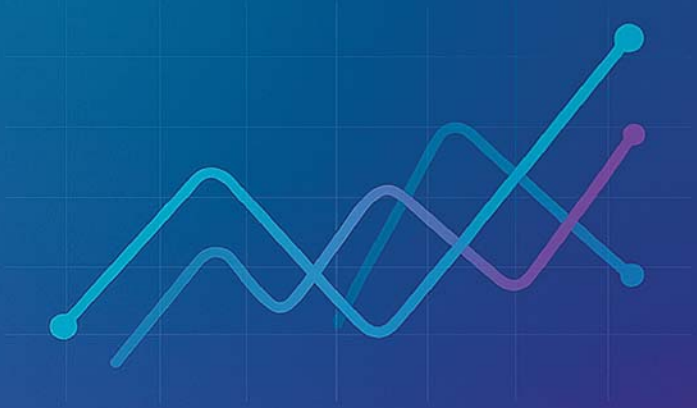
m, for detailed information on medical statistics books, visit our website www.ejaypee.com, for detailed information on medical statistics books

11th Edition

Mahajan's Methods in Biostatistics

for Medical Students and Research Workers



Revised and Edited by
Yogesh K Singhal

diginerve®
A Jaypee Initiative
• Video Lectures • Dr. Wise AI Chatbot
• Question Bank • Notes
Download App

JAYPEE

Contents

SECTION 1: STATISTICS

1. **History and Evolution of Statistics 3**
   - Origins and Evolution of Statistics 3
   - Ancient Civilizations (3000 BCE–500 CE) 3
   - Middle Ages to Enlightenment (500–1800 CE) 4
   - Development of Modern Statistics (1800–1900 CE) 4
   - 20th Century Advances (1900–2000 CE) 5
   - Contemporary Statistics (2000 CE)–Present) 5
2. **Introduction 15**
   - Biostatistics 16
3. **Data Types, Sources, Collection and Presentation 25**
   - Types of Data 25
   - Qualitative and Quantitative Data 26
   - Scales of Measurement 29
   - Primary and Secondary Data 30
   - Sources of Data 32
   - Data from Multiple Sources 33
   - Collection of Medical Data 35
   - Quality of Data 39
   - Presentation of Data 42
   - Presentation or Illustration of Qualitative Data 50
   - Presentation or Illustration of Quantitative Data 55
4. **Sampling and Sampling Variability 70**
   - Definitions 70
   - Sample Characteristics 72
   - Sample Size Determination 73
   - Sampling Techniques 80
   - Probability Sampling Techniques 83
   - Nonprobability Sampling Techniques 96
   - Sampling Variability and Significance 98
   - Sampling Distribution 100
5. **Measures of Central Tendency and Location 103**
   - Measures of Central Tendency: Averages 103
   - Measures of Location: Percentiles 117

6. Variability (Dispersion) and its Measures 128
- Types of Variability 128
- Measures of Variability 132
- Appendix on Chapters 5 and 6 146

7. Normal Distributions and Normal Curve 163
- Normal Distribution 163
- Demonstration of a Normal Distribution 165
- Normal Curve 167
- Standard Normal Variate (Z) 169
- Standard Normal Curve 170
- Asymmetrical Distributions 173
- Mathematical Explanation of Normal Distribution 176

8. Significance and Test of Significance..... 181
- Estimation of Population Parameter 181
- Tests of Significance 195
- Significance of Difference in Means 198
- Standard Error of Mean ($Se_{\bar{x}}$) 198
- Significance of Difference Between Means of Large Samples By Z Test 204
- Significance of Difference Between Means of Small Samples by Student's T-Test 211
- Significance of Difference Between Means of More Than Two Samples by F-Test and ANOVA 222
- Analysis of Variance Test 223
- Significance of Difference In Proportion of Large Samples 226
- Standard Error of Proportion 228

9. Nonparametric Tests 245
- The Chi-Square Test 245
- Other Nonparametric Tests 261
- Tests for Making Inferences from Binomial and Ordinal Data 261

10. Correlation and Regression 281
- Measures of Relationship Between Continuous Variables 281
- Types of Correlation 283
- Characteristics of a Relationship 286
- Correlation Coefficient 286
- Calculation of Correlation Coefficient from Ungrouped Series 287
- Calculation of Correlation Coefficient From Grouped **Series 292**
- Regression 294
- Calculation of Regression Coefficient 296

11. Probability (Chance)..... 305
- Terms Used in Probability 306
- Basics on Probability 307
- Laws of Probability 309
- Conditional Probability 319
- Use of Probability in Assessing Screening Test 320
- Other Types of Standard Distribution 324

12. Predictive Modeling..... 329
- Basics of Predictive Modeling 329
- Techniques Used for Predictive Modeling 330
- Models of Predictive Modeling 330
- Benefits Offered by Predictive Modeling 331
- Application of Predictive Modeling in Healthcare: A Simplistic Example for Better Understanding of the Concepts 331

SECTION 2: POPULATION SCIENCE

13. Demography and Vital Statistics 339
- Collection of Demographic Data 339
- Medical Certification of the Cause of Death 348
- Compilation and Presentation 366

14. Measures of Population and Vital Statistics 368
- Measures of Population 368
- Measures of Vital Statistics 377
- Various Indicators of Population and Health 380

15. Life Table 407
- Uses and Application 407
- Construction of a Life Table 408
- Modified Life Table 416
- Kaplan–Meier Method of Survival Analysis 421

SECTION 3: RESEARCH METHODOLOGY

16. Designing and Methodology of Quantitative Medical Research Study..... 431
- Epidemiological Study Designs in Quantitative Research 431
- Steps in Methodology and Designing 436
- Presentation Format 441

17. Data Analysis in Quantitative Medical Research Study..... 443
- Estimation of the Risk of Developing a Disease 443
- Estimation of the Reliability of Test Result 446
- Estimation of the Performance of a Test 449
- Prediction from Test Result 451
- Calculating the Various Rates 452
- Estimating Strength of Evidence 452

18. Designing and Conducting Qualitative Medical Research Studies 457
- Approaches to Qualitative Research 457
- Planning a Qualitative Research Study 459
- Data Collection Approaches in Qualitative Research 462
- Data Analysis in Qualitative Research 464
- Data Analysis Software in Qualitative Research 467
- Ensuring Rigor in Qualitative Research 467
- Reporting Qualitative Research 468
- Uses of Qualitative Research in Health Care 469

