Demystifying Medical Data

Demystifying Medical Data:

Unlocking the Power of Statistics in Healthcare

By

Dr Geeta Arora and Dr Sarabjit Singh

Cambridge Scholars Publishing

Demystifying Medical Data: Unlocking the Power
of Statistics in Healthcare

By Dr Geeta Arora and Dr Sarabjit Singh

This book first published 2025

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-5335-0
ISBN (Ebook): 978-1-0364-5336-7

TABLE OF CONTENTS

PREFACE

The fields of material science, business analysis, sports, management, and even biological research all depend primarily on the results of statistical analyses. While there are many books on statistical analysis for mathematicians, business majors, and management students, there is not a single book for biomedical researchers and basic medical scientists who did not major in mathematics in college. This comprehensive statistics book aims to help readers comprehend statistical concepts and their practical applications in diverse real-life scenarios, with a particular emphasis on the medical sciences. The writing style of this book is simple and straightforward. This book is an excellent resource for students pursuing any course of medical sciences, be it dentistry, pharmacy, or graduate and undergraduate clinical courses.

I hope that teachers and researchers in the field of medicine and basic medical sciences will enjoy this book.

This work comprises brief topic details, multiple-choice questions (MCQs), and some unsolved problems in the form of fill-in-the-blank and true/false questions. This book covers all the important topics related to the field of statistics. After a brief discussion of each topic, students will feel confident when they attempt the variety of questions provided.

We have verified the answers to the provided problems. We have made efforts to check the content for typos and grammatical errors, but it's possible that a few have slipped through. I would appreciate it if readers could point out these errors for improvement. I always appreciate suggestions and comments from readers to enhance the content of the book. Readers are welcome to share their experience of their learning with the provided material at my email, geetadma@gmail.com.

Dr. Geeta Arora

CHAPTER 1

DESCRIPTIVE STATISTICS

When you're done with this chapter, you will be able to:

- Know about statistics and its role in data handling.
- Explain the difference between nominal, ordinal, and metric data.
- Figure out the type of value that a given variable can take.
- Explain why ordinal data is not a number.
- Explain a frequency distribution of the given data.
- Construct a frequency table from the raw data.
- Construct the tables for data with relative frequency, cumulative frequency, and relative cumulative frequency.

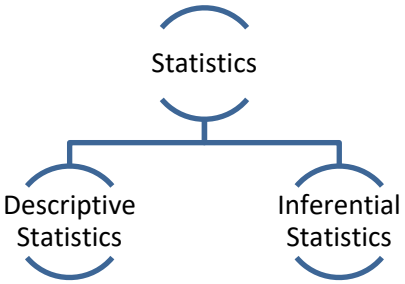
One of the key discussions in this chapter revolves around the definition of statistics. Let's begin with the first topic of discussion: What do you understand by the term "statistics"?

Statistics is a field of study that focuses on the analysis of data. It is a branch of mathematics that manages data and provides useful information. It is a tool to process and analyze numbers. The various steps involved in the analysis of a fact include the collection of data, organization of data, and finally implementation of the statistical tools to interpret the data.

Decision-makers use statistics to reduce the uncertainty involved in various aspects of business decision-making. It helps from the initial stage of presenting and describing information to drawing valid conclusions about large groups of individuals or items. Furthermore, we utilize statistics to create dependable projections for data related to health, economy, education, or any other data-driven activity. Furthermore, statistics plays a crucial role in enhancing processes by making informed decisions. Therefore, people use statistics for a variety of decision-making purposes.

For instance, statistical analysis of data serves the following purposes:

- One can analyze the effectiveness of the marketing strategies used to promote the new surgical procedure.
- The aim could be to assess the effectiveness of a recently introduced medication in the market for managing blood pressure.
- Statistics can be used to determine the effectiveness of a recently introduced gadget in tracking the risk of heart attacks in elderly individuals.



1.1 Branches of Statistics

Statistics can be categorized into two types based on the type of analysis to be performed on the data:

Descriptive Statistics: It involves organizing, summarizing, and displaying data. It is used to describe the important characteristics of the data with the help of tables, charts, averages, percentages, etc. The measures of central tendency, the creation of graphs, and dispersion are the essential parts of these statistics.

1. Inferential Statistics: This branch of statistics involves drawing conclusions about the population or making inferences about a large group based on a partial analysis.
2. We commonly refer to a group of entities as a population, and the small portion under study as a sample. Therefore, we can define statistics as the process of interpreting the characteristics of a population through the investigation of a small sample.

