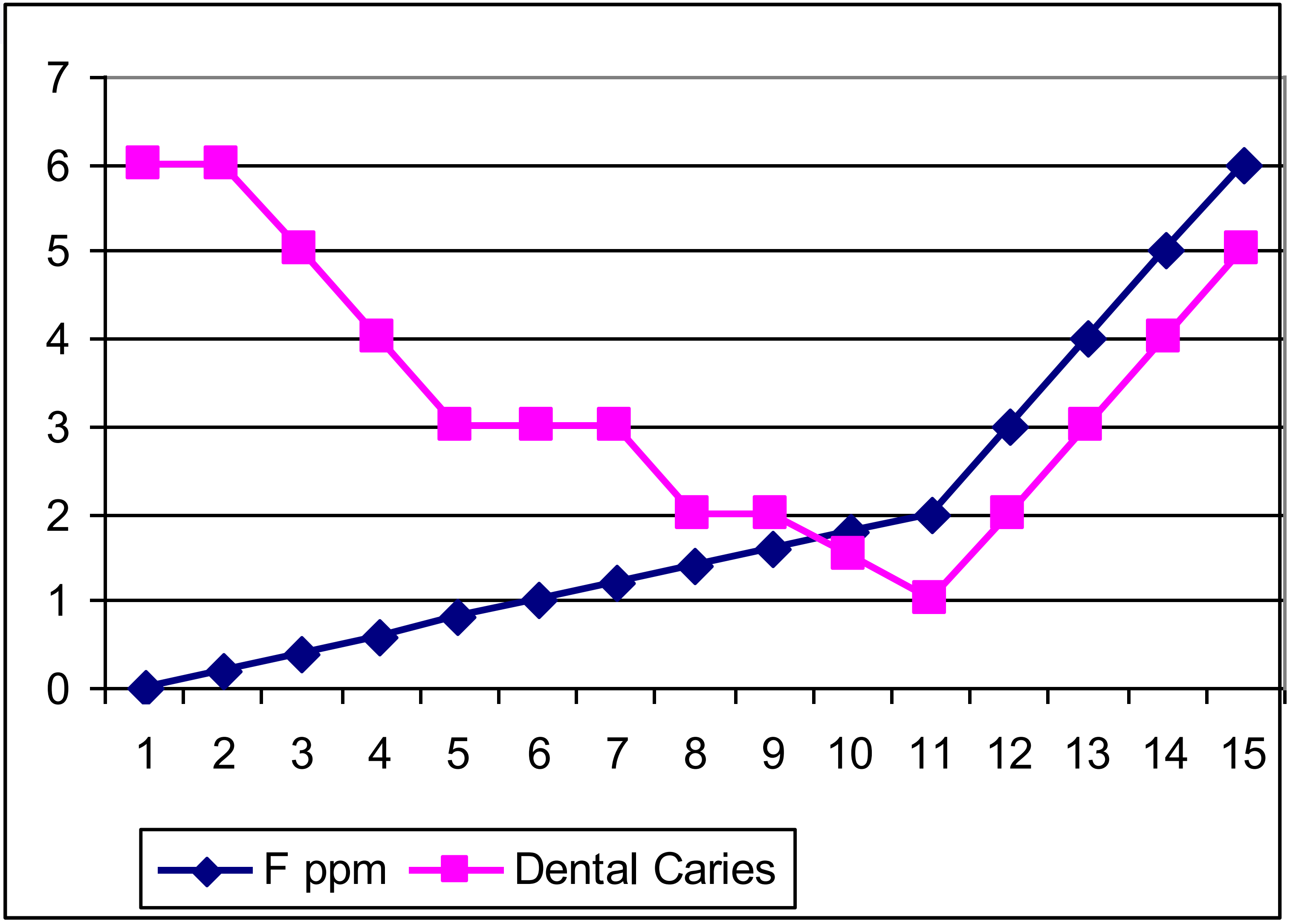
Frequency polygon



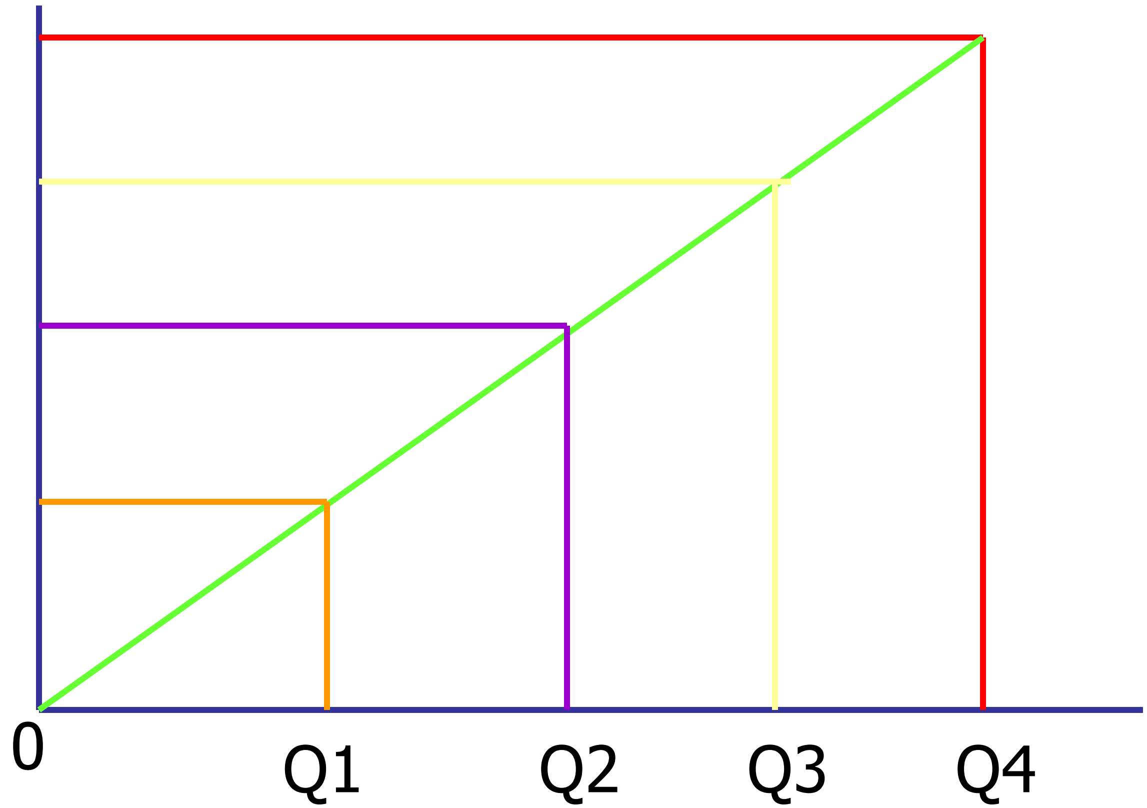
Frequency Curve



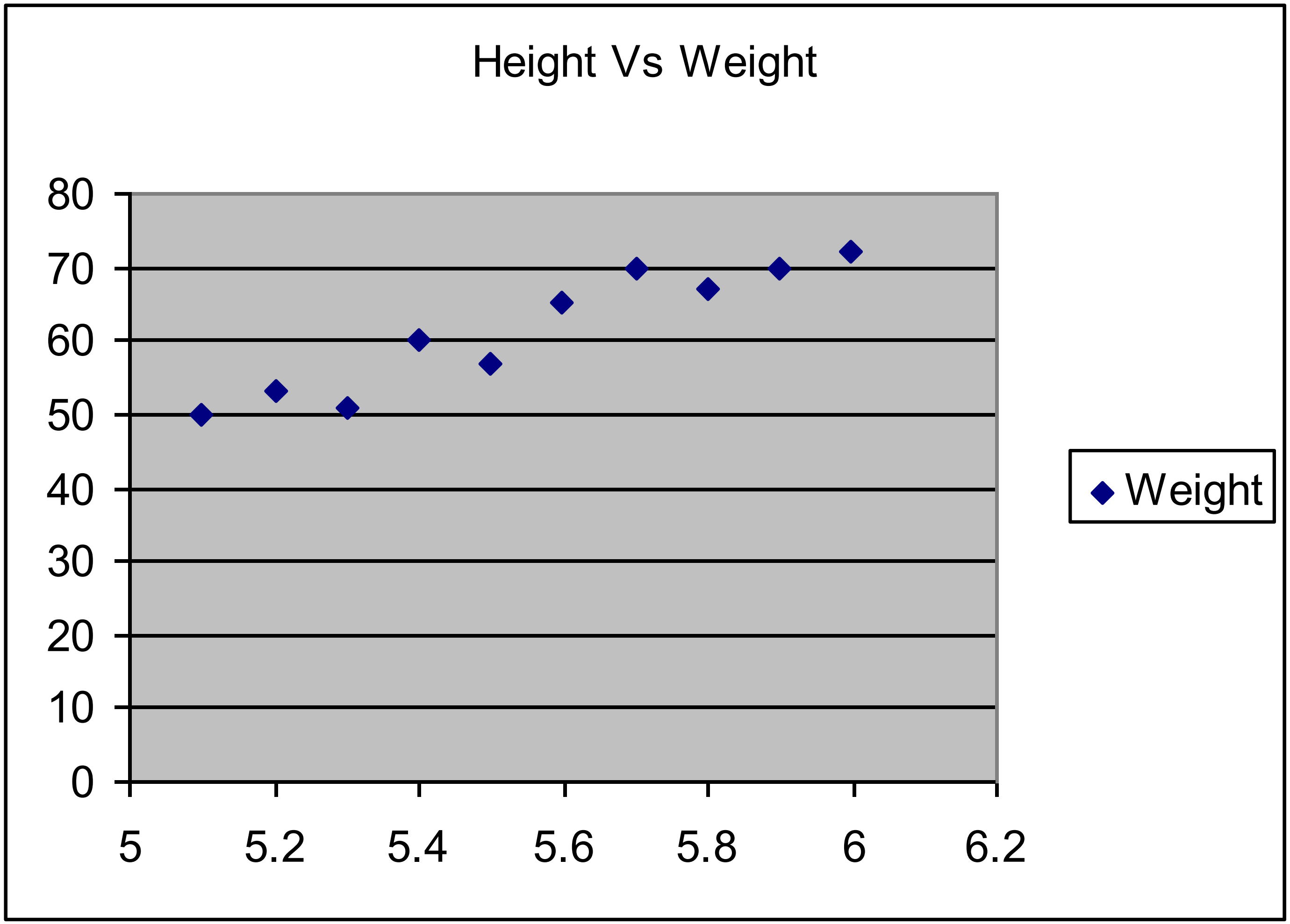
Line graph



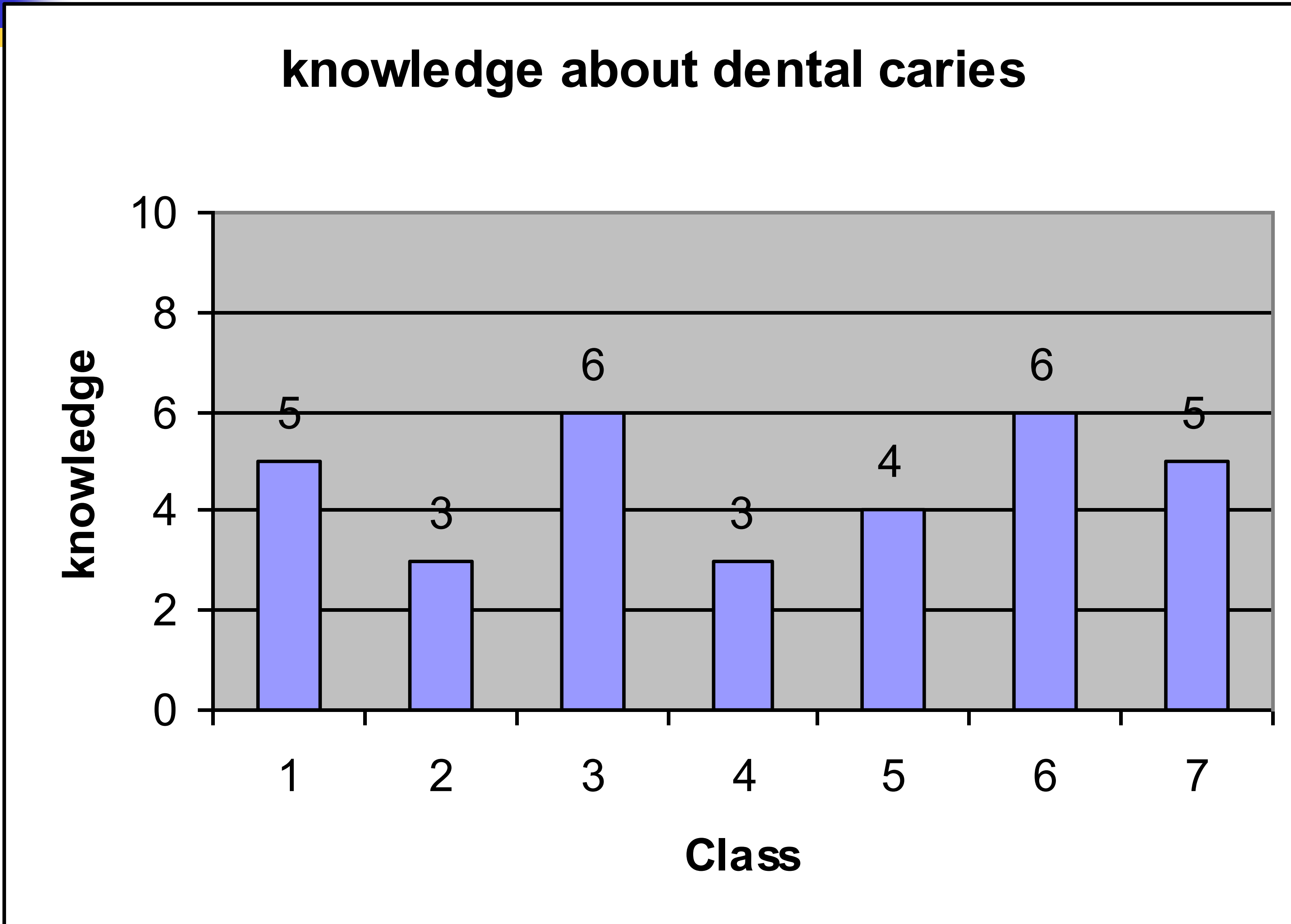
Cumulative curve or Ogive



Scatter or dot Diagram



Bar diagram



Dr Sandesh N

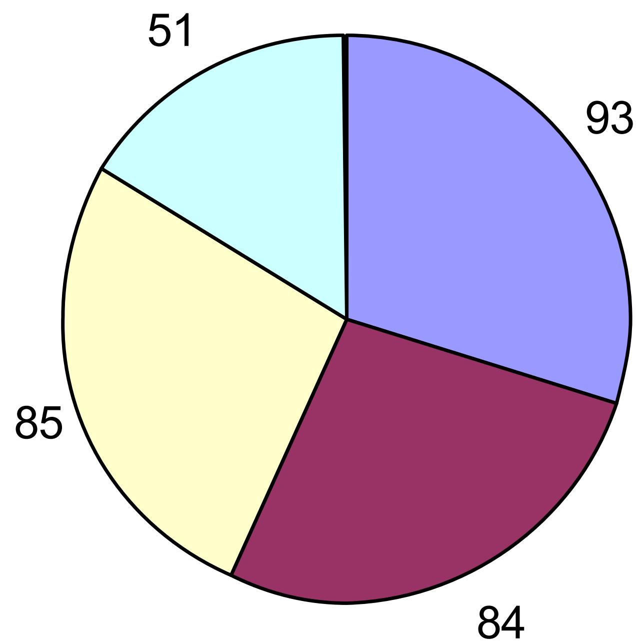
Bar diagram is of three types

1. *Simple Bar diagram*
2. *Multiple Bar diagram*
3. *Proportional Bar diagram*

Pie or Sector Diagram

Size of the angle = $\frac{\textit{Class interval}}{\textit{Total Observation}} \times 360$

Class	NO.	Angles
FIRST year	93	107
SECOND year	84	97
THIRD year	85	98
FOURTH year	51	59
Total	313	



Pictogram or Picture diagram

Map diagram or Spot map

Summarizing the data

Measure of central tendency

1. *Mean*
2. *Median*
3. *Mode*

Measure of Dispersion

1. *Range*
2. *Mean deviation*
3. *Standard deviation*
4. *Coefficient of variation*

Mean

It is a arithmetic mean or arithmetic average which is obtained by dividing the total of all observations by the number of observations

$$\bar{x} \quad \frac{x_1 \quad x_2 \quad x_3 \quad .... \quad x_n}{n} \quad \frac{x}{n}$$