19. Basics of Research Reporting 471
- Scientific Documents Before Initiating Research 471
- Scientific Documents After Completion of Research 477

SECTION 4: COMPUTER APPLICATION IN MEDICAL SCIENCE

20. Computers in Medicine 485
- Application of Computers in Biostatistics 486
- Application of Computers in Medical Education and Training 486
- Application of Computers In Community or Public Health Care and Management 486
- Application of Computers in Healthcare Establishments 488
- Application of Computers in Research 489
- Basics of Referencing Software: Zotero 490

21. Computer Application for Citation Management: Zotero .. 490
- Basics of Referencing Software: Zotero 490
- Using Zotero 492

22. Computer Application for Data Collection: EpiCollect..... 500
- Basics of Epicollect 500
- Epicollect5 Portal 502
- Data Collection Using Epicollect5 502

23. Computer Application for Data Entry and Analysis: Microsoft Excel 508
- Basics of Microsoft Excel 508
- Data Entry in MS Excel 509
- Data Analysis in MS Excel 509
- Preparing Graphs in Microsoft Excel 513

24. Computer Application for Biostatistical Analysis: Statistical Package for the Social Sciences 518
- Basics of Statistical Package for the Social Sciences 518
- Statistical Analysis in SPSS 530

Appendices 567
- *Appendix 1:* Tables of Unit Normal Distribution (UND) for *Z* Values of α and β 569
- *Appendix 2:* Table of '*t*' and Probability of Larger Value of '*t*' 570
- *Appendix 3:* Variance Ratio 5% and Points of *e* 571
- *Appendix 4:* Table of χ^2 and Probability (*p*) 573
- *Appendix 5:* The Correlation Coefficient Table Probability (*p*) 574
- *Appendix 6:* Random Numbers 575
- *Appendix 7:* Critical Values of the Mann–Whitney U 578
- *Appendix 8:* Critical Values of the Wilcoxon Signed Ranks Test 582
- *Appendix 9:* Upper Critical Values of the Kruskal–Wallis Test 583

Bibliography 587

Index 589

5 CHAPTER

Measures of Central Tendency and Location

Having learnt the methods of collection and presentation of data, it is important to understand the application of mathematical techniques involved in analysis and interpretation of data, such as calculation of averages, percentiles, standard deviation, standard error, correlation and regression coefficients, etc. They are mostly used to assess what is normal in health as in the case of height, weight, blood pressure, cholesterol in blood, pulse rate, red blood cell (RBC) count, etc., which is expressed as the average or central value of the group, most commonly used average being the mean. However, normal is not the mean or central value but the accepted range of variation on either side of the average of mean, e.g., normal blood pressure is not the mean but mean 120 ± 20, i.e., a range between 100 and 140. Chances of even higher and lower observations being normal are there. This range of variation is known as the Confidence Interval, which will be discussed in Chapter 8.

MEASURES OF CENTRAL TENDENCY: AVERAGES

Information from a series of observations presented by rank in a sample or frequency distribution table is summarized by an observer to get answers to the following three questions:

1. What is its average or central value?
2. How are the other values dispersed around this value or what is the degree of scatter?
3. What is the shape of the distribution? Is it normal?

Average value of a characteristic is the one central value around which all other observations are dispersed. In any large series, nearly 50% observations lie above while the remaining 50% lie below the central value. It indicates how the values lie near the center. In other words, it is a measure of central tendency or concentration of all other observations around the central value. Thus, an average value helps:

JAYPEE BROTHERS

- Firstly, to find the central value around which most of the normal observations lie. Few of the values may be too large or too small and lie far away at both ends. These are known as outliers.
- Secondly, to find which group is better off by comparing the average of one group with that of the other, e.g., average body weight of one group of children is higher than another group; average daily attendance of one hospital is higher than that of another and so on. After finding the difference, one may reason out why in one group it is more than that in the other.

Average is a general term which describes the center of a series. There are three common types of averages or measures of central position or central tendency—mean, median and mode. They are summary indices describing the 'central point or the most characteristic value' of a set of measurements.

Mean

This measure implies arithmetic average or arithmetic mean which is obtained by summing up all the observations and dividing the total by the number of observations.

Example

Erythrocyte sedimentation rates (ESRs) of 7 subjects are 7, 5, 3, 4, 6, 4, 5. Calculate the mean.

$$\text{Mean} = \frac{7+5+3+4+6+4+5}{7} = \frac{34}{7} = 4.86$$

It is the one central value that is most commonly used as average in statistical methods.

Median

When all the observations of a variable are arranged in either ascending or descending order, the middle observation is known as median. It implies the mid-value of a series. ESRs of seven subjects in the above example are arranged in ascending order: 3, 4, 4, (5), 5, 6, 7. The 4th observation (5) is the median in this series. Thus, in this example, median is almost equal to the mean. Consider another example of 7 observations in absenteeism of school children in the series: 4, 6, 8, (10), 12, 14, 32.

- Mean 86/7 = 12.3
- Median value = 10

JAYPEE BROTHERS

In this case, mean value gives a distorted result as one observation 32 is too large, so the mean as a measure of central tendency should not always be considered appropriate. In a case with outliers, to have a better idea of average, one should ignore outliers, i.e., unduly high or low observations, such as 32 in the above example. Mean of the remaining observations will be 54/6 = 9.0 which is much closer to the median, i.e., 10, rather than the mean 12.3 calculated with all seven observations.

Median, therefore, is a better indicator of central value when one or more of the lowest or the highest observations are wide apart or not so evenly distributed, e.g.

	Year	
	1974	*1977*
Mean age at marriage	16.40	15.70
Median age at marriage	15.70	16.00

As per mean, the age at marriage in 1977, has decreased while as per median it has increased which is nearer the truth.

Another good example is the duration of hospital stay. In such cases median is a better indicator because the stay may be unduly long in some cases.

Mode

This is the most frequently occurring observation in a series, i.e., the most common value of the observation or one of highest frequency. In the previous example of ESRs of seven subjects as: 3, 4, 4, 5, 5, 6, 7, there are two values which are occurring twice and thus more frequent than the other values. Hence, mode will be an average of these two values, i.e., 4 + 5/2 = 9/2 = 4.5.