1.2 Raw Material of Statistics

Data is the **raw material of statistics.** Data can exist in the form of unarranged facts and figures that can be sorted out, arranged and classified according to the common attributes. Data which is collected from a survey for the analysis of the cause from some source to be interpreted is called as *raw data*.

For Example:

- The birthweights of a sample of 100 babies.
- The age and blood groups of the infants in a hospital in the last six months.
- The number of doctors available in the hospital with higher education.
- The number of nursing staff in a medical college/university.

The following is an example of data in raw form that has not yet been arranged and is collected from the infants of age 1-20 days who are diagnosed with increased bilirubin.

Yes	**No**	**Yes**	**Yes**	**No**	**Yes**	**No**
Yes	**No**	**Yes**	**No**	**No**	**Yes**	**No**
No	**Yes**	**No**	**No**	**Yes**	**No**	**Yes**
Yes	**No**	**No**	**Yes**	**No**	**Yes**	**Yes**

A single observation in the data is called a **data point,** and the collection of data is referred to as **a data set.**

We can further understand data from the perspective of how it can be collected:

- The **elements** are the entities on which data are collected.
- A **variable** represents a characteristic that is of interest to the elements.
- An **observation** is the collection of measurements for a specific element.
- The total **number of data values** in a data set is the number of elements multiplied by the number of variables.

Example 1: Examine the data gathered from various educational streams, including medical, law, sciences, and humanities, regarding the number of students, their placements, and the startups they initiated.

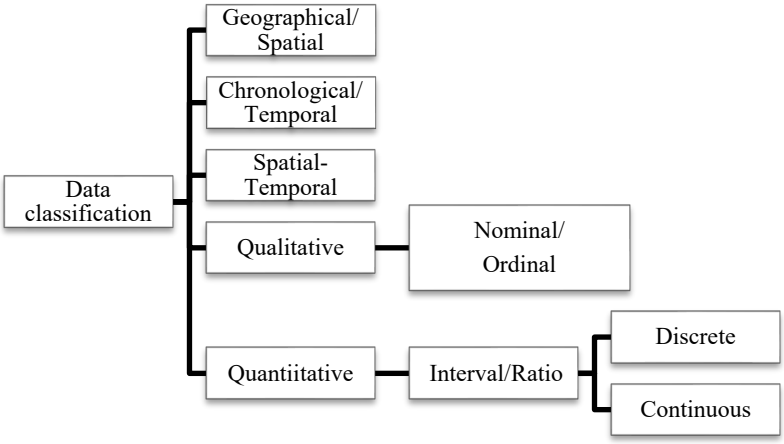
Educational streams	Number of students	Placement held	Startup initiated
Medical	102	45	10
Law	45	30	5
Sciences	54	27	4
Humanities	40	32	5

The four elements in this case are the educational streams given as medical, law, sciences, and humanities. The three variables that determine the collection of observations are the number of students, placement held, and startup initiated. The total number of data values in a data set is twelve.

1.3 Classification of Data

A. We can classify data based on the amount of information it contains. As discussed below, we can categorize the data's scale of measurement and represent it in the form of a hierarchy.

1. **Nominal Data**- We refer to data that we can categorize and write in text form as nominal data.



For example:

- **Gender:** Male and Female
- **Income status:** high class and middle class
- **Brands of laptops:** Dell, HP, Lenovo, and Apple

You can also use a numeric code for the variables.
For instance, you can use 1 for the male and 2 for the female.

Also, the data under the nominal category is independent of the units of measurement, and the ordering of the categories is arbitrary.

For example, assume data on "live births" in February 2022 is available. Out of a total of 100 births, 44 were boys and 56 were girls. The order of the data regarding the number of boys and girls can vary based on the number of values each category contains.

Example: Female: 44; Male: 56

2. **Ordinal Data- When** data can be categorized and arranged in some order, it is called ordinal data. Thus, the data possess the same properties as nominal data, and their order or rank also holds significance. But the ordinal data are not real numbers; it is not appropriate to apply any of the rules of basic arithmetic to this sort of data. One cannot add, subtract, multiply, or divide the ordinal values.

For example:

- The quality of surgical instruments ranges from excellent to best.
- There are three stages of cancer: first, second, and third.
- The distance between the OPD and the emergency department is either near or far away.
- Reviews of patients range from 1 for excellent, 2 for good, 3 for neutral, and 4 for bad.