Eg. calculate the mean of DMFT scores 2.3, 2.0,2.7,3.0,2.0.

$$\bar{x} \quad \frac{2.3 \quad 2.0 \quad 2.7 \quad 3.0 \quad 2.0}{5} \quad \frac{12}{5} \quad 2.4$$

Geometric mean (GM) *nth root of the product*

$$GM \quad \sqrt[n]{x_1 \quad x_2 \quad x_3 \ldots x_n} \qquad \frac{\log x}{n}$$

When the variation between the lowest and the highest value is very high, geometric mean is advised & preferred

Harmonic mean (HM) *is the reciprocal of the arithmetic mean of the reciprocal of the observations*

$$HM \qquad \frac{1}{\frac{1}{n} \qquad \frac{1}{x_i}} \qquad \frac{n}{\frac{1}{x_i}}$$

Median

is the middle value, which divides the observed values into two equal parts, when the values are arranged in ascending or descending order

$$\frac{n \quad 1}{2}$$

Eg. calculate the median of DMFT scores 2.3, 2.0,2.7,3.0,2.0.

arrange in asc order,

$$2.0,2.0.2.3,2.7,3.0 \quad \frac{5 \quad 1 \quad 6}{2} \quad 3^{rd\ value} \quad ie\ \ 2.3$$

Mode

is the value of the variable which occurs most frequently

$$Mode = (3median \quad 2mean)$$

Eg. calculate the mode of DMFT scores 2.3, 2.0,2.7,3.0,2.0.

