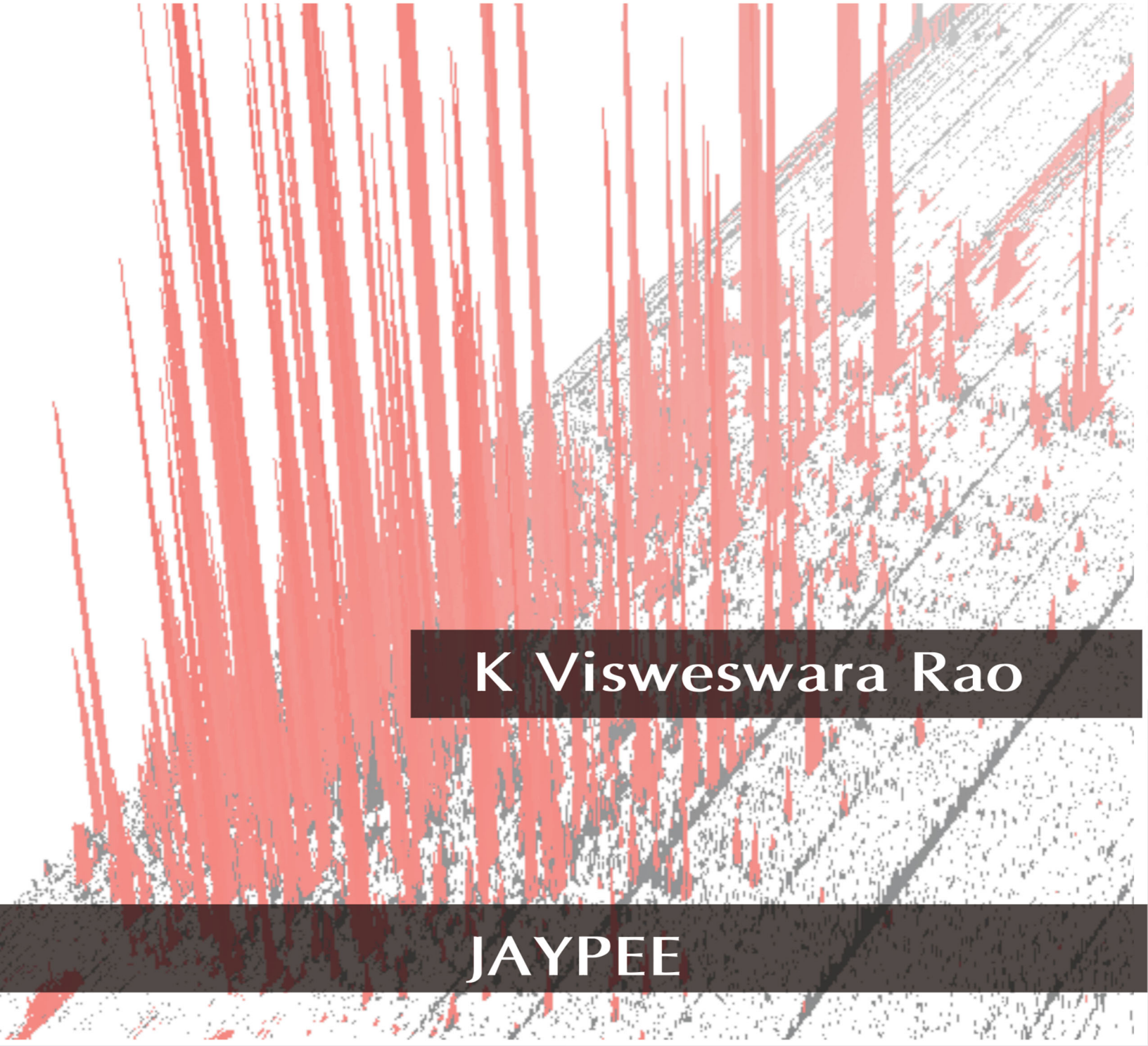
BIOSTATISTICS IN BRIEF MADE EASY®



K Visweswara Rao

JAYPEE

Contents

1. Introduction 1
2. Varieties of Data and Methods of Presentation 5
3. Methods of Sampling Population 8
4. Measures of Averages and Variability 15
5. Probability: Concepts and Utility 24
6. Frequency Distributions: Continuous and Discrete Variables 38
   - 6.1 Normal Distribution / Normal Curve 42
   - 6.2 Student's 't' Distribution 50
   - 6.3 'F' Distribution 54
   - 6.4 Binomial Distribution 56
   - 6.5 Poisson Distribution 57
   - 6.5 a Poisson Distribution: Examples 60
   - 6.6 Chi-square Distribution 62
7. Correlation and Regression 68
8. Quality of the Data and Measurement 77
9. Experimental Designs and Investigations 90
10. Inferential Procedures and Tests of Hypotheses 95
    - 10.1 Comparison of Two Means with Small Samples: 't' Test . 101
    - 10.2 Comparison of Two Means with Large Samples: Normal Test 106
    - 10.3 Comparison of Two Proportions with Large Samples: Normal Test 108
    - 10.4 Comparison of Sample Mean with a Known Value of Population Mean: t Test/ Z Test 110
    - 10.5 Comparison of Sample Proportion with Ideally Needed Values:t Test/Z Test 113
    - 10.6 Differentials of Variances Between Samples: 'F' Test 115
    - 10.7 Comparison of Several Proportions: Chi-square Test 118
11. Analysis of Variance and Covariance Procedures 122
    - 11.1 Analysis of Variance 122
    - 11.2 Analysis of Covariance 143
    - 11.3 Multiple-group Comparision: Test Procedures 153

12. Multivariate Analytical Procedures 164
13. Logistic Regression Procedures: Utility 169
14. Meta-analysis 175
15. Rank Correlation and Partial Correlations 179
15.1 Multiple Correlation Coefficient 186
16. Non-parametric Tests 189
17. Confidence Limits for Variance and Chi-square Test: Examples . 209
18. Vital Statistics and Morbidity 226
19. Life Table Methodology 239
20. Sources and Availability of Vital Statistics 250
21. Sample Size Needs and Determination 255
22. Research Methodology 268
23. Recollection of Concepts of Biostatistics in Brief 270
24. Examples and Exercises 273
25. Useful References 297
Appendices 298
I : Basic Mathematics 298
II : Statistical Tables 306
III : Glossary of Biostatistical Terms 323

Author Index *347*
Subject Index *349*

CHAPTER 4

Measures of Averages and Variability

Information from samples is usually presented in the form of various statistics. Measures of averages and variability or variation are most useful. There are several measures of averages and variation. Most frequently used are presented below.