3. **Interval and Ratio**

We can further classify data as **interval** and **ratio**.

The interval data have the properties of ordinal data, and the interval between the observations can be expressed in terms of a fixed unit of measure.

You can obtain the difference between the values in interval data, but there is no real zero. It means that zero has a value.

Consider example of **temperature**: This level of measurement encompasses temperature data, specifically the range between 35°F and 85°F for outdoor temperatures. We sort those values and define the difference between them as 50°F. However, there is no natural starting point. Although the 0°F figure may seem like a starting point, it is arbitrary and does not represent the absolute absence of heat.

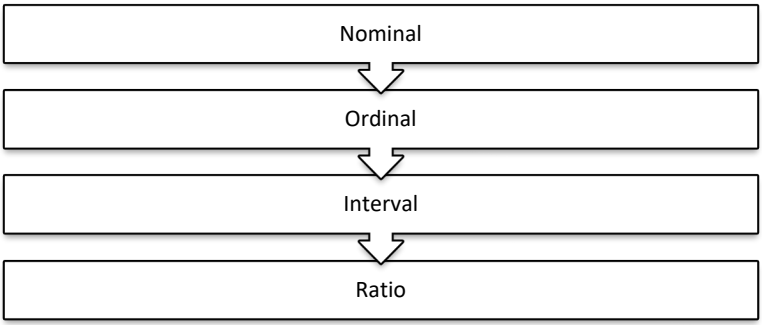
Another example can be considered of **years**: If we consider the years 2022 and 1980 in chronological order, the difference of 42 years is significant. However, because time did not start in the year 0, the year 0 is arbitrary rather than a natural zero beginning point indicating "no time."

Ratio-type data pertains to the enumeration of data in which zero signifies nothing.

Example: The number of students in the class can be 10, 20, or 0. In this case, zero indicates that there are no students present.

The ratio data has all the properties of interval data, and a ratio of two values is meaningful. Variables such as distance, height, weight, and time use the ratio scale. Another example is that if Ram weighs 80 kg and Kailash weighs 40 kg, then Ram's weight is twice that of Kailash.

Following is the hierarchy of the above defined data types:



And can be summarized as follows:

Data type	Nominal	Ordinal	Interval	Ratio
	Categories only. Data cannot be arranged in order.	Data can be sorted, but gaps cannot be detected or have no significance.	Differences have significance, but ratios are meaningless since there is no inherent zero reference point.	Ratios have meaning, and zero is a suitable beginning point.
Example:	Gender: Male/Female	Stages of cancer	Temperature	Weight, Height, lengths

B. Generally, we classify data based on the types of values a variable can take. This category encompasses five distinct classification types:

1. **Geographical/ Spatial classification-** The term "geographic/spatial classification" refers to the categorization of data based on its geographic location.

   For example:

   - Population of the different towns in a city.
   - The residential status of the doctors of a medical Institute.
   - The location of the WHO help centers located all over India.

2. **Chronological/ Temporal classifications-**

   Chronological data refers to events that occurred within a specific time frame. You can measure time in any unit, including years, months, weeks, hours, minutes, and seconds.

   For example:

   - A patient's blood pressure fluctuates during specific hours of the day.
   - The population of infected hens in a poultry farm in the last three years.

- The number of reported patients who have suffered from acute heart attacks in the last two weeks.

3. **Spatial-temporal classifications:** Data is said to be of spatial-temporal type when the data is recorded for the time interval with respect to the geographical location as well.

   For example:

   - Population of the different towns in a city during the last ten years.
   - The residential status of the research fellows in an institute in the last one year.
   - The location of the COVID help centers created in the last two years.

4. **Qualitative classifications:** Qualitative data is defined based on the characteristics or attributes of an entity.

   For example:

   - The gender of the person appointed as member secretary in an election.
   - The history of the father of a patient regarding diabetics.
   - Record of the immunity of the people living in a society.

5. **Quantitative classifications:** We refer to any data that provides a quantity of the discussed entity. It can be numbers, values, amounts, or any unit that can be measured and compared.

   For example:

   - Numbers of the health care centers in a city.
   - The number of dialysis done on a day-to-day basis in a hospital.
   - Number of patients suffering from kidney problems in a hospital.