Mode   2.0

Mode   3   2.3   2   2.4   2.1

Measure of Dispersion

Range

It is the difference between highest and the lowest values in the series

Variance or mean deviation

Is the appropriate measure of dispersion for interval or ratio level data

Computes how far each score is from the mean

This is done by $x \quad \bar{x}$

Each score will have a deviation from the mean, so to find the average deviation => we have to add all the deviations and divide it by number of scores (just like calculating mean)

$$i.e.\frac{x \quad \bar{x}}{N}$$

$$but.... \quad x \quad \bar{x} \quad 0$$

So to eliminate this zero, square the deviations which eliminates the (-) sign

$$i.e.\frac{x \quad \bar{x}^{\;2}}{N} \quad S^2$$

- is the average of the squared deviations

Standard deviation (Root Mean Square deviation)

Is defined as the square root of the arithmetic mean of the squared deviations of the individual values from their arithmetic mean

$$SD \qquad \sqrt{\frac{x \quad \bar{x} \quad ^{2}}{N \quad 1}}$$

For small samples

$$SD \qquad s \qquad \sqrt{\frac{x \quad \bar{x} \quad ^{2}}{N}}$$

For large samples

For frequency distribution

$$SD \qquad \sqrt{\frac{f \;\; x \quad \bar{x} \;\; ^{2}}{N \quad 1}}$$

For small samples

$$SD \qquad s \qquad \sqrt{\frac{f \;\; x \quad \bar{x} \;\; ^{2}}{N}}$$

For large samples

Uses of SD

1. *Summarizes the deviations of a large distribution from mean in one figure used as unit of freedom*
2. *Indicates whether the variation from the mean is by chance or real*
3. *Helps finding standard error- which determines whether the difference b/n means of two samples is by chance or real*
4. *Helps finding the suitable size of the sample for valid conclusions*

Standard error

Standard deviation of mean values

Used to compare means with one another

$$SE = \frac{\text{Standard deviation}}{\sqrt{\text{sample size}}} = \frac{SD}{\sqrt{n}}$$

Coefficient of variation

is a measure used to compare relative variability

I.e,

Variation of same character in two or more different series .

(eg pulse rate in young & old person)

Variation of two different characters in one & same series .

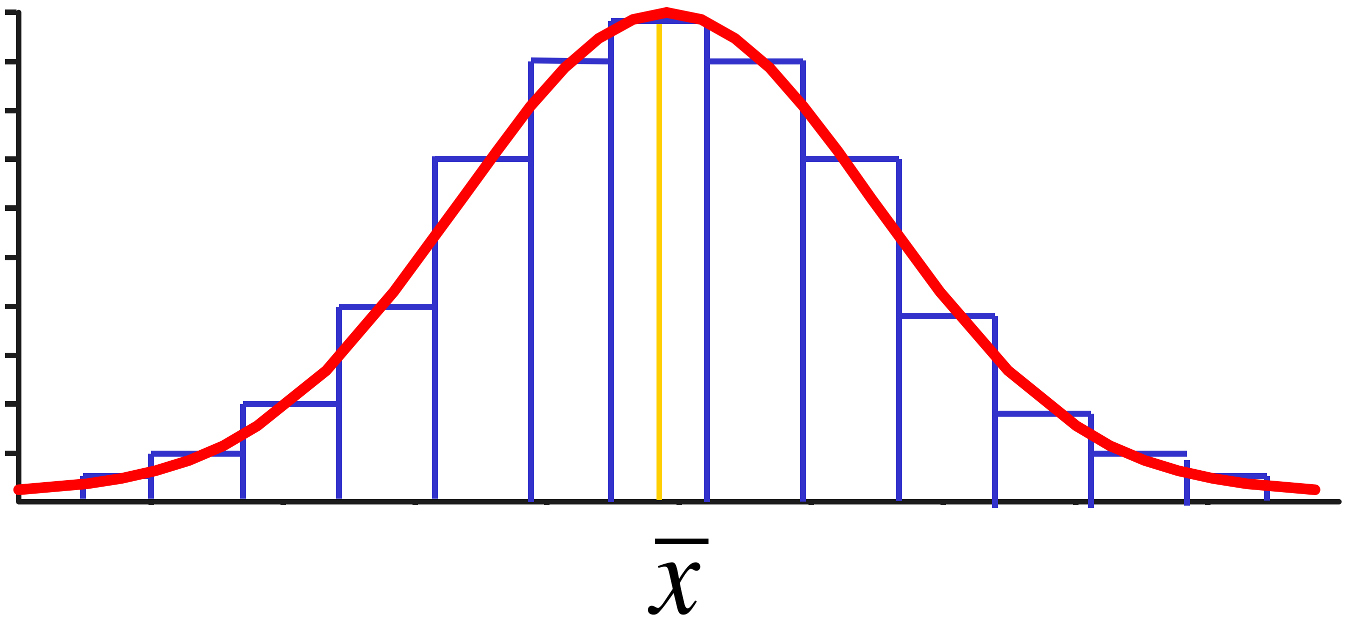
(eg height & weight in same individual)

$$CV \quad \frac{\text{Standard Deviation}}{\text{Mean}} \quad 100$$

Normal curve and distribution

The histogram of the same frequency distribution of heights, with large number of observations & small class intervals   gives a frequency curve which is symmetrical in nature   ***Normal curve or Gaussian curve***

Normal curve



Characteristics of normal curve

Bell shaped

Symmetrical

Mean, Mode & Median coincide

Has two inflections the central part is convex, while at the point of inflection the curve changes from convexity to concavity

On preparing frequency distribution with small class intervals of the data collected, we can observe

1. *Some observations are above the mean & others are below the mean*
2. *If arranged in order, maximum number of frequencies are seen in the middle around the mean & fewer at the extremes decreasing smoothly*
3. *Normally half the observations lie above & half below the mean & all are symmetrically distributed on each side of mean*

A distribution of this nature or shape is called ***Normal or Gaussian distribution***

Arithmetically

$mean \quad 1SD$ limits , include 68.27% observatio ns

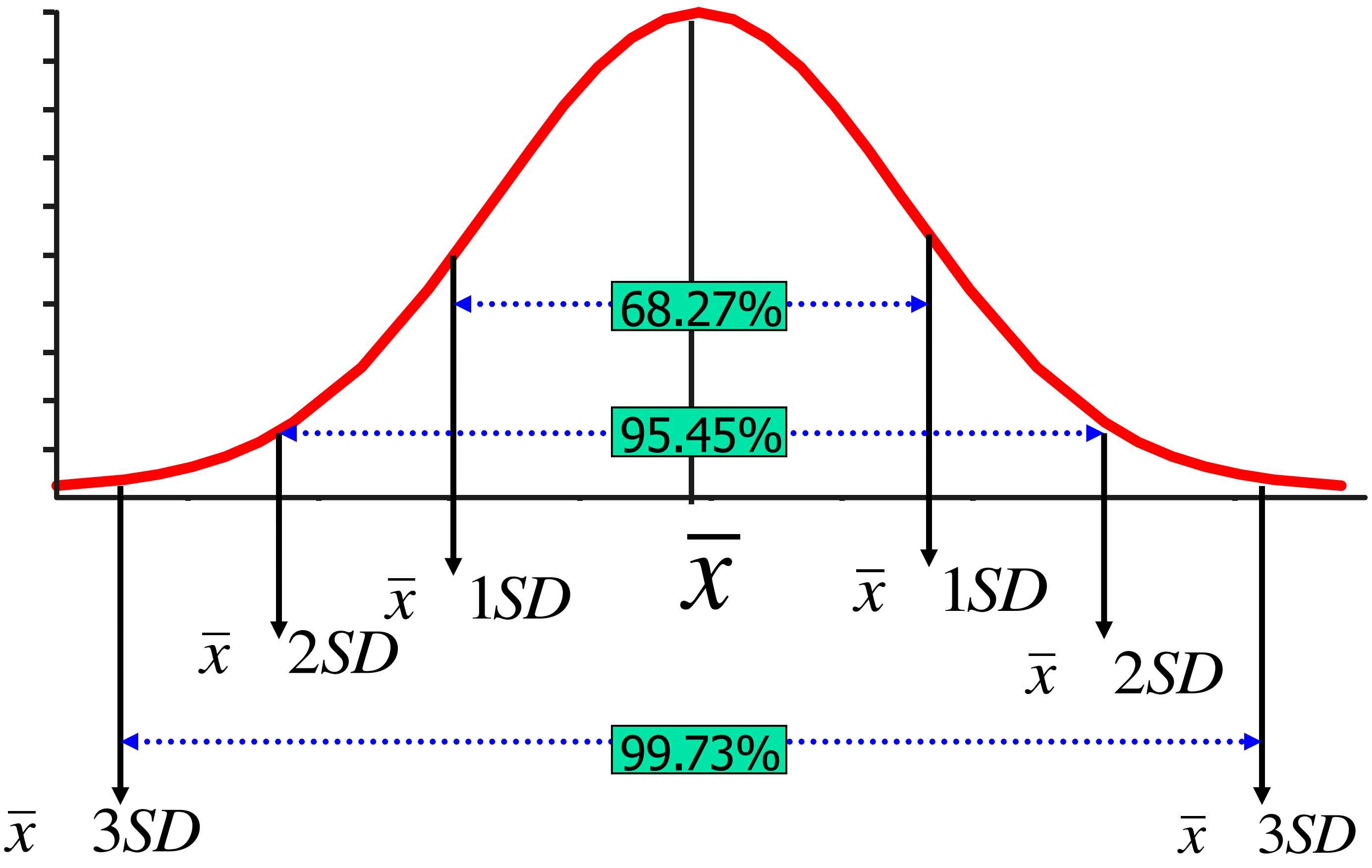
$mean \quad 2SD$ limits, include 95.45% observatio ns