Measures of Averages

(i) Arithmetic Mean (AM), (ii) Median, and (iii) Mode

(i) **Arithmetic mean (or AM):** The most common one is the arithmetic mean (or mean). If $x_1, x_2, \ldots, x_i, \ldots, x_n$ are the 'n' individual values, the arithmetic mean, $\bar{x}$, is given by the formula as given below:

$$\bar{x} \quad \frac{x_1 \quad x_2 \quad \ldots \quad x_i \quad \ldots \quad x_n}{n} \quad \frac{\sum_{i=1}^{n} x_i}{n} \quad \frac{x_i}{n} \quad \frac{x}{n}$$

where $\sum_{i=1}^{n} x_i \quad x_i \quad x$ is the sum of all the individual values of (x) and 'n' is the sample size.

This is the sample mean ($\bar{x}$). The mean (μ or $\bar{x}$) for the entire population of x_i individuals can also be had with same formula as below:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \frac{x_i}{N} \quad \frac{x}{N} \quad \frac{x_1 \quad x_2 \quad \ldots \quad x_i \quad \ldots \quad x_N}{N}$$

where $\sum_{i=1}^{N} x_i \quad x_i \quad x$ is the sum or total of all the 'N' individual values.

(ii) **Median:** Median is the middle value of the numbers when all the values are arranged in order of magnitude from smallest to the largest. If the number of data points is odd, the value of the middle number is the median. However, if the number of data points is even, the two values of the middle numbers are averaged to obtain the median. When medians are determined from the experimental data points, they may be referred to as sample medians. A population median, usually unknown, is the value of the variable such that 50 per cent of the population values are above (or below) the median value.

JAYPEE BROTHERS

(iii) **Mode:** The mode of a sample is the most frequently observed observation. Some samples have more than one mode. In many situations, this implies that the distribution has two or more peaks. This sometimes happens, when two or more samples are mixed and the pieces of the combined distribution represent the centres of each sample. The formula, given below, can be utilised for the calculation of the mode.

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median})$$

Or

$$\text{Mode} = \text{Mean} - 3\,[\text{Mean} - \text{Median}] \text{ or}$$

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

Measures of Variation

Most commonly used measures of variation are: (i) Range, (ii) Variance and Standard deviation, (iii) Standard Error, (iv) Coefficient of variation, and (v) Mean deviation.