   Quantitative data can further be classified according to the values that it can take. It can be of discrete or continuous type.

i. **Discrete Data**-Data is said to be of discrete type when the data values are integers and cannot be taken in fractions. Thus, data that may be represented by a single integer value is called **discrete data or metric data**.

For example:

- Number of deaths in a city in the last month.
- Number of patients admitted to a hospital due to heart attacks.
- Number of hospitalizations in the emergency ward during the festive season.

Thus, metric or discrete variables are those that can be reduced to a count or a set of values (the "numbers of things") that can be quantified in some way. The numerical information they supply is always a whole number (an integer).

ii. **Continuous Data**: Data is said to be of continuous type when it can take any value between a range.

For example:

- Reading of the blood pressure (mmHg) of the patients.
- The record of the blood cholesterol (µg/ml) of a person in the last month.
- The body mass index (kg/m2) measured during the last year at regular intervals.
- Weight (kg) of a young man can lie anywhere between 30-70 kg.

Thus, metric continuous data are those that have been measured and hence have units of measurement. The information is based on actual measurements.

Note: Another classification of the data can be done based on the category of the data. There are two major types of variables: **categorical variables and metric variables**; each of them can be further divided into two subtypes as shown below and are already discussed above.

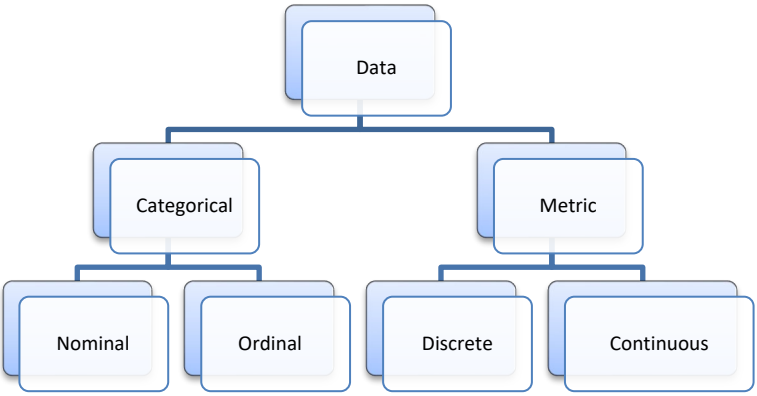


Fig. 1-1 Classification of data

1.4 Descriptive Statistics

To answer the issues of interest related to the data and to gain insight into the situation, there are several strategies for organizing and presenting the data. The general term for all these techniques is **"descriptive statistics."** They are a collection of methods that can be used on raw data to draw out the most salient aspects and qualities. Size-based sorting, tabularization, graphical representation, and numeric summarization are all possible ways to accomplish this.

After obtaining the data from a reliable source, the next step is to find out the type of data and to manage it. Managing data refers to the sorting and presentation of data in a tabulated form. The common term used for the data management is **frequency distribution,** which involves the following steps:

- Arrange data in ascending or descending order.
- Arrange data in tabular form.

The data can further be classified in a grouped frequency form known as **grouped frequency distribution**.

1.4.1 Frequency Distribution

Frequency refers to how often something occurs.

For example:

- The number of times a meal is served in a day.
- The number of patients suffering from diabetes is based on their age.
- The number of times a day a CT scan machine is operated.

A frequency distribution table can be created by counting the number of times a value is repeated. Rewriting the data in the class interval form gives the grouped frequency table.

Frequency tables for ordinal data

When the information available is ordinal, it can be sorted into categories. Consider a study in which researchers polled 500 patients to gauge their opinion of the quality of treatment they received from nurses. It's quite evident that "level of satisfaction" is an ordinal scale. While "satisfaction" itself cannot be quantified, the categories can be arranged in meaningful ways. As can be seen from the table given below, more than half of the patients (335 patients) who received nursing care have reported that they are satisfied with the care. Dissatisfaction was indicated by a far smaller percentage of people.

Satisfaction with nursing care	Number of patients (n=500)
Very satisfied	180
Satisfied	155
Neutral	100
Dissatisfied	25
Very dissatisfied	40

1.4.2 Definitions related to the grouped frequency distribution

The class interval: It refers to the class in which the values lie between. It has two values that refer to the lower-class limit and upper-class limit.

Lower-class limit: These are the smallest numbers that can belong to different classes.

Upper-class limit: These are the largest numbers that can belong to different classes.

Mid-points: The mid-points can be found by adding the lower-class limit to the upper-class limit and dividing the sum by two.