$mean \quad 1.96SD$ limits, include 95% observatio ns

$mean \quad 3SD$limits, includes 99.73% observatio ns

$mean \quad 2.58SD$limits, includes 99% observatio ns

Normal curve and distribution



Dr Sandesh N

Height in cm	frequency of each group	frequency with in height limits of
142.5	3	
145.0	8	
147.5	15	
150.0	45	
152.5	90	
155.0	155	
157.5	194	
160.0(M)	195	
162.5	136	
165.0	93	
167.5	42	
170.0	16	
172.5	6	
175.0-177.5	2	

Mean $\pm$1SD
680
68%

Mean $\pm$2SD
950
95%

Mean $\pm$3SD
995
99%

Mean 160.0 SD 5cm

Skewness

Skewness as the static to measure the asymmetry

coefficient of skewness is 0

Positively (right) skewed

Negatively (left) skewed

Bimodal

kurtosis

Kurtosis is a measure of height of the distribution curve

Coefficient of kurtosis is 3

Leptokurtic(high)

Platykurtic (flat)

Mesokurtic (normal)

Tests of significance

Population

is any finite collection of elements

I.e individuals, items, observations etc,.

Sample

is a part or subset of the population

Parameter

is a constant describing a population

Statistic

is a quantity describing a sample, namely a function of observations

	Statistic (Greek)	Parameter (Latin)
Mean	$\bar{x}$	
Standard Deviation	s	
Variance	s^2	^{2}
Correlation coefficient	r	
Number of subjects	n	N

Dr Sandesh N

Hypothesis testing

Hypothesis H

is an assumption about the status of a phenomenon or is a statement about the parameters or form of population

Null hypothesis or hypothesis of no difference

States no difference between statistic of a sample & parameter of population or b/n statistics of two samples

This nullifies the claim that the experiment result is different from or better than the one observed already

Denoted by H_0

Alternate hypothesis

Any hypothesis alternate to null hypothesis, which is to be tested

Denoted by H_1

Note : the alternate hypothesis is accepted when null hypothesis is rejected

Type I & type II errors

	H_0 Accept	H_1 Accept
H_0 is true	No error	Type I error
H_1 is true	Type II error	No error

Type I error =

Type II error =

When primary concern of the test is to see whether the null hypothesis can be rejected such test is called Test of significance

The probability of committing type I error is called P value

Thus p-value is the chance that the presence of difference is concluded when actually there is none

Type I error important- fixed in advance at a low level such upper limit of tolerance of the chance of type I error is called

Level of Significance ( )

Thus

of type I error

Difference b/n level of significance & P-value -

LOS

1) Maximum tolerable chance of type I error is fixed in advance

P-value

1) Actual probability of type I error

2) calculated on basis of data following procedures

The P-value can be more than

When P-value is < than results is statistically significant

Dr Sandesh N

The level of significance is usually fixed at *5% (0.05) or 1% (0.01) or 0.1% (0.001) or 0.5% (0.005)*

Maximum desirable is 5% level

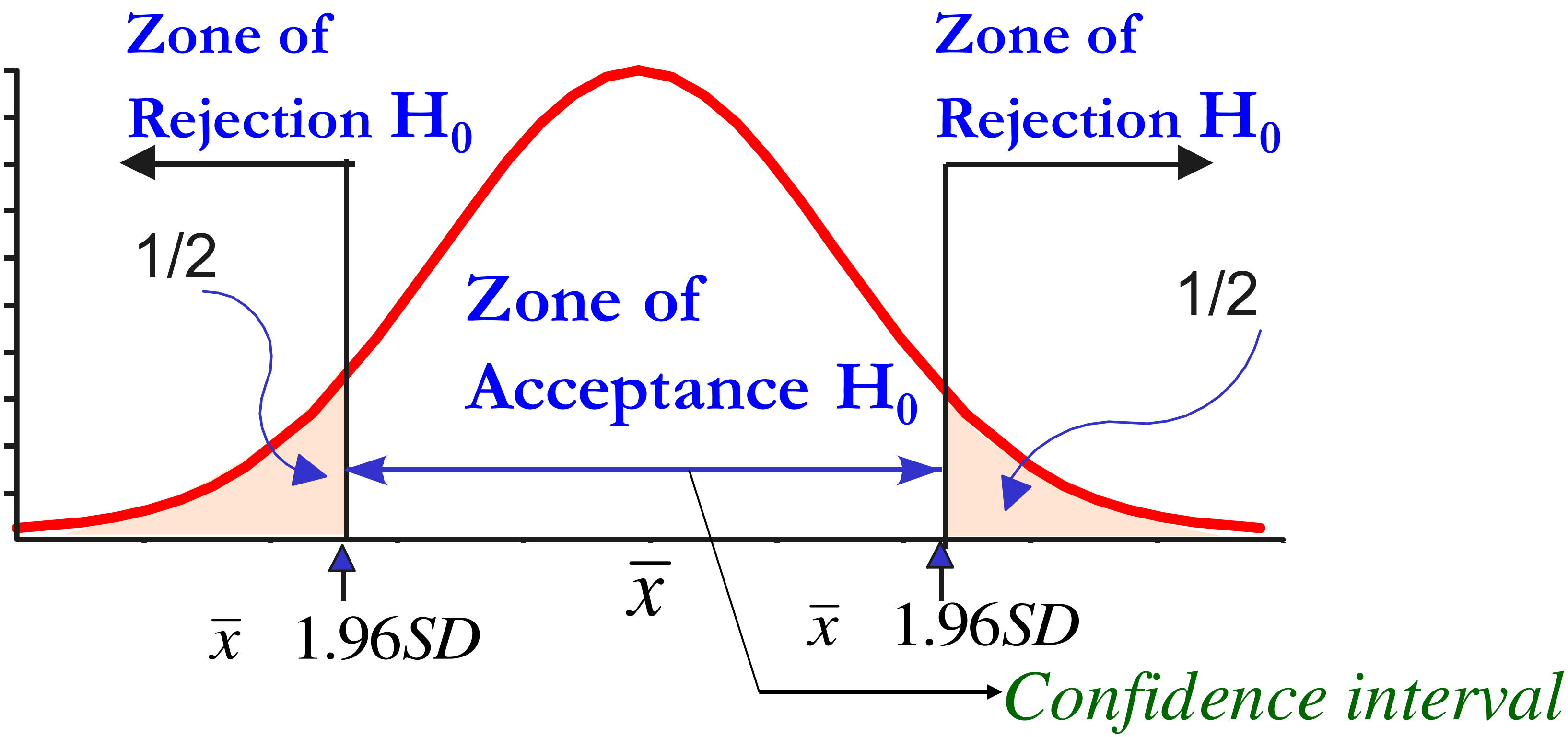
When P-value is b/n

0.05-0.01 = statistically significant

< than 0.01= highly statistically significant

Lower than 0.001 or 0.005 = very highly significant

Sampling Distribution



Confidence limits 95%

Tests of significance

Are mathematical methods by which the probability (P) or relative frequency of an observed difference, occurring by chance is found

Steps & procedure of test of significance

1. *State null hypothesis* H_0
2. *State alternate hypothesis* H_1
3. *Selection of the appropriate test to be utilized & calculation of test criterion based on type of test*

4. *Fixation of level of significance*
5. *Select the table & compare the calculated value with the critical value of the table*
6. *If calculated value is > table value,* H_0 *is rejected*
7. *If calculated value is < table value,* H_0 *is accepted*
8. *Draw conclusions*

TESTS IN TEST OF SIGNIFICANCE

Parametric
(normal distribution &
Normal curve)