(i) **Range:** The range of a set of values is the difference between the largest and smallest values in the set. Sometimes, the range is given by quoting the smallest and largest values. It is less useful for its intended purpose than are the measures like standard deviation.

(ii) **Variance and Standard Deviation:** The variance (or mean square) for a set of values is defined as the sum of squared deviations of individual values from the mean, divided by the sample size reduced by one. The sample variance is usually denoted by S^2, and the standard deviation defined as the positive square root of S^2, is denoted by 'S' and sometimes by SD. The formula for the sample variance is as follows:
S^2 = Sum of squares of deviations of individual values from mean ÷ (Sample size – 1)

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_n - \bar{x})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum x_i\right)^2}{n-1} \quad \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

where, $\bar{x}$ = Arithmetic Mean; Σx_i^2 = Sum of squares of 'n' individual values; $\sum x_i$ = Total of all the 'n' individual values.

$$SD = \sqrt{S^2} \quad \sqrt{SD^2}$$

JAYPEE BROTHERS

The standard deviation is used more frequently than the variance to express the variability of the observed values. SD and Variance are the indicators of variation between individual values compared to mean. Variance has the mathematical advantage over SD, since the variances can be added or subtracted from each other. Because of this advantage and facility, analysis of variance (ANOVA) has been developed and well-utilised in the field of hypothesis testing and decision making. This facility cannot be availed with SD.

(iii) **Standard Error:** It is a measure of variation in the sample mean. It is used to compare values of means with one another. The formula is:

$$\text{Standard Error} = (SE) = \sqrt{\frac{SD^2}{n}} = \frac{SD}{\sqrt{n}}$$

SE^2 is also known as the *"Variance of arithmetic mean".*

(iv) **Coefficient of Variation:** This is known as CV. It is expressed as a percentage.

$$\text{The formula is } CV = \frac{SD}{Mean} \times 100.$$

The Coefficient of Variation is also called relative variation as SD is represented as percentage of the mean. This helps to compare the degree of variation between various samples. The more diffusely the observations are scattered from their mean, the larger is the CV. It can be stated by comparison that CV is higher in nutrient content of fresh food samples than in dried food samples. CV varies between nutrients/food items.

Examples 1 and 2 on measures of averages and variation are given below:

Example 1

The mineral content of samples of millets and leafy vegetables are given below. Calculate the mean, median, mode, variance, SD, SE and CV for both samples of foods.

Sl. No.	*Foods*	*Mineral Contents*	*Sample size (n)*	*Total* Sx_i	*Sum of squares* Sx_i^2
1.	Millets	1, 5, 4, 2, 4, 3	6	19	71
2.	Leafy vegetables	2, 10, 8, 4, 8, 6	6	38	284

$$\text{A.M.} = \text{Mean} = \bar{x} = \frac{x_i}{n}; \quad SD^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}{n-1}$$

JAYPEE BROTHERS

Calculations are provided below:

S.No.	Foods	Sample size (n)	Total: $Sx_i = Sx$	Sum of squares = $Sx_i^2 = Sx^2$
1.	Millets	6	1+5+4+2+4+3 = 19	$1^2+5^2+4^2+2^2+4^2+3^2 = 71$
2.	Leafy Vegetables	6	2+10+8+4+8+6 = 38	$2^2+10^2+8^2+4^2+8^2+6^2 = 284$

Millets

$$AM = \bar{x} = \frac{\sum x}{n} = \frac{19}{6} = 3.1667$$

$$SD^2 \quad \frac{\sum_{i\ 1}^{n} x_1^2 \quad n\bar{x}^2}{n \quad 1}$$

$$= \frac{71 - 6\,(3.1667)^2}{(6-1)}$$

$$= \frac{71 - 60.1667}{5}$$

$$= \frac{10.8333}{5} = 2.1667$$

$$SD = \sqrt{SD^2} = \sqrt{2.1667}$$

$$= 1.472$$

$$SE = \frac{SD}{\sqrt{n}} \quad \frac{1.472}{\sqrt{6}} = \frac{1.472}{2.449}$$

$$= 0.6008$$

Leafy Vegetables

$$AM = \bar{x} = \frac{38}{6} = 6.3333$$

$$SD^2 = \frac{284 - 6\,(6.3333)^2}{6-1}$$

$$= \frac{284 - 240.6641}{6-1}$$

$$= \frac{43.3359}{5}$$

$$= 8.6672$$

$$SD = \sqrt{SD^2} = \sqrt{8.6672} = 2.944$$

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

$$SE = \frac{2.944}{\sqrt{6}} = \frac{2.944}{2.449}$$

$$= 1.2016$$

Coefficient of Variation

$$CV \quad \frac{SD}{Mean} \times 100 \quad \frac{1.472}{3.1667} \times 100$$

$$= 46.5\%$$

$$CV \quad \frac{SD}{Mean} \quad (100) \quad \frac{2.944}{6.334} \times 100$$

$$= 46.5\%$$

Median: Value of the variable below and above of which there are 50% of the observations:

	Millets	*Leafy Vegetables*
Values in order are:	1, 2, 3, 4, 4, 5	2, 4, 6, 8, 8, 10
Median:	(3 + 4) / 2 = 3.5	(6 + 8) / 2 = 7.0

JAYPEE BROTHERS

Mode: Value occurring most frequently is 4 in millets and 8 in leafy vegetables. It can also be calculated using the formula,

Mode = Mean – 3 (Mean – Median) also.

Measures of averages and variation in mineral contents are as below:

Sl. No.	*Food item*	*Sample size*	*Mean*	*Median*	*Mode*	SD^2	*SD*	*SE*	*CV*
1.	Millets	6	3.1667	3.5	4.0	2.1667	1.4720	0.6008	46.5%
2.	Leafy Vegetables	6	6.3334	7.0	8.0	8.6672	2.9440	1.2016	46.5%

Comment: Though CV is same, mean, median, mode, variance, SD and SE are different between millets and leafy vegetables. It is worth testing the significance of difference with an appropriate test of significance.

Example 2: Calculations of arithmetic mean, median, mode, variance, SD, range, SE and CV are given (Table 4.1) for the protein content of the 25 pulse samples.

Table 4.1: Calculation of SD^2

Protein (gm%)(x_i)	$(x_i - \bar{x}) = d$	$(x_i - \bar{x})^2 = d^2$
8	(8 – 20) = – 12	144
9	(9 – 20) = – 11	121
10	(10 – 20) = – 10	100
10	(10 – 20) = – 10	100
10	(10 – 20) = – 10	100
10	(10 – 20) = – 10	100
10	(10 – 20) = – 10	100
10	(10 – 20) = – 10	100
11	(11 – 20) = – 9	81
12	(12 – 20) = – 8	64
19	(19 – 20) = – 1	1
20	(20 – 20) = 0	0
20	(20 – 20) = 0	0
20	(20 – 20) = 0	0
21	(21 – 20) = 1	1
28	(28 – 20) = 8	64
29	(29 – 20) = 9	81
30	(30 – 20) = 10	100
30	(30 – 20) = 10	100
30	(30 – 20) = 10	100
30	(30 – 20) = 10	100
30	(30 – 20) = 10	100
30	(30 – 20) = 10	100
31	(31 – 20) = 11	121
32	(32 – 20) = 12	144
Total 500		2022

JAYPEE BROTHERS

Sample size = n = 25

$$\text{Mean} = \bar{x} \quad \frac{\bar{x}_i}{n} \quad \frac{8 \quad 9 \quad 10 \quad \ldots \quad 32}{25} \quad \frac{500}{25} \quad 20$$

Median value is the value pertaining to the individual sample of 13, in order. Median value = 20; Mode is the value occurring most frequently; the formula is:

Mode = Mean – 3 (Mean – Median) i.e., Mode = 20.

$$\text{Variance} = \frac{x_i \quad n\bar{x}^2}{n \quad 1} \quad \frac{2022}{24} \quad 84.25$$

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{SD^2} = \sqrt{84.25} = 9.1788 = 9.18$$

$$\text{SE} = \sqrt{\frac{SD^2}{n}} \quad \sqrt{\frac{84.25}{25}} \quad \sqrt{3.37} \quad 1.8357 \quad 1.84$$

$$\text{CV} = \frac{SD}{\text{Mean}} \times 100 = \frac{9.1788}{20} \times 100 = 45.89\%$$

Range = Highest Value – Lowest value = 32 – 8 (or 24)

Range is also stated as 8 – 32, i.e. smallest value to highest value.

(v) Mean Deviation

The mean deviation is the arithmetic mean of the absolute values of the deviations of individual values from some average like median or mode or arithmetic mean. The formula is given below:

$$\text{Mean deviation: } \frac{\sum_{i\ 1}^{n} |x_i \quad \bar{x}|}{n}$$

where x_i is the "ith" individual value, $\bar{x}$ is the arithmetic mean value and n is the sample size.