Mode is rarely used in medical studies. However, it is useful in certain cases where it is essential to know the value of the observation that influences the dataset the most. To study sexual behavior for planning strategies for prevention of HIV infection, mode is more useful than mean for knowing the age of sexual debut as this indicates the age at which most individuals start their sex behavior. Hence preventive strategies should be targeted at population of age group indicated by mode. Mean age may become too late for most of the population thus making the strategy ineffective.

JAYPEE BROTHERS

Out of the three measures of central tendency, mean is better and utilized more often because it uses all the observations in the data and is further used in the tests of significance. However, it is applicable most effectively for data following normal distribution. The concept of normal distribution has been explained later in Chapter 7.

Calculation of Mean

A series of observations is indicated by the letter X and individual observations by X_1, X_2, ..., X_n. The mean of series is denoted by $\overline{X}$ ($\overline{X}$ bar); the number of observations by n and the sum of observations by Σ (sigma) which means sum-up or add-up of all the results.

Formula for calculation of mean is:

$$Mean = \frac{Total\ or\ sum\ of\ the\ observations}{Number\ of\ observations}$$

$$\overline{X} = \frac{(X_1 + X_2 \cdots + X_n)}{n} = \frac{\Sigma X}{n}$$

The mean is calculated by different methods in two types of series, ungrouped and grouped.

Ungrouped Series

In such series, the number of observations is small and there are two methods for calculating the mean. The choice depends upon the size of observations in the series.

- When the dataset of observations is small in size, simply add them up and divide by the number of observations.

  Example: Tuberculin test reaction of 10 boys is arranged in ascending order being measured in millimeters. Find the mean size of reaction.

  3, 5, 7, 7, 8, 8, 9, 10, 11, 12

  $$\text{Mean or } \overline{X} = \frac{\Sigma X}{n}$$

  $$\text{Mean} = \frac{3 + 5 + 7 + 7 + 8 + 8 + 9 + 10 + 11 + 12}{10}$$

  $$= \frac{80}{10} = 8 \text{ mm}$$

  When the values of individual observations are large in size, either use a calculator for addition and proceed as direct method

JAYPEE BROTHERS

described above, or assume an arbitrary mean or working origin (w) to reduce the value of each observation. Find out the difference between each measurement and the assumed mean or working origin, sum up the differences and divide the total by 'n', i.e., the number of observations. This gives the mean of differences or deviations denoted by $\bar{x}$ (small *x* bar).

$$\text{Now} \quad \bar{x} = \frac{(X_1 - w + X_2 - w, \cdots, + X_n - w)}{n}$$

$$= \frac{\Sigma(X - w)}{n} = \frac{\Sigma x}{n}$$

Here, X stands for the original series of observations and x for the differences between X and the working origin or assumed mean.
Mean of the original observations or $\bar{X} = w + \bar{x}$
Actual mean = Assumed mean + (sum of new values/number of observations)

- **Example No. 1:** Find the mean incubation period of 9 cases of a disease given below. In this case 20 may be taken as the working origin or working mean and 0 may be put opposite that.

X	*X – w*	*x*
23	23–20	+3
22	22–20	+2
20	20–20	0
24	24–20	+4
16	16–20	–4
17	17–20	–3
18	18–20	–2
19	10–20	–1
21	21–20	+1
180		+ 10 –10 = 0

– *By direct method*:
Number of observations $n = 9$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{180}{9} = 20$$

JAYPEE BROTHERS

– *By assumed mean (w) method*:

$$\bar{x} = \frac{\Sigma(X - w)}{n} = \frac{10 - 10}{9} = 0$$

so, $\bar{X} = w + \bar{x} = 20 + 0 = 20$

- **Example No. 2:** Heights in centimeters for 7 school children are given below. Find the mean. 140 may be taken as the working origin to calculate the mean height.

X	X – w	x
148	148–140	8
143	143–140	3
160	160–140	20
152	152–140	12
157	157–140	17
150	150–140	10
155	155–140	15
1065		85

– *By direct method*:
Number of observations n = 7

$$\bar{X} = \frac{\Sigma X}{n} = \frac{1065}{7} = 152.1$$

– *By assumed mean (w) method*:

$$\bar{x} = \frac{\Sigma(X - w)}{n} = \frac{85}{7} = 12.1$$

$\bar{X} = w + \bar{x} = 140 + 12.1 = 152.1$

If $w = 150$, calculations will be still easier.

Then, $$\bar{x} = \frac{\Sigma(X - 150)}{n}$$

$$\bar{x} = \frac{-2 - 7 + 10 + 2 + 7 + 0 + 5}{7} = \frac{15}{7} = 2.1$$

$\bar{X} = w + \bar{x} = 150 + 2.1 = 152.1$

- **Example No. 3:** The heights of 7 boys aged 14 years are given in cm with fractions in the table that follows. Find the mean.

JAYPEE BROTHERS

Take 0 for any convenient working origin or assumed mean, such as 150 in the example below:

X (height)	*X – w = x*	*Sum of x*
145.8	–4.2	
14.9	–3.1	– 8.5
148.8	–1.2	
150.0(w)	0	
152.1	+2.1	
153.6	+3.6	+ 12.7
157.0	+7.0	

$$\bar{x} = \frac{\Sigma(X - w)}{n}$$

$$= \frac{12.7 - 8.5}{7} = \frac{4.2}{7} = 0.6$$

$$\overline{X} = w + \bar{x} = 150 + 0.6 = 150.6 \text{ cm}$$

Grouped Series

When the number of observations is large, the data are arranged in groups and frequency distribution table is prepared first. In all grouped series, only weighted mean has to be found and not the ordinary mean. Make convenient groups as per the characteristic values and prepare the frequency distribution table. The groups may be either of a single value or a group of values of the variable. Find the value or weight, contributed by each group separately and multiply the mid value of group with its frequency. Add these product values and then divide by the total number of observations in the sample. This mean is called *weighted mean, or grand mean, or mean of means*.

- Grouped data with single value in each group
  - *Direct method:* This method is similar to the direct method of grouped data, but in this case the value (x) of each group is multiplied by its frequency of occurrence (f) and the products (fx) are added together ∑fx and finally divided by the total number of observation (N) to give the mean ∑fX/N.
  - *Short cut or code method:* Sometimes the values of variables or the frequencies in each group may be large, making the

JAYPEE BROTHERS

product fx too large and difficult to manage. In such cases, short cut method may be used to reduce the figures by taking an assumed mean and subtracting it from the value of observations as explained in case of ungrouped data. This new figure is multiplied with the frequency and divided by total number of observations. The figure thus obtained is added to the assumed mean to get the actual mean.