Quantitative data

1) *Student t test*
   1) *Paired*
   2) *Unpaired*
2) *Z test*
   (for large samples)
3) *One way ANOVA*
4) *Two way ANOVA*

Qualitative data

1) *Z prop test*
2)

Non-parametric
(not follow
normal distribution)

Qualitative
(quantitative converted
to qualitative)

1. *Mann Whitney U test*
2. *Wilcoxon rank test*
3. *Kruskal wallis test*
4. *Friedmann test*

Parametric	Uses	Non-parametric
Paired t test	Test of diff b/n Paired observation	*Wilcoxon signed rank test*
Two sample t test	Comparison of two groups	*Wilcoxon rank sum test* *Mann Whitney U test* *Kendall s s test*
One way Anova	Comparison of several groups	*Kruskal wallis test*
Two way Anova	Comparison of groups values on two variables	*Friedmann test*
Correlation coefficient	Measure of association B/n two variable	*Spearman s rank Correlation* *Kendall s rank correlation*
Normal test (Z test)		*Chi square test*

Student t test

Small samples do not follow normal distribution as the large ones do => will not give correct results

Prof W.S.Gossett Student t test pen name student

It is the ratio of observed difference b/n two mean of small samples to the SE of difference in the same

Types

- *Unpaired t test*
- *Paired t test*

Actually, t-value Z-value of large samples, but the probability (P) of this is determined by reference ***t table***

Degree of freedom (df)- is the quantity in the denominator which is one less than independent number of observations in a sample

For unpaired t test $= n_1 \quad n_2 \quad 2$

For paired t test $= n\text{-}1$

Criteria for applying t test

Random samples

Quantitative data

Variable follow normal distribution

Sample size less than 30

Application of t test

1. *Two means of small independent sample*
2. *Sample mean and population mean*
3. *Two proportions of small independent samples*

Unpaired t test

I) Difference b/n means of two independent samples

Data

	Group 1	Group 2
Sample size	n_1	n_2
Mean	$\bar{x}_1$	$\bar{x}_2$
SD	SD_1	SD_2

1) Null hypothesis $H_0 \quad \bar{x}_1 \quad \bar{x}_2 \quad 0$

2) Alternate hypothesis $H_1 \quad \bar{x}_1 \quad \bar{x}_2 \quad 0$

3) Test criterion $t \quad \frac{\left| \bar{x}_1 \quad \bar{x}_2 \right|}{SE \ \bar{x}_1 \quad \bar{x}_2}$

here SE of $\bar{x}_1 \quad \bar{x}_2$ is calculated by

$$SE \text{ of } \bar{x}_1 \quad \bar{x}_2 \qquad SD\sqrt{\frac{1}{n_1} \quad \frac{1}{n_2}}$$

$$\text{where} \quad SD \qquad \sqrt{\frac{n_1 \quad 1\, SD_1^2 \qquad n_2 \quad 1\, SD_2^2}{n_1 \quad n_2 \quad 2}}$$

$$SE \ \bar{x}_1 \quad \bar{x}_2 \qquad \sqrt{\frac{n_1 \quad 1\, SD_1^2 \qquad n_2 \quad 1\, SD_2^2}{n_1 \quad n_2 \quad 2} \quad \frac{1}{n_1} \quad \frac{1}{n_2}}$$

4) Calculate degree of freedom

$$df \quad n_1 \quad 1 \quad n_2 \quad 1 \quad n_1 \quad n_2 \quad 2$$

5) Compare the calculated value & the table value

6) Draw conclusions

Example *difference b/n caries experience of high & low socioeconomic group*

Sl no	Details	High socio economic group	Low socio economic group
I	Sample size	$n_1 \quad 15$	$n_2 \quad 10$
II	DMFT	$\bar{x}_1 \quad 2.91$	$\bar{x}_2 \quad 2.26$
III	Standard deviation	$SD_1 \quad 0.27$	$SD_2 \quad 0.22$

$$t \quad \frac{\left|\bar{x}_1 \quad \bar{x}_2\right|}{SE\ \bar{x}_1 \quad \bar{x}_2} \quad \frac{0.65}{0.1027} \quad 6.34, \quad df \quad 23$$

$$t_{0.001} \quad 3.76 \qquad\qquad t_c \quad t_{0.001}$$

There is a significant difference

Table A3 Percentage points of the *t* distribution. **T table**

Adapted from Table 7 of White *et al.* (1979) with permission of authors and publishers.

	One-sided P value								
	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	Two-sided P value								
d.f.	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62
2	0.82	1.89	2.92	4.30	6.96	9.92	14.09	22.33	31.60
3	0.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92
4	0.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61
5	0.73	1.48	2.02	2.57	3.36	4.03	4.77	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96
7	0.71	1.42	1.90	2.36	3.00	3.50	4.03	4.78	5.41
8	0.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04
9	0.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78
10	0.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59
11	0.70	1.36	1.80	2.20	2.72	3.11	3.50	4.02	4.44
12	0.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32
13	0.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22
14	0.69	1.34	1.76	2.14	2.62	2.98	3.33	3.79	4.14
15	0.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07
16	0.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02
17	0.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.96
18	0.69	1.33	1.73	2.10	2.55	2.88	3.20	3.61	3.92
19	0.69	1.33	1.73	2.09	2.54	2.86	3.17	3.58	3.88
20	0.69	1.32	1.72	2.09	2.53	2.84	3.15	3.55	3.85
21	0.69	1.32	1.72	2.08	2.52	2.83	3.14	3.53	3.82
22	0.69	1.32	1.72	2.07	2.51	2.82	3.12	3.50	3.79
23	0.68	1.32	1.71	2.07	2.50	2.81	3.10	3.48	3.77
24	0.68	1.32	1.71	2.06	2.49	2.80	3.09	3.47	3.74
25	0.68	1.32	1.71	2.06	2.48	2.79	3.08	3.45	3.72
26	0.68	1.32	1.71	2.06	2.48	2.78	3.07	3.44	3.71
27	0.68	1.31	1.70	2.05	2.47	2.77	3.06	3.42	3.69
28	0.68	1.31	1.70	2.05	2.47	2.76	3.05	3.41	3.67
29	0.68	1.31	1.70	2.04	2.46	2.76	3.04	3.40	3.66
30	0.68	1.31	1.70	2.04	2.46	2.75	3.03	3.38	3.65
40	0.68	1.30	1.68	2.02	2.42	2.70	2.97	3.31	3.55
60	0.68	1.30	1.67	2.00	2.39	2.66	2.92	3.23	3.46
120	0.68	1.29	1.66	1.98	2.36	2.62	2.86	3.16	3.37
∞	0.67	1.28	1.65	1.96	2.33	2.58	2.81	3.09	3.29

Other applications

II) Difference b/n sample mean & population mean

$$t \quad \frac{\left|\bar{x} \qquad \right|}{SE \quad SD/\sqrt{n}} \qquad df \quad n \quad 1$$

III) Difference b/n two sample proportions

$$t \quad \frac{\left|p_1 \quad p_2\right|}{\sqrt{PQ \quad \frac{1}{n_1} \quad \frac{1}{n_2}}} \qquad \text{where} \quad P \quad \frac{n_1 p_1 \quad n_2 p_2}{n_1 \quad n_2}$$

$$Q \quad 1 \quad P$$

$$df \quad n_1 \quad n_2 \quad 2$$

Paired t test

Is applied to paired data of observations from one sample only when each individual gives a paired of observations

Here the pair of observations are correlated and not independent, so for application of t test following procedure is used-

1. *Find the difference for each pair* $y_1 \quad y_2 \quad x$
2. *Calculate the mean of the difference* (x) *ie* $\bar{x}$
3. *Calculate the SD of the differences & later SE*

$$SE \quad \frac{SD}{\sqrt{n}}$$

4. Test criterion $t \quad \frac{\bar{x} \quad 0}{SE \ d} \quad \frac{\bar{x}}{SD \ x \big/ \sqrt{n}}$

5. Degree of freedom $df \quad n \quad 1$

6. Refer t table & find the probability of calculated value

7. Draw conclusions

Example *to find out if there is any significant improvement in DAI scores before and after orthodontic treatment*

Sl no	DAI before	DAI after	Difference	Squares
1	30	24	6	36
2	26	23	3	9
3	27	24	3	9
4	35	25	10	100
5	25	23	2	4
Total			24	158

$$Mean\ x = \frac{x}{n} = \frac{24}{5} = 4.8$$

$$\text{sum of squares,} \quad \left(x \quad \bar{x}\right)^2 = \left(6 \quad 4\right)^2 + \left(3 \quad 4\right)^2 + \left(3 \quad 4\right)^2 + \left(10 \quad 4\right)^2 + \left(2 \quad 4\right)^2$$

$$4 \quad 1 \quad 1 \quad 36 \quad 4$$

$$46$$

$$SD = \sqrt{\frac{\left(x \quad \bar{x}\right)^2}{n \quad 1}} = \sqrt{\frac{46}{4}} = \sqrt{11.5} = 3.391$$

$$SE = \frac{SD}{\sqrt{n}} = \frac{3.391}{\sqrt{5}} = 1.5179$$

$$t_c = \frac{\bar{x}}{SE} = \frac{4.8}{1.5179} = 3.162 \qquad df = n \quad 1 = 4$$

$$\text{but} \quad t_{0.5} = 2.78$$

$t_c > t_{0.5}$ Hence significant

Dr Sandesh N

Z test (Normal test)