Class width: It is the difference between two consecutive lower-class limits. To determine the width of a class interval, divide the range (highest value–lowest value) of the data by the number of classes.

Example 2: A survey was conducted by a researcher on 150 diabetic patients to find the number of people of different ages that suffer from diabetes. The results can be written in the frequency distribution table as follows:

Age	Number of people
20-30	11
30-40	22
40-50	33
50-60	84

Here, the number of people aged between 20-30 years is 11. For this class interval, the lower limit is 20, the upper limit is 30, and the class width is 10, with the mid-point of class interval as 25.

In general, a frequency distribution should have at least 5 but not more than 15 classes.

1.5 Methods of Data Classification

There are two ways in which the observations of a data set can be classified based on the class intervals: the exclusive method and the inclusive method.

Exclusive Method:

In the exclusive method, the data is presented in a way such that the upper limit of a class interval is the lower limit of the succeeding class interval.

Example 3: Consider the data of ages obtained for 20 patients who had been recommended for heart transplantation as follows:

60	42	53	44	50	39	28	52	35	63
58	56	64	47	35	77	46	40	24	57

To arrange the data into class intervals, the data can first be arranged in either ascending or descending order, and then the data can be classified in class intervals using an exclusive approach. Following is the data in ascending form:

24	28	35	35	39	40	42	44	46	47
50	52	53	56	57	58	60	63	64	77

Data arranged by exclusive method with class width 10 can be written as:

Class Interval	Class Interval	Frequencies
20-30	20 but less than 30	2
30-40	30 but less than 40	3
40-50	40 but less than 50	5
50-60	50 but less than 60	6
60-70	60 but less than 70	3
70-80	70 but less than 80	1

Inclusive Method:

In the inclusive method, the data is classified in a way such that both the lower and upper limits of a class interval are included in the interval itself.

Example 4: The number of patients falling under the class interval can be presented as:

Class Interval	Class Interval	Frequencies
20-30	Ranging from 21 to 30	2
30-40	Ranging from 31 to 40	4
40-50	Ranging from 41 to 50	5
50-60	Ranging from 51 to 60	6
60-70	Ranging from 61 to 70	2
70-80	Ranging from 71 to 80	1

NOTE: An exclusive method is used to work with continuous data, and an inclusive method is used to work with discrete data.

1.6 Types of Frequency Distributions

A) Cumulative Frequency Distribution

A cumulative frequency distribution is of two types:

(i) More than type- In this type of cumulative frequency distribution, the frequency of each class interval is added successively from bottom to top and represents the cumulative number of observations greater than or equal to the lower limit of class intervals.

(ii) Less than type- In this type of cumulative frequency distribution, the frequency of each class interval is added successively from top to bottom to give the number of observations less than or equal to the upper limit of class intervals.

Here is an example of research reporting on a study of a drug's adverse effect. In this case, 30 people are used to determine an average increase in weight in kilograms (kg).

Weight (In Kgs)	***Number of patients (Frequency)***	***Lower limit***	***Cumulative frequency (More than)***	***Upper limit***	***Cumulative freq (Less than)***
1-2	5	More than 1	30	Less than 2	5
2-3	11	More than 2	25	Less than 3	16
3-4	14	More than 3	14	Less than 4	30

B) Relative Frequency Distribution

Relative frequency distribution is used to show the percentage of observations that fall within each class of a distribution. For calculating the relative frequency distribution, divide each class frequency by the total number of observations in the entire distribution. Following is the calculation of the relative frequency of the data discussed above.

Weight (In Kgs)	*Number of patients (Frequency)*	*Relative Frequency*	*Percentage Frequency*
1-2	5	$\frac{5}{30} = 0.1667$	16.67
2-3	11	$\frac{11}{30} = 0.3667$	36.67
3-4	14	$\frac{14}{30} = 0.4667$	46.67
Total	30	1	100

Chapter based Quiz

A) MCQs

1. Collecting the data for the choice of the flavour of ice cream falls in the category of____data.

   a) Nominal
   b) Ordinal
   c) Interval
   d) Ratio

2. Rating of the movie by the reviewers as Excellent, Best, Good, Boar can be taken in the _____data type.

   a) Interval
   a) Ratio
   b) Discrete
   c) Ordinal

3. The number of clothes needed to be purchased for a ceremony is of _____data type.

   a) Discrete
   a) Continuous
   b) Irrational
   