- Grouped data with arrange of values
  - *Middle point method:* When the range of values of observations is wide the data is tabulated as a frequency distribution table with the values grouped with class intervals and number of observations displayed against each class. To calculate mean in such cases it is assumed that the frequencies in a class is uniformly distributed on either side of the mean. The midpoint of each class is determined and is multiplied with the respective frequency and the mean is calculated as explained in direct method for grouped data with single value in each group.
  - *Alternative method:* This is similar to the code method described above by taking an assumed mean or a working group. It is explained below with examples.
- **Example No. 1:** In an office the average income of 10 subordinate staff is ₹40,000 per month and that of 20 superior staff is ₹60,000 per month, calculate the weighted mean or average income of all staff. One may consider the average income of all staff as ₹50,000—by adding the averages of ₹40,000 and ₹60,000 and then dividing by 2. It is however advisable to compute after taking into account the incomes or weights contributed by each group of workers as shown below:

  Total income of 10 subordinate staff
  = *mean income of the group* × *frequency*
  $= \overline{X}_1 f_1 = 40{,}000 \times 10 = ₹4{,}00{,}000$
  Total income of 20 superior staff
  = *mean income of the group* × *frequency*
  $= \overline{X}_2 f_2 = 60{,}000 \times 20 = ₹12{,}00000$
  Total income of all the 30 staff
  $= \overline{X}_1 f_1 + \overline{X}_2 f_2 = 4{,}00{,}000 + 12{,}00000 = ₹16{,}00000$
  The weighted mean income of all staff

$$= \frac{\Sigma f \overline{X}}{n} = \frac{16{,}00000}{30} = ₹53333.33$$

JAYPEE BROTHERS

Weighted mean is computed in the same way when an event in a qualitative data is expressed in percentages.

- **Example No. 2:** Calculate overall fatality rate of a disease from the age wise fatality rate given below:

Age group in years	*No. of cases*	*Fatality rate percent*
0–1	150	35.33
2–4	304	21.38
5–9	421	16.86
Above 9	170	14.17

It will be wrong to compute the overall rate by adding the rates and dividing by 4.
$(35.33 + 21.38 + 16.86 + 14.17) \div 4 = 21.94$
Instead, we will have to calculate the total deaths in each age group from the fatality rate percent given above.
No. of deaths

$$= \frac{35.33}{100} \times 150 + \frac{21.38}{100} \times 304 + \frac{16.86}{100} \times 421$$
$$+ \frac{14.17}{100} \times 170 = 213.06$$

213.06 deaths have occurred out of 1,045 cases of a disease. Hence overall fatality rate percent

$$= \frac{213.17}{1045} \times 100 = 20.39\%$$

Calculation can be simplified by directly multiplying the frequency of cases with the death rate in each age group and then dividing the total deaths by total cases. This eliminates division as well as multiplication by 100.

$$\frac{150 \times 35.33 + 304 \times 21.38 + 421 \times 16.86 + 170 \times 14.17}{1045} = 20.39$$

Weighted mean is easy when groups are fewer, size of observations is small and they are not in fractions in a frequency distribution table. No working origin is necessary, one has to simply find the value of each group, i.e., weight, add the group weights and divide by the total number in the sample.

JAYPEE BROTHERS

- **Example No. 3:** Find mean days of confinement after delivery in the following series:

Days of confinement *x*	*No. of patients group (weights)* *f*	*Total days of each group (weights)* *fx*
6	5	30
7	4	28
8	4	32
9	5	27
10	2	20
	18	**137**

Apply the formula for the weighted mean ∑fX/n where X denotes the series of observations, i.e., days of confinement and f is the frequency of patients in each group and n is the total frequency, i.e., the number of patients confined.
Mean days of confinement, therefore

$$= \frac{\Sigma fX}{n} = 137 \div 18 = 7.61$$

Class interval in the days of confinement is taken as 1.

- **Example No. 4:** Calculate mean when class interval is again one, but group characteristics and frequencies are in fractions and large in size as in the weight of large number of boys in the table that follows:

Weight of children (kg) *X*	*Midpoint or average weight of each group* X_g	*No. of children in each group* *f*	*Weight contributed by each group* fX_g
60– <61	60.5	10	10 × 60.5
61– <62	61.5	20	20 × 61.5
62– <63	62.5	45	45 × 62.5
63– <64	63.5	50	50 × 63.5
64– <65	64.5	60	60 × 64.5
65– <66	65.5	40	40 × 65.5
66– <67	66.5	15	15 × 66.5
Total		**240**	**15310.0**

JAYPEE BROTHERS

Finding the average weight of each group from the original records may be laborious and they may not be available. The group wise frequency distribution table is prepared as in example 3. To find the average or mean of each group the mid value of the group is taken as its mean, such as 60.5 kg for weight group 60 – <61. Now the weight contributed by each group (Xg) is found by multiplying with the frequency (f). Apply the formula as in example 3.

$$= \frac{\Sigma f X_g}{n} = \frac{15310.0}{240} = 63.79$$

This method of calculation is still laborious and can be further simplified by assuming an arbitrary mean or working origin as was done in the ungrouped series when size of observations was large. This working origin will be the mid value of any group which is taken as 0.

The group values above and below this assumed mean (0) may be reduced to working units of 1, 2, 3, ..., etc., which are multiples of class or group interval in the frequency distribution table.

Now proceed stepwise as follows:

- Place 0 opposite any convenient group which may be taken as the working origin (w). It should be the mid value of that group.
- Find the weight contributed by the group by multiplying frequency (f) with weight in working units (x –), i.e., f x where x = $X_g - w$.
- Sum up the group weights and divide by the number of persons to calculate mean in working units x = $\Sigma fx/N$
- Multiply this mean ($\bar{x}$) by the class interval, 1, 5 or 10 as the case may be; $1 \times \bar{x}$ or $5 \times \bar{x}$ or $10 \times \bar{x}$.
- Add this to the working origin. This gives the mean ($\overline{X}$) in real units $\overline{X} = \bar{x} + w$.

Now mean weight may be found as below:

Weight (kg) x	*Mid-value Xg*	*Frequency or no. f*	*Working units x*	*Group weight fx*	*Sum of fx (fx)*
60–61	60.5	10	–2	10 × –2	
61–62	61.5	20	–1	20 × –1	
62–63	62.5(w)	45	0	45 × 0	

Contd...