Similar to t test in all aspect except that the sample size should be > 30

In case of normal distribution, the tabulated value of Z at -

5% level	$Z_{0.05}$	1.960
1% level	$Z_{0.01}$	2.576
0.1% level	$Z_{0.001}$	3.290

Z test can be used for

1. **Comparison of means of two samples**

$$Z \quad \frac{\bar{x}_1 \quad \bar{x}_2}{SE\ \bar{x}_1 \quad \bar{x}_2}$$

where $SE\ \bar{x}_1 \quad \bar{x}_2 \quad \sqrt{SE_1{}^2 \quad SE_2{}^2}$

$$\sqrt{\frac{SD_1^2}{n_1} \quad \frac{SD_2^2}{n_2}}$$

2. **Comparison of sample mean & population mean**

$$Z \quad \frac{|\bar{x} \quad |}{\sqrt{\frac{SD^2}{n}}}$$

3. Difference b/n two sample proportions

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $P = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

$Q = 1 - P$

4. Comparison of sample proportion (or percentage) with population proportion (or percentage)

$$Z = \frac{p - P}{\sqrt{PQ\left(\frac{1}{n}\right)}}$$

Where p = sample proportion

P = populn proportion

Analysis of variance (ANOVA)

Useful for comparison of means of several groups

Is an extension of student s t test for more than two groups

R A Fisher in 1920 s

Has four models

1. *One way classification (one way ANOVA)*
2. *Single factor repeated measures design*
3. *Nested or hierarchical design*
4. *Two way classification (two way ANOVA)*

One way ANOVA

Can be used to compare like-

Effect of different treatment modalities

Effect of different obturation techniques on the apical seal , etc,.

Groups (or treatments)	1	2	i	k
Individual values	x_{11}	x_{21}	x_{i1}	x_{k1}
	x_{12}	x_{22}	x_{i2}	x_{k2}
	x_{1n}	x_{2n}	x_{in}	x_{kn}
Calculate				
No of observations	n	n	n	n
Sum of x values	$_1 \quad x_{11} \quad x_{12} \quad \dots \quad x_{1n}$	T_2	T_i	T_k
Sum of squares	$S_1 \quad x_{11}{}^2 \quad x_{12}{}^2 \quad .. \quad x_{1n}{}^2$	S_2	S_i	S_k
Mean of values	$\bar{x}_1 \quad \frac{T_1}{n}$	$\bar{x}_2$	$\bar{x}_i$	$\bar{x}_k$

ANOVA table

Sl no	Source of variation	Degree of freedom	Sum of squares	Mean sum of squares	F ratio or variance ratio
I	Between Groups	$k \quad 1$	$_{i}\, x_i \quad \bar{x}\ ^{2} \qquad _{i}\, x_i^2 \quad \frac{T^2}{N}$	$S_B^2 \quad \frac{_{i}\, x_i \quad \bar{x}\ ^{2}}{k \quad 1}$	$\frac{S_B^2}{S_W^2} \quad k \quad 1, N \quad k$
II	With in groups	$n \quad k$	$_{i}\ _{j}\, x_{ij} \quad \bar{x}_i\ ^{2} \qquad _{i}\ _{j}\, x_{ij}^2 \quad _{i}\, \frac{T_i^2}{n_i}$	$S_W^2 \quad \frac{_{i}\ _{j}\, x_{ij}^2 \quad _{i}\, T_i^2/n_i}{N \quad k}$	
III	Total	$n \quad 1$	$_{i}\ _{j}\, x_{ij} \quad \bar{x}\ ^{2} \qquad _{i}\ _{j}\, x_{ij}^2 \quad \frac{T^2}{N}$	$S_T^2 \quad \frac{_{i}\ _{j}\, x_{ij}^2 \quad T^2/N}{N \quad 1}$	

Table A4 Percentage points of the F distribution.

Adapted from Table 4 of Armitage (1971) and Table 18 of Pearson & Hartley (1966) with permission of the authors and publishers and the Biometrika Trustees.

The table gives a one-sided significance test for the comparison of two variances, as appropriate for use in analysis of variance. A two-sided test may be obtained by doubling the P values.

ANOVA

$d.f._1$ = d.f. for numerator; $d.f._2$ = d.f. for denominator

		$d.f._1$														
$d.f._2$	P value	1	2	3	4	5	6	7	8	9	10	20	40	60	120	∞
1	0.05	161	200	216	225	230	234	237	239	241	242	248	251	252	253	254
	0.025	648	800	864	900	922	937	948	957	963	969	993	1006	1010	1014	1018
	0.01	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6209	6287	6313	6339	6366
	0.005	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24836	25148	25253	25359	25465
	0.001	405300	500000	540400	562500	576400	585900	592900	598100	602300	605600	620900	628700	631300	634000	636600
2	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.47	19.48	19.49	19.50
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.45	39.47	39.48	39.49	39.50
	0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.45	99.47	99.48	99.49	99.50
	0.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5
	0.001	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.4	999.4	999.5	999.5	999.5	999.5
3	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.59	8.57	8.55	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.17	14.04	13.99	13.95	13.90
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.69	26.41	26.32	26.22	26.13
	0.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	42.78	42.31	42.15	41.99	41.83
	0.001	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9	129.2	126.4	125.0	124.5	124.0	123.5
4	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.72	5.69	5.66	5.63
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.56	8.41	8.36	8.31	8.26
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.02	13.75	13.65	13.56	13.46
	0.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.17	19.75	19.61	19.47	19.32
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	46.10	45.09	44.75	44.40	44.05

***Example-** see whether there is a difference in number of patients seen in a given period by practitioners in three group practice*

Practice	A	B	C
Individual values	268	387	161
	349	264	346
	328	423	324
	209	254	293
	292		239
Calculate			
No of observations (n)	5	4	5
Sum of x values	1441	1328	1363
Sum of squares	426899	462910	393583
Mean of values	288.2	332.0	272.6

Between group sum of squares

$$\frac{x_A{}^{2}}{n_A} \quad \frac{x_B{}^{2}}{n_B} \quad \frac{x_C{}^{2}}{n_C} \quad \frac{x_A \quad x_B \quad x_C{}^{2}}{n_A \quad n_B \quad n_C}$$

8215.71

Total sum of squares

$$x_A^2 \quad x_B^2 \quad x_C^2 \quad \frac{x_A \quad x_B \quad x_C{}^{2}}{n_A \quad n_B \quad n_C}$$

63861.71

With in group sum of squares

total SS - between SS

55646.0

ANOVA table

Sl no	*Source of variation*	*Degree of freedom*	*Sum of squares*	*Mean sum of squares*	*F ratio or variance ratio*
I	*Between Groups*	3 1 2	8215.71	$\frac{8215.71}{2}$ 4107.86	$\frac{4107.86}{5088.73}$ 0.81
II	*With in groups*	14 3 11	55646	$\frac{55646}{11}$ 5088.73	
III	*Total*	14 1 13	63861.71		

F 0.81 $F_{0.05}$ 3.98 df 2,11

Because $F_C < F_T$, there is no significant difference in the number of patients attending 3 different practice

Further, any particular pair of treatments can be compared using SE of difference b/n two means

Eg $\bar{x}_d \ \& \ \bar{x}_c$

$$SE \ \bar{x}_d \quad \bar{x}_c \qquad \sqrt{MSE \ \frac{1}{n_d} \quad \frac{1}{n_c}}$$

& difference $\bar{x}_d \quad \bar{x}_c$ may be tested by using

't' test criterion

$$t \qquad \frac{\bar{x}_d \quad \bar{x}_c}{SE \ \bar{x}_d \quad \bar{x}_c}$$

Dr Sandesh N

Two way ANOVA

Is used to study the impact of two factors on variations in a specific variable

Eg Effect of age and sex on DMFT value

Sample values							
blocks	***Treatments***				***sample size***	***Total***	***Mean value***
i	x_{11}	x_{21}	x_{31}	x_{k1}	k	T_1	$\bar{x}_1$
ii	x_{12}	x_{22}	x_{32}	x_{k2}	k	T_2	$\bar{x}_2$
..							
n	x_{1n}	x_{2n}	x_{3n}	x_{kn}	k	T_n	$\bar{x}_n$
Sample size	n	n	n	n	nk N		
Total	T_1	T_2	T_3	T_k		T	
Mean value	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_k$			$\bar{x}$