This measure of variation ignores the sign of the deviations of individual values from the arithmetic mean. This measure of variation is hardly utilised in literature.

Standard deviation is better defined and is accepted as an adequate measure of variability. It is the theoretically well-defined and practically well utilised.

Example 2A: Calculate the mean deviation for the mineral content of millets in Example 1:

Mineral content: 1, 5, 4, 2, 4, 3

Answer:

$$\text{AM} \quad \bar{x} \quad \frac{\sum_{i\ 1}^{n} x_i}{n} \quad \frac{x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_6}{6}$$

JAYBEE BROTHERS

$$\frac{1 \quad 5 \quad 4 \quad 2 \quad 4 \quad 3}{6} \quad \frac{19}{6} \quad 3.1667$$

Mean deviation from AM $\quad \frac{|x_i \quad \bar{x}|}{n}$

$$= \frac{|x_1 \quad \bar{x}| \quad |x_2 \quad x| \quad \ldots \quad |x_i \quad \bar{x}| \quad \ldots \quad |x_n \quad \bar{x}|}{n}$$

$$= \frac{\sum_{i=1}^{6} |x_i - \bar{x}|}{6}$$

$$\frac{|(1 \quad 3.1667)| \quad |5 \quad 3.1667)| \quad |4 \quad 3.1667)| \quad |2 \quad 3.1667)| \quad |4 \quad 3.1667)| \quad |3 \quad 3.1667)|}{6}$$

$$\frac{|\ 2.1667| \quad |1.8333| \quad |0.8333| \quad |\ 1.1667| \quad |0.8333| \quad |\ 0.1667|}{6}$$

$$= \frac{2.1667 + 1.8333 + 0.8333 + 1.1667 + 0.8333 + 0.1667}{6} = \frac{7.0}{6} = 1.1667$$

Example 2B: Check whether the answers given for AM and mean deviation are correct with the protein content of five samples of wheat studied.

Protein content (gm): 10.0, 15.0, 12.0, 13.0, 10.0

$$\text{AM} = \text{Mean} = \bar{x} = \frac{60}{5} = 12.0$$

$$10 - 12.0 \ + 15 - 12.0 \ + \ 12 - 12.0 \ + \ 13 - 12.0 \ + 10 - 12.0$$

$$\text{Mean deviation} = \frac{|10 - 12.0| + |15 - 12.0| + |12 - 12.0| + |13 - 12.0| + |10 - 12.0|}{5}$$

$$= \frac{|-2.0| + |3.0| + |0| + |1.0| + |-2.0|}{5} = \frac{2.0 + 3.0 + 0 + 1.0 + 2.0}{5} = \frac{8}{5} = 1.6$$

Example 2C: Calculate the mean, median, mode, range, SD^2, SD, CV and SE for the systolic BP (in mm Hg) of the two groups of subjects studied.

Group I	128	132	129	130	131	$n_1 = 5$	$n_1 = n_2 = n = 5$
Group II	140	150	120	130	110	$n_2 = 5$	

JAYPEE BROTHERS

S.No.	*Group I*	*Ranks*	$(x_i - x)$	$(x_i - x)^2$
1	128	(1)	128 – 130 = –2	4
2	132	(5)	132 – 130 = 2	4
3	129	(2)	129 – 130 = – 1	1
4	130	(3)	130 – 130 = 0	0
5	131	(4)	131 – 130 = 1	1
				10

Number (n) : 5 Mean deviation $\sum (x_i - \bar{x})^2 = 10$

Total : 650 = $|x_i \quad \bar{x}|$ n $SD^2 = (x_1 \quad x)^2$ [n 1]

AM : 130 = [| – 2 | + | 2 | + | – 1| + | 0 | + | 1 |] ÷ 5 = 10 ÷ 4 = 2.5

Median : 130 = [2 + 2 + 1 + 0 + 1] ÷ 5 $SD = \sqrt{2.5} = 1.5811$

Mode : 130 = 6/5 = 1.2 $SE^2 = \frac{SD^2}{n} = \frac{2.5}{5} = 0.5$

Range : 128-132 $SE = \sqrt{0.5}$ 0.7071

CV (SD $\bar{x}$) 100
= (1.5811÷130)×100
= 1.2162%

S.No.	*Group II*	*Ranks*	$(y - \bar{y})$	$(y - \bar{y})^2$
1	140	(4)	140 – 130 = 10	100
2	150	(5)	150 – 130 = 20	400
3	120	(2)	120 – 130 = –10	100
4	130	(3)	130 – 130 = 0	0
5	110	(1)	110 – 130 = –20	400
				1000