JAYPEE BROTHERS

Contd...

Weight (kg) x	*Mid-value Xg*	*Frequency or no. f*	*Working units x*	*Group weight fx*	*Sum of fx (fx)*
		75			–40
63–64	63.5	50	+1	50 × 1	
64–65	64.5	60	+2	60 × 2	
65–66	65.5	40	+3	40 × 3	
66–67	66.5	15	+4	15 × 4	
		165			+350
Total		**240**			**+310**

Mean in working units, $\overline{X} = \frac{310}{240} = 1.29$

Mean in real units, $\overline{X} = w + \overline{x} = 62.50 + 1.29 = 63.79$.

Thus, mean in real units = central value (62.5) in real units of the group against which 0 is placed, + × class interval (1.29 × 1). If the class interval is 5 as in the next example, multiply the mean in working units by 5 and add it to the central value of the group or arbitrary working origin (w).

- Find the mean age at death in years in the following series:

Age of persons at death X	*Age in working units x*	*No. died f*	*Group wise year lived fx*	*Sum of fx Σ(fx)*
0–<5	– 6	100	–600	
5–<10	– 5	40	–200	
10<15	– 4	20	–80	
15–<20	– 3	25	–75	
20–<25	– 2	20	–40	
25–<30	– 1	25	–25	
30–<35 (w)	0	30	0	–1020
35–<40	+1	15	+15	
40–<45	+2	18	+36	
45–<50	+3	16	+48	
50–<55	+4	20	+80	

Contd...

JAYPEE BROTHERS

Contd...

Age of persons at death X	*Age in working units x*	*No. died f*	*Group wise year lived fx*	*Sum of fx Σ(fx)*
55–<60	+5	15	+75	
60–<65	+6	5	+30	
65–<70	+7	1	+7	+291
Total		350		–729

Mean age at death in working units. $\bar{x} = 5 \times \frac{-729}{350} = 5 \times -2.08$

= –10.40 (5 is the group interval with which mean in working units has to be multiplied)

Thus, real mean age, $\overline{X}$ = 32.5 (mid value opposite which 0 is placed) – 10.4 = 22.1 years.

- Find the mean birth weight of 470 infants born in a hospital, in one year, from the following table:

Weight of infants (kg) x	*Mid value*	*Weight in working units x*	*No. of infants f*	*Group weight fx*	*Sum of fx Σ(fx)*
2.0–2.4	2.2	– 2	17	–34	
2.5–2.9	2.7	– 1	97	–97	–131
3.0–3.4	w = 3.2	0	187	0	
3.5–3.9	3.7	+1	135	+135	
4.0–4.4	4.2	+2	28	+56	+209
4.5 +	4.7	+3	6	+18	
Total			470		+78

Mean weight $\overline{X} = w + \bar{x} \times$ group interval where x is the mean in working units.

$$\overline{X} = 3.2 + \frac{78}{470} \times 0.5 = 3.20 + 0.083 = 3.283$$

Essential Points to be Noted in Calculation of Mean

- Ordinary mean is the arithmetic mean of ungrouped series.
- To simplify calculations, put 0 against any convenient individual observation in ungrouped series, taking it as an arbitrary working origin or assumed mean.

JAYPEE BROTHERS

- In grouped series put 0 opposite any convenient group, preferably central group and find the mean in working units.
- Keep in view the mid value of the group opposite which 0 is placed, while finding the mean in real units.
- Do not forget to multiply the mean in working units with class interval when it is more than 1.
- Weighted mean is the average of grouped series in a frequency distribution table.

Calculation of Median

Ungrouped Series

Arrange the observations in the series in ascending or descending order of magnitude. The central observation of the arranged series gives the median.

Formula, $\frac{n+1}{2}$ gives the serial number of median, irrespective of the number (n) of observation in the series, odd or even.

In odd number of observations, e.g., 39, the serial number of median will be $\frac{39+1}{2} = 20$, i.e., the 20th observation is median.

In case of even number of observations, such as 40, apply the same formula

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5.$$

Median will be the mean of 20th and 21st observations, e.g., if in 40 patients, the period of stay of 20th patient was 10 days and that of 21st patient was 12 days placed in order, the median value of 40 patients as regards period of stay will be $\frac{10+12}{2} = 11$ days.

Example

- Duration of sickness absenteeism in a year, of 10 school children is given below. Calculate the median duration of absence.
  4, 8, 5, 5, 3, 9, 2, 14, 7, 3
  Arranged in ascending order: 2, 3, 3, 4, 7, 7, 7, 8, 9, 14
  Median is the value of $\frac{n+1}{2}$ observation = $\frac{10+1}{2} = 5.5$, i.e., mean of 5th and 6th observation, i.e. 7.

JAYPEE BROTHERS

Grouped Series

To find the median in grouped series, simply divide the total observation by 2. If total is 200, 100th observation is median. Even if the total is 201, the 100th observation may be taken as median.

Value of median observation, can also be obtained by graphic and arithmetical methods. Taking the example given in **Table 3.10** in the chapter it comes to 171.50 cm and 171.46 cm, respectively, as explained later under the section of percentiles.

Calculation of Mode

Mode is the most frequently occurring value. It can be determined in the following ways:

- Mode is located from frequency distribution table as the value of observation with maximum frequency.
- In a frequency polygon, mode lies at the point where the curve turns from increase to decrease, i.e., the highest peak.
- Mode can also be calculated as, Mode = Mean – 3 (Mean – Median)

MEASURES OF LOCATION: PERCENTILES

Percentiles

Averages discussed so far are measures of central value, therefore, they locate the center or midpoint of a distribution. It may also be of interest to locate other points in the range. Percentiles do that by dividing the entire population into 100 equal parts and then drawing an imaginary line that divides the data into two parts to obtain the percentage required such as 10% and 90%, or 25% and 75%, etc. In all, there are 99 percentiles. Centiles or percentiles are values in a series of observations arranged in ascending order of magnitude which divide the distribution into 100 equal parts. Thus, the median is 50th percentile. The 50th percentile will have 50% observations on either side. Accordingly, 10th percentile should have 10% observations to the left and 90% to the right. Values of a variable such as height, weight, age, etc., may be expressed in this way. If children at age 3½ years form 10th percentile, it means 10% of entire population is below 3½ years of age and 90% is above that age. Age 20 may be median or 50th percentile and divide the people into one half below 20 years and the other half above 20 years of age. Thus, percentiles are used to divide a distribution into convenient groups. Those in common use are described below:

JAYPEE BROTHERS

Deciles

Deciles, 9 in number, divide the distribution into 10 equal parts. First decile or 10th percentile will divide the distribution into 10% and 90% while 9th decile will divide into 90% and 10% and 5th decile will be same as median. All the 9 deciles are described below:

1. First decile or 10th percentile, will have 90% observations on its left and 10% on its right.
2. Second decile or 20th percentile, will have 20% observations on its left and 80% on its right.
3. Third decile or 30th percentile, will have 30% observations on its left and 70% on its right.
4. Fourth decile or 40th percentile, will have 40% observations on its left and 60% on its right.
5. Fifth decile or 50th percentile, will have 50% observations on its left and 50% on its right. This is also the median.
6. Sixth decile or 60th percentile, will have 60% observations on its left and 40% on its right.
7. Seventh decile or 70th percentile, will have 70% observations on its left and 30% on its right.
8. Eighth decile or 80th percentile, will have 80% observations on its left and 20% on its right.
9. Ninth decile or 90th percentile, will have 90% observations on its left and 10% on its right.

Quintiles

Quintiles, 4 in number, divide the distribution into 5 equal parts. In an example of height of a group of people first quintile will have 10% observations of heights falling on its left and 90% on its right.

- First quintile or 20th percentile, will have 20% observations on its left and 80% on its right.
- Second quintile or 40th percentile, will have 40% observations on its left and 60% on its right.
- Third quintile or 60th percentile, will have 60% observations on its left and 40% on its right.
- Fourth quintile or 80th percentile, will have 80% observations on its left and 20% on its right.

Quartiles

These are three different points located on the entire range of a variable Q_1, Q_2 and Q_3. In an example of height of a group of people

JAYPEE BROTHERS

Q_1 or lower quartile will have 25% observations of heights falling on its left and 75% on its right.

- First quartile or 25th percentile, will have 25% observations on its left and 75% on its right.
- Second quartile or 50th percentile, will have 50% observations on its left and 50% on its right. This is also the median.
- Third quartile or 75th percentile, will have 75% observations on its left and 25% on its right.

So, median of a variable can also be called as second quartile Q_2, 5th decile D5 or 50th percentile P_{50}.

Difference between Percentage and Percentile

Percentage means proportion out of the whole, while percentile means location or position or rank in a data series. For example, in an examination a student scored 80 percentage of marks and attained 90th percentile. This means out of the total marks in the examination the student has scored 80% marks, while out of the total students who appeared for the examination there are 10% students who scored more than him.

Graphic Method of Location of a Percentile in the Range of a Variable

Median can be located from the curve of cumulative frequency which is known as Ogive, which has been explained in Chapter 3. The curve is plotted with the value of observations on the X-axis and cumulative frequency on the Y-axis.

For example of graphic location of percentiles, refer to Cumulative Frequency **Table 3.10** and to Ogive shown in **Figure 3.16** in Chapter 3.

To find the median height or Q_2 in the above example, draw a perpendicular on Ogive **(Fig. 3.16)** from the midpoint of frequency N/2, i.e., 200/2 = 100 and note the point where it cuts the Ogive. Perpendicular line from this point on to the baseline or X-axis locates the median value. It comes to 171.50 cm.

Repeat the same process to find the height of the first or lower quartile Q_1, i.e., of 50th observation (200/4) or the third quartile Q_3, i.e. of 150th observation (200 × ¾).

Similarly, any percentile value can be found, e.g., to find 10th percentile determine the value of 20th observation (200/10).

The values for Q_1, median, Q_3 and 10th percentile are shown in the cumulative frequency diagram **(Fig. 3.16)**.

JAYPEE BROTHERS

From the same diagram, one can also find how any given value of a variable divides a distribution into two parts.

Thus, the percentage of people who are taller or shorter than one's own height can be found. Suppose it is 175 cm. Draw a perpendicular from this height on the baseline on to the Ogive **(Fig. 3.16)** and from Ogive to the vertical line. We find 50 persons out of 200 or 25% are taller than 175 cm.

Arithmetic Method of Finding Value of Any Percentile from Cumulative Frequency Table

The median, quartile and percentile values of the characteristic can be calculated from the cumulative frequency table. Find the variable group in which the particular observation lies and then raise the lower value of the variable of that group proportionately to the value of that particular observation, with the presumption that the rise in the variable group from the lower value to the higher value is uniform.

For illustration, median, Q_1, Q_3 and 10th percentile are calculated from the cumulative frequency **(Table 3.10)**.

- Median (Q_2), i.e., 200/2 or the 100th observation lies in the height group 170.0–171.9 cm. The cumulative frequency up to the height less than 170 cm is 81. The frequency rises by 19 from 81 to 100, i.e., median, or middle observation. For 26 observations, the attribute value as per the table, rises by class interval of 2 cm from 170 cm to 171.9 cm. Therefore, the proportionate rise in the attribute for 19 observations.

  $$= 2 \times \frac{(100 - 81)}{26} = \frac{2 \times 19}{26} = 1.46$$

  Thus, median or second quartile Q_2 value = 170 + 1.46 = 171.46 cm which is almost equal to the graphic value 171.50 cm. Median can thus be calculated from the formula,

  $$\text{Median} = L \frac{\frac{N}{2} - c.f.}{f} \times h$$

  Where,

  L = Lower limit of the median class

  $N = \sum f$ = Sum of the frequencies

  c.f. = Cumulative frequency of the class preceding the median class

  f = Frequency of the median class

  h = Class size = Class interval

JAYPEE BROTHERS

Quartiles can also be calculated in the same way as explained below.

- First quartile (Q_1), i.e., 200/4 = 50th observation, lies in the group 166–168 cm. Cumulative frequency up to height 166 cm is 42.

$$Q_1 = 166 + 2 \times \frac{(50 - 42)}{19} \text{ (19 is the group frequency)}$$

$$= 166 + \frac{16}{19} = 166.84 \text{ cm}$$

- Third quartile (Q_3), i.e., 200 × 3/4 = 150th observation, lies in group 174–176 cm. Cumulative frequency up to height 174 is 136.

$$Q_3 = 174 + 2 \times \frac{(150 - 136)}{30} \text{ (30 is the group frequency)}$$

$$= 174 + \frac{28}{30} = 174.93 \text{ cm}$$

- Tenth (10th) percentile, i.e., 200/10 or 20th observation, lies in the group 162 to 164 cm.

$$P_{10} = 162 + 2 \times \frac{(20 - 10)}{15}$$

$$= 162 + \frac{20}{15} = 163.33 \text{ cm}$$

To find how a particular observation or percentile will divide the entire frequency into two parts of percentages, one has to find in which group of the frequency distribution table this particular observation lies and then raise the frequency value from the lower limit proportionately, again making the presumption that rise is uniform from lower to higher value.