2 way ANOVA table

Sl no	Source	Sum of squares	Degree of freedom	Mean sum of squares (MSS)	Variance ratio F
I	Blocks	SS_{blocks}	$n \quad 1$	$MS_{\text{blocks}} \quad \frac{SS_{blocks}}{n \quad 1}$	$F_1 \quad \frac{MS_{blocks}}{MS_{residual}}$
II	Treatments	$SS_{treatments}$	$k \quad 1$	$MS_{treatments} \quad \frac{SS_{treatments}}{k \quad 1}$	$F_2 \quad \frac{MS_{treatment}}{MS_{residual}}$
III	Residual or error	$SS_{residual}$	$n \quad 1 \quad k \quad 1$	$MS_{residual} \quad \frac{SS_{residual}}{n \quad 1 \quad k \quad 1}$	
IV	Total	SS_{total}	$nk \quad 1 \quad N \quad 1$		

F_1 variance ratio of blocks with *df* of $n \quad 1$ Vs $n \quad 1 \quad k \quad 1$

F_2 variance ratio of treatment 's with *df* of $k \quad 1$ Vs $n \quad 1 \quad k \quad 1$

Multiple comparison tests

1. Fisher s procedure *student s t test*
2. Least significant difference method (LSD)

   Just like student s t test

   To test significant difference b/n two groups or variable means
3. Scheffe s significant difference procedure

   Is applicable when groups having heterogeneous variance or variations
4. Tukey s method

   For comparison of the differences b/n all possible pairs of treatments or group means

5. Duncan s multiple comparison test

For all comparisons of paired groups only

6. Dunnet s comparison test procedure

For comparison of one control and several treatment groups

Non parametric tests

Here the distribution do not require any specific pattern of distribution. They are applicable to almost all kinds of distribution

- *Chi square test*
- *Mann Whitney U test*
- *Wilcoxon signed rank test*
- *Wilcoxon rank sum test*
- *Kendall s S test*
- *Kruskal wallis test*
- *Spearman s rank correlation*

Chi square test

By Karl Pearson & denoted as

Application

1. *Alternate test to find the significance of difference in two or more than two proportions*
2. *As a test of association b/n two events in binomial or multinomial samples*
3. *As a test of goodness of fit*

Requirement to apply chi square test

Random samples

Qualitative data

Lowest observed frequency not less than 5

Contingency table

Frequency table where sample classified according to two different attributes

2 rows ; 2 columns => 2 X 2 contingency table

r rows : c columns => rXc contingency table

$$^{2} \qquad \frac{O \quad E^{\;2}}{E}$$

O observed frequency

E expected frequency

Steps

1. *State null & alternate hypothesis*
2. *Make contingency table of the data*

$$r \quad c$$

3. *Determine expected frequency by*

$$E \quad \frac{r \quad c}{N \ \text{total frequency}}$$

4. *Calculate chi-square of each by-*

$$2 \quad \frac{O \quad E^{2}}{E}$$

5. *calculate degree of freedom*

$$df \qquad c \quad 1 \;\; r \quad 1$$

6. *Sum all the chi-square of each cell   this gives chi-square value of the data*

$$^{2} \qquad\qquad \frac{O \quad E^{\,2}}{E}$$

7. *Compare the calculated value with the table value at any LOS*

8. *Draw conclusions*

Example from a dental health campaign

School	Oral hygiene				Total
	G	F_+	F_-	P	
Below avg	62 (85.9)	103 (93.0)	57 (45.2)	11 (8.9)	233
Avg	50 (43.9)	36 (47.5)	26 (23.1)	7 (4.6)	119
Above avg	80 (62.3)	69 (67.5)	18 (32.8)	2 (6.5)	169
Total	192	208	101	20	521

$$E \quad \frac{r \quad c}{N \ \text{total frequency}} \qquad\qquad df \quad c \quad 1 \quad r \quad 1 \quad 3 \quad 2 \quad 6$$

$${}^{2}_{c} \quad \frac{O \quad E^{2}}{E} \quad 31.4 \qquad\qquad \text{Table} \ {}^{2}_{t} \ \text{at P} \quad 0.001 \ \text{is} \ 22.46$$

Hence significant difference

Dr Sandesh N

Table A5 Percentage points of the χ^2 distribution.

Adapted from Table 8 of White *et al.* (1979) with permission of the authors and publishers.

d.f. = 1. In the comparison of two proportions (2×2 χ^2 or Mantel–Haenszel χ^2 test) or in the assessment of a trend, the percentage points give a two-sided test. A one-sided test may be obtained by halving the *P* values. (Concepts of one- and two-sidedness do not apply to larger degrees of freedom, as these relate to tests of multiple comparisons.)

	P value							
d.f.	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.13
9	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	69.33	77.58	85.53	90.53	95.02	100.43	104.22	112.32
80	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	89.33	98.65	107.57	113.15	118.14	124.12	128.30	137.21
100	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45

Alternate formulae

If we have contingency table

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d=N

$$^{2} \quad \frac{N \; ad \quad bc^{\;2}}{a \quad b \;\; c \quad d \;\; a \quad c \;\; b \quad d} \quad \text{with} \quad df \quad 1$$

If one of the value is below 5 => Yate s correction formula

$$^{2} \quad \frac{\left(N \; \left|ad \quad bc\right| \; \frac{N}{2}\right)^{2}}{a \quad b \;\; c \quad d \;\; a \quad c \;\; b \quad d} \quad \text{with} \quad df \quad 1$$

If the table is larger than 2X2, Yate s correction cannot be applied then the small frequency (<5) can be pooled or combined with next group or class in the table

Chi square test only tells the presence or absence of association, but does not measure the strength of association

If degree of association as to be calculated then

1. ***Yule s coefficient of association*** $Q \quad \dfrac{ad \quad bc}{ad \quad bc}$

2. ***Yule s coefficient of colligation*** $Y \quad \dfrac{1 \quad \sqrt{bc/ad}}{1 \quad \sqrt{bc/ad}}$

3. **-** $V \quad \dfrac{ad \quad bc}{\sqrt{a \quad b \quad a \quad c \quad b \quad d \quad c \quad d}}$

4. ***Pearson s coefficient of contingency***

$$C \quad \sqrt{\frac{\quad^{2}}{N \quad\quad ^{2}}}$$

Wilcoxon signed rank test

Is equivalent to paired t test

Steps

Exclude any differences which are zero

Put the remaining differences in ascending order, ignoring the signs

Gives ranks from lowest to highest

If any differences are equal, then average their ranks

Count all the ranks of positive differences T_+

Count all the ranks of negative differences $T-$

If there is no differences b/n variables then T_+ & T_- will be similar, but if there is difference then one sum will be large and the other will be much smaller

T= smaller of T_+&T_-

Compare the T value with the critical value for 5%, 2% & 1% significance level

A result is significant if it is smaller than critical value

Example: *Results of a placebo-controlled clinical trail to test the effectiveness of sleeping drug*

Patients	Sleep hrs	
	Drug	Placebo
1	6.1	5.2
2	7.0	7.9
3	8.2	3.9
4	7.6	4.7
5	6.5	5.3
6	8.4	5.4
7	6.9	4.2
8	6.7	6.1
9	7.4	3.8
10	5.8	6.3

Difference
0.9
-0.9
4.3
2.9
1.2
3.0
2.7
0.6
3.6
-0.5

Rank with signs	
+	-
3.5	-
-	-3.5
10	-
7	-
5	-
8	-
6	-
2	-
9	-
-	-1
50.5	-4.5

Calculated T=-4.5 df =10,
Table value at 5% (n=10)= 8

Cal T< table value, H_0 is rejected

We conclude that sleeping drug is more effective than the placebo

Mann Whitney U test

Is used to determine whether two independent sample have been drawn from same sample

It is a alternative to student t test & requires at least ordinal or normal measurement

$$U \quad n_1 n_2 \quad \frac{n_1 \;\; n_1 \quad 1}{2} \quad R_1 \; or \; R_2$$

Where, $n_1 n_2$ are sample sizes

R_1 R_2 are sum of ranks assigned to I & II group

Procedure

All the observation in two samples are ranked numerically from smallest to largest without regarding the groups

Then identify the observation for I and II samples

Sum of ranks for I and II sample determined separately

Take difference of two sum $T = R_1 - R_2$

Comparison of birth weights of children born to 15 non smokers with those of children born to 14 heavy smokers

NS	3.9	3.7	3.6	3.7	3.2	4.2	4.0	3.6	3.8	3.3	4.1	3.2	3.5	3.5	2.7
HS	3.1	2.8	2.9	3.2	3.8	3.5	3.2	2.7	3.6	3.7	3.6	2.3	2.3	3.6	