Number : 5 Mean deviation

Total : 650 $= \frac{|y_i \quad \bar{y}|}{n}$

AM : 130 $= \frac{10 + 20 + 10 + 0 + 20}{5}$

Median : 130

Mode : 130

Range : 110-150 = 60 ÷ 5 = 12

JAYPEE BROTHERS

$$\sum (y_i - \bar{y})^2 = 1000$$

$$SD^2 = \sum (y_1 - \bar{y})^2 \div (n - 1) = 1000 \,/\, 4 = 250$$

$$SD = \sqrt{SD^2} = \sqrt{250} = 15.8114$$

$$SE^2 = \frac{SD^2}{n} = \frac{250}{5} = 50$$

$$SE = \sqrt{SE^2} = \sqrt{50} = 7.0711$$

$$CV = \frac{SD}{AM} \times 100 = \frac{15.8114}{130} \times 100 = 12.1626$$

Measures of averages and variation are as below:

Groups	*n*	*AM*	*Median*	*Mode*	*Range*	*SD²*	*SE²*	*SD*	*SE*	*CV*
i	5	130	130	130	128 – 132	2.5	0.5	1.5811	0.7071	1.2069
ii	5	130	130	130	110-150	250	50	15.811	7.0711	12.1626

Inference: The groups I and II have equal measures of averages, but measures of variation are unequal and different.

JAYPEE BROTHERS

BIOSTATISTICS IN BRIEF
MADE EASY®

Biostatistics or Statistics has been felt by the students and readers as the subject of difficult exposure. The author attempted the presentation of various topics with a simple and easy clarity covering with useful examples and exercises. The topics covered are: Varieties of Data, Data Quality, Methods of Data Generation and Presentation, Measures of Averages and Variation, Graphic Depictions, Concepts and Utility of Probability, Frequency Distributions: Normal, Binomial, Poisson, Student's 't' and Chi-square Test, Measures of Inter-relationship and Predictions, Tests of Inference and Decision Making, Multivariate Analytical Procedures and Logistic Regression, Meta-analysis and Research Methodology, Vital and Health Statistics, and Designs of Experiments and Investigations.

These topics and the needful basic principles of mathematics and glossary of biostatistical terms provided in the book are useful as a good reference for perusal. Appendix provided covers the useful reference tables for the analysis and interpretation of the generated data.

The Univariate and Multivariate Analytical Procedures covered with examples and exercises are presented with clarity in an easy, simple and understandable way. One will have an impression of the subject with improved clarities.

K Visweswara Rao holds BA in Statistics (Andhra University), MA in Statistics (Delhi University), MS Hygiene in the field of Biostatistics (Harvard University, USA), and PhD in Statistics (Osmania University, India).

He has good research collaboration and contributions in the fields of Biostatistics, Nutrition, Physical Anthropology, and Vital Statistics. To his credit are several articles in scientific journals, research bulletins and books.

He has brought out as the co-author or editor, the following books: *Diet Atlas of India (1969, 1972), Statistics in Health and Nutrition (1990), Biostatistics: A Manual of Statistical Methods for Use in Health, Nutrition and Anthropology (1996),* and *Biostatistics in Brief Made Easy (2009)*.

He was with the WHO of the United Nations as a consultant in Malaysia and the Philippines and as a Fellow in USA, Western Europe, Central America and West Indies. He was conferred as a Fellow of the Institute of Statisticians (England), Indian Society for Medical Statistics, and the United Writers Association of India.

He was a recipient of Harvard Gold Key (USA), Prof Rangachari's Gold Medal, Prof BG Prasad's Gold Medal and Padmabushan Prof PV Sukatme's Gold Medal. He retired as a Senior Deputy Director from the Department of Statistics, National Institute of Nutrition, Indian Council of Medical Research, Hyderabad, India. He is now an honorary Emeritus Scientist at the Osmania University in the Department of Microbiology and Statistics.

Available at all medical book stores
or buy directly from Jaypee Brothers through online shopping
at www.jaypeebrothers.com

or call + 91-11-32558559

JAYPEE BROTHERS
Medical Publishers (P) Ltd.
www.jaypeebrothers.com

ISBN 978-81-8448-760-2