For illustration, let us find how the height 163.33 cm divides the entire frequency. As seen from the frequency distribution **(Table 3.10)**, this height lies in cumulative frequency of group 162 to 164 cm, i.e., somewhere between 10 and 25, proportional rise of frequency from 10 in this group would be calculated as follows:

Rise in frequency for 2 cm height = 15

Rise in height = 163.33 – 162.00 = 1.33 cm

Rise in frequency for the height 1.33 cm

$$= \frac{15}{2} \times 1.33 = 10$$

JAYPEE BROTHERS

Therefore, the number of observations up to the height 163.33 cm = 10 + 10 = 20, which forms the dividing point where 19 observations lie below 20 and 180 above that. Thus, we make use of the *cumulative frequency distribution table* and not the cumulative frequency polygon.

To summarize, we can locate the size of a given percentile by employing either of the two methods, graphic as well as arithmetic. Conversely, we can also find how a given size of an observation divides the frequency distribution into two parts by employing either of the methods.

Application and Use of Percentiles

- *Location of a percentile that divides the frequency distribution into two parts.*
  1. **Example:** The 25th percentile is located in the entire height range as 166.84 cm graphically and arithmetically.
  2. **Example:** Height 163.33 cm is found to be 10th percentile graphically and arithmetically. It divides the distribution of 200 heights into 10% and 90%. If a boy's height is less than 163.33 cm, he is short in size for his age. 90% boys are taller than he is.
- *Preparation of a standard percentile,* such as quartile Q_1 or median Q_2, etc., for particular age, gender, etc.
  Example: Standard median is made use of in the preparation of a weight card, recommended by the World Health Organization to assess the nutritional status of children, which has been adopted by India for monitoring growth and development of children and adolescents, separately for boys and girls, under the National Health Mission **(Fig. 5.1)**. The same is used in Integrated Child Development Services (ICDS) Scheme, India.

  Figure 5.1 shows growth chart for 0–3 year-old girl and boy child. The line demarcating the lower limit of green zone indicates 80% of the standard median and that below the yellow zone indicates 70% of the standard median, at different ages. Color coding is done to identify at a glance moderately malnourished children lying in the yellow zone and those lying in the red zone as severely malnourished. In the diagram, thus prepared, the vertical axis indicates the weight in kg and horizontal axis gives the age in months. Weight at any age is plotted to note between which lines it lies.
- *Comparison of one percentile value of a variable* of one sample with that of another sample, drawn from the same population

JAYPEE BROTHERS

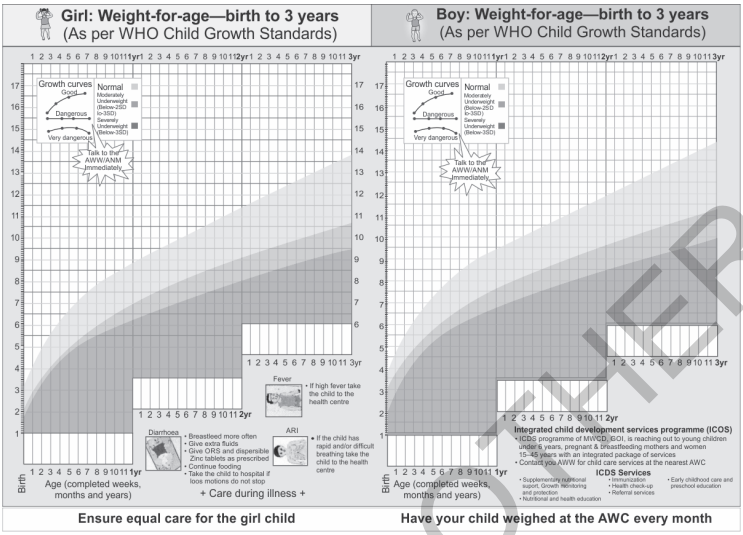


Fig. 5.1: Weight card used for grading weight of children in age group 0–3 years in the ICDS Scheme, India.

or from different population, median, quartile or any other percentiles can be compared.

Example: Cumulative frequency polygons or Ogives for 1-year-old Indian and Harvard children are drawn taking an identical sample in both the cases **(Fig. 5.2)**. It is seen that Harvard 50th percentile corresponds to Indian 60th percentile. Weight up to 9 kg covers 50% of Harvard children but it covers 60% of the Indian children, showing former children are heavier than the latter.

Thus, size of any percentile of one sample can be compared with other, and deviation found by the difference between the two will be clearly indicated.

- *To study growth in children*

Examples:

– A cumulative frequency table of heights of male children, who have completed 4th, 5th and 6th year of age, is given below:

Age in year	*Height in cm up to different percentiles*						
	5	*10*	*25*	*50*	*75*	*90*	*95*
4	94.0	96.0	99.8	103.7	106.9	110.5	112.9
5	97.3	99.2	102.6	106.2	110.3	114.2	115.8
6	102.8	104.9	108.1	110.4	115.2	118.4	119.4

JAYPEE BROTHERS

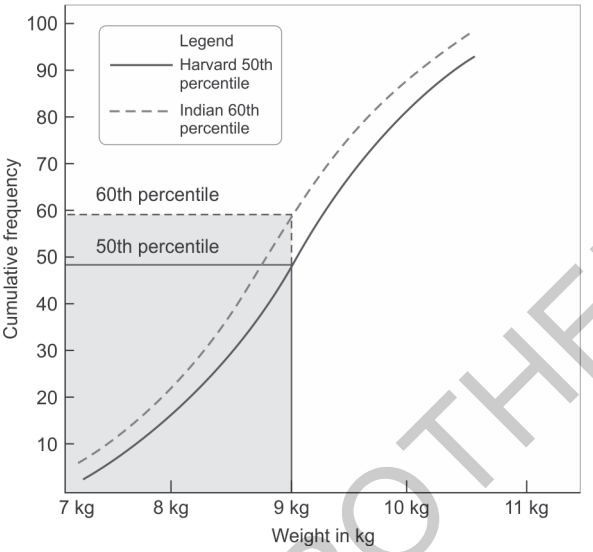


Fig. 5.2: Comparison of weight in Indian and Harvard children at 1 year of age by percentiles.