Ranks assignments

R1	26	23	16	21	8	29	27	17	24	12	28	10	15	13	03
R2	7	5	6	11	25	14	9	4	20	22	19	2	1	18	

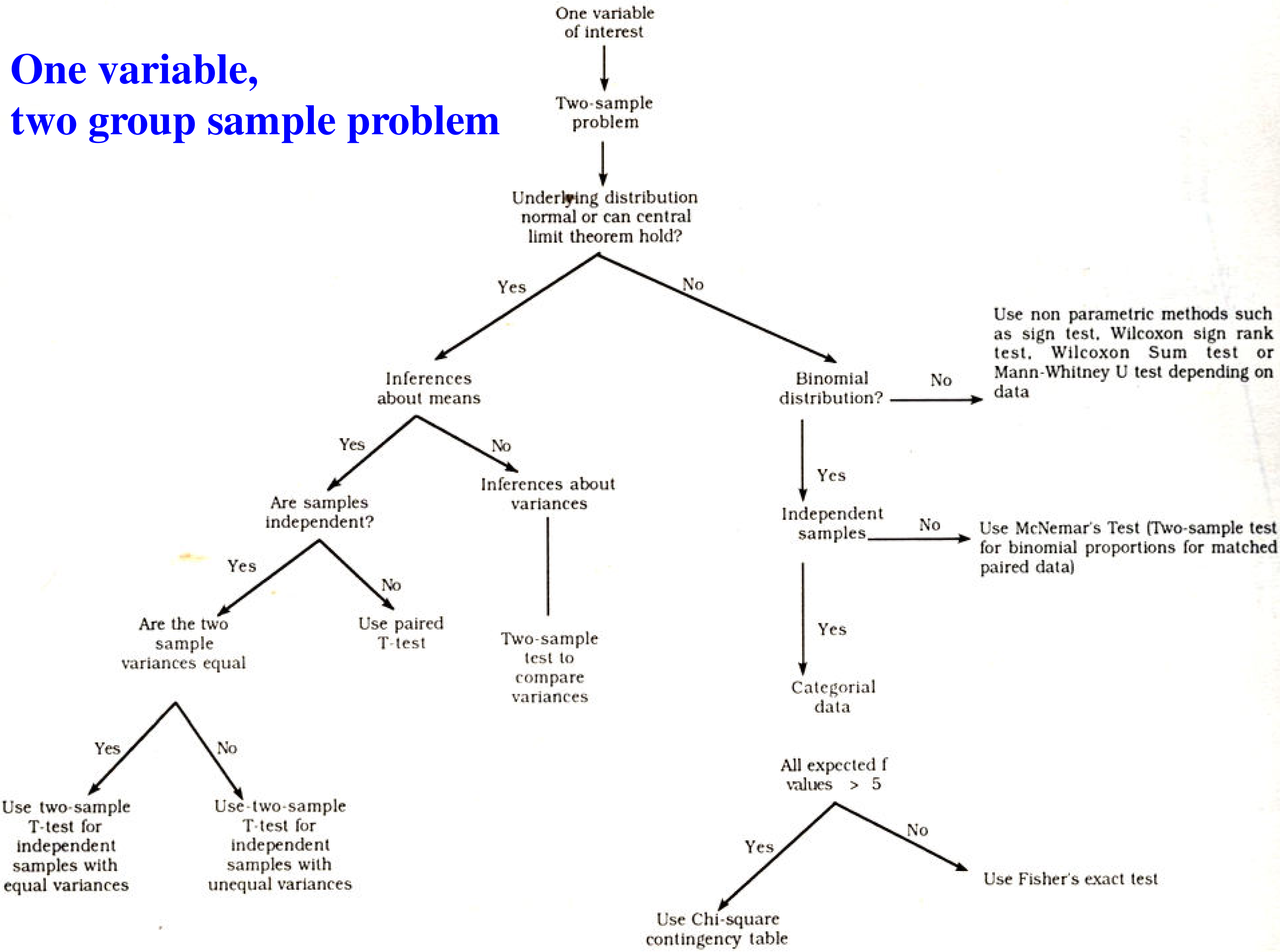
Sum of R_1= 272 and Sum of R_2=163
Difference $T=R_1 \quad R_2$ is 109

The table value of $T_{0.05}$ is 96 , so reject the H_0

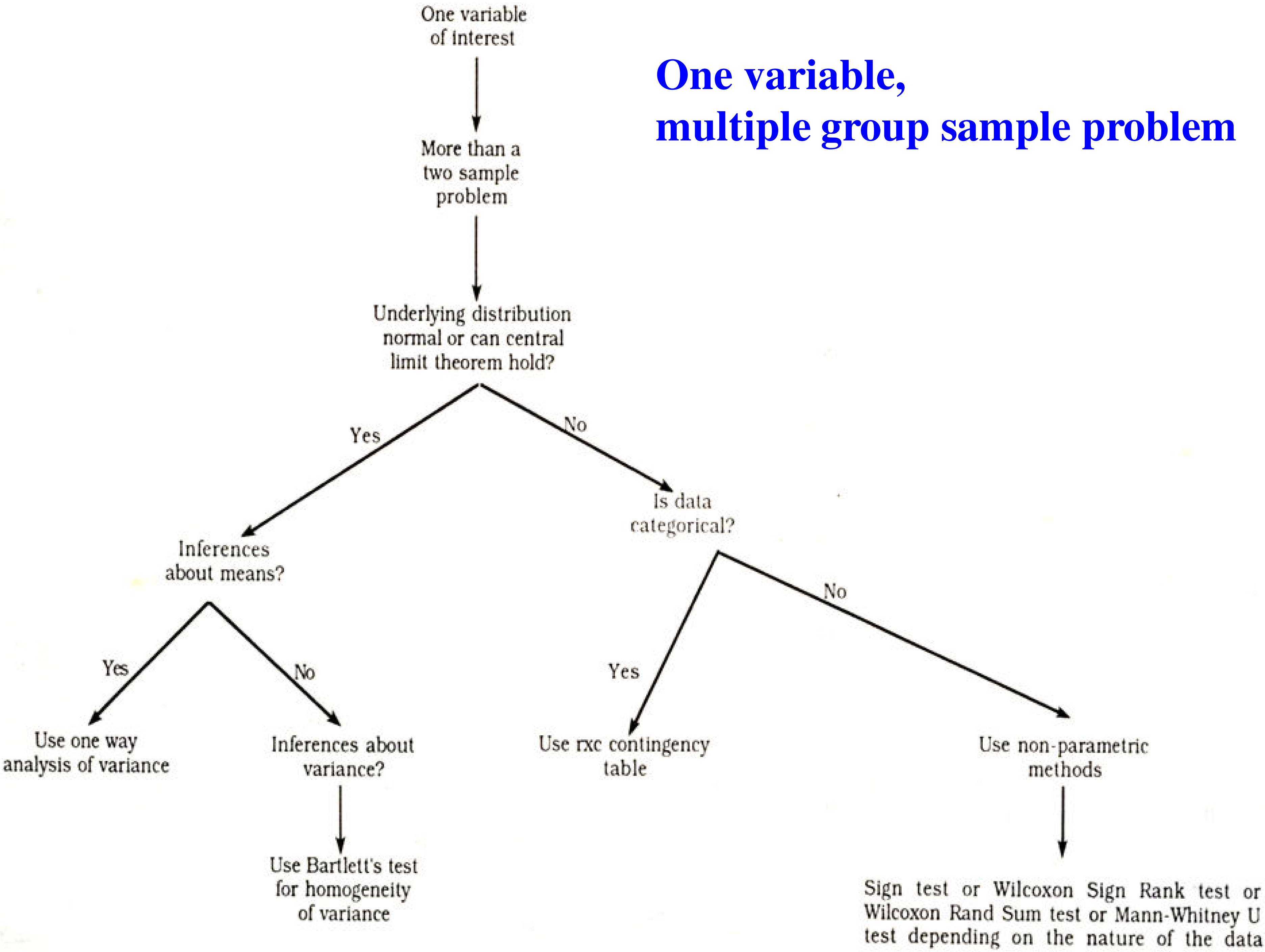
We conclude that weights of children born to the heavy smokers are significantly lower than those of the children born to the non-smokers ($p<0.05$)

Applications of statistical tests in Research Methods

One variable, two group sample problem



One variable, multiple group sample problem



Two variable problem

Interested in relationship b/n two variables

Are both continuous variables

Yes → *Use linear regression*

No → *Is one variable continuous & other categorical*

Yes → *Use analysis of variance Or Logistic regression*

No → *Both variables categorical* → *Use contingency table analysis*

Multiple variable problem

Research interested in relationship
B/n more than two variables

↓

Use multiple regression
Or
Multivariate analysis

Conclusion

Statistics are excellent tools in research data analysis; how ever, if inappropriately used they may make the results of a well conducted research study un-interpretable or meaningless

Bibliography

Biostatistics

Rao K Vishweswara, Ist edition.

Methods in Biostatistics

Dr Mahajan B K, 5th edition.

Essentials of Medical Statistics

Kirkwood Betty R, 1st edition.

Health Research design and Methodology

Okolo Eucharia Nnadi.

Simple Biostaistics

Indrayan,1st edition.

Statistics in Dentistry

Bulman J S


Thank U