Percentiles at different ages indicate growth which is almost uniform when different percentiles are compared. It is about 3 cm from age 4 to 5 years and about 4 cm from age 5 to 6 years. The table also shows how different percentiles divide the distribution of children in two parts.

At age of 4 years the height 94 cm shows that 5% children were shorter and 95% taller than height 94 cm. Only 5% children are taller than 112.9 cm in this group.

– The 10th percentile height of 200 children at age 4 years was 96 cm. Half of the children (100) were put on vitamins A and D, and the other 100 children were kept as control and given placebo.

After one year, it was found that 10th percentile height was 97.3 cm in control group and 99 cm in intervention group. Thus, it was observed that vitamins A and D boosted the height in intervention group **(Fig. 5.3)**.

- *As a measure of dispersion*: Interquartile and semi-interquartile ranges are sometimes used as measures of dispersion, as explained in Chapter 6.

JAYPEE BROTHERS

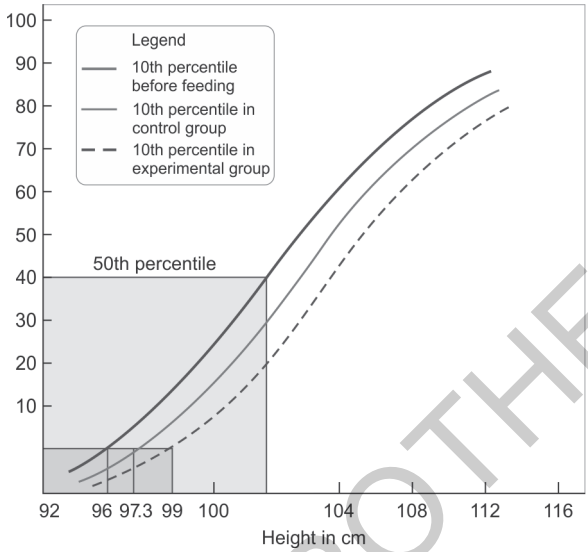


Fig. 5.3: Increase in height after feeding, indicated by 10th percentile in control and experiment groups.

QUESTION BANK

1. **Name and define the statistical measures of central tendency (centering constants) and variability.**
2. **Following stillbirth rates per 1,000 total births were repeated by 30 towns in 2024. Calculate mean, median and mode of the given data.**

   27, 28, 40, 32, 30, 36, 25, 29, 30, 29, 26, 30, 20, 35, 32
   36, 37, 29, 29, 42, 32, 27, 35, 36, 29, 33, 27, 41, 49, 34
3. **Determine the range, mean deviation, variance, standard deviation and coefficient of variation of the data in exercise 2.**
4. **(a) Sex ratio of males and females in India in 1981, was 1000:933, literacy rate in the same year was 40% in males and 20% in females. Calculate general literacy rate.**

   (b) If birth rate in the same year was 32/1,000 population for males and 28 for females, calculate combined birth rate.
5. **Calculate the mean number of living children per woman from the following table:**

JAYPEE BROTHERS

No. of living children (x)	*No.of women (f)*
0	42
1	49
2	57
3	40
4	31
5	22
Total	**241**

6. **Calculate mean and standard deviation of the following data on protein content of local sorghum varieties: 9, 8, 8, 9, 8, 9, 10.**
7. **From the frequency table prepared under Exercise 13 show that the mean calculated from a grouped data does not exactly agree with the mean calculated directly from the total number of observations.**
8. **Compare variability of systolic blood pressure in children of age group 5–10 years, with that of adults of age group 30–40 years. Their means and SDs were 100 and 8 in children and 120 and 12 in adults, respectively.**
9. **Define percentiles and discuss their applications. Compute first, second and third quartiles and semi-interquartile range for the following data:**

Height (cm)	*No. of students*
145.0–147.4	2
147.5–149.9	4
150.0–152.4	8
152.5–154.9	18
155.0–157.4	30
157.5–159.9	40
160.0–162.4	40
162.5–164.9	28
165.0–167.4	24
167.5–169.9	6

JAYPEE BROTHERS

Answers

2. Mean = 32.167, median = 31, mode = 29
3. Range = 29, mean deviation = 4.5, Var. = 34.69, SD = 5.89, CV = 18.31%
4. a. GLR = 30.346 b. CBR = 30.06
5. 2.145 per woman
6. $\overline{X}$ = 8.71, SD = 0.760
7. $\overline{X}$ from grouped data = 32.3, from n observations = 32.167
8. Adults10%, children 8%
9. q_1=156.5, q_2=159.88, q_3=163.2, SIQR= 3.35

JAYPEE BROTHERS

for detailed information on medical statistics books, visit our website www.ejaypee.com, for detailed information on medical statistics books, vis

Mahajan's Methods in Biostatistics for Medical Students and Research Workers

Salient Features

- *Mahajan's Methods in Biostatistics for Medical Students and Research Workers* is one of the bestselling biostatistics books and a great asset for medical students, research workers, and medical teachers of all disciplines.
- This book is written in simple language, with complex statistical procedures being explained lucidly.
- It is handy in size while including all principles of biostatistics in a concise manner.
- A special feature of the 11th edition is that several new chapters on statistical methods have been added, namely history and evolution of statistics, test statistics of Mann–Whitney U test, Wilcoxon test, Kruskal–Wallis test, and Kaplan–Meier method of survival analysis. All the chapters have been thoroughly revised and edited. Question bank has been added at the end of each chapter.

Yogesh K Singhal MBBS MD (PSM) MBA (HHSM) is an Associate Professor, Department of Community Medicine at American International Institute of Medical Sciences, Udaipur, Rajasthan, India, with over 15 years of experience in teaching medical, paramedical, and nursing field. He is a faculty of community medicine at DigiNerve—digital e-learning platform for medical students. He has more than 15 publications in national and international scientific journals. He is keenly interested in research and has been supervised many students in several funded research projects. His areas of interest are epidemiology, research methodology, health economics, healthcare management, hospital quality programs, and project management.

Buy from ejaypee

JAYPEE BROTHERS

Shelving Recommendation
MEDICAL STATISTICS

ISBN 978-93-6616-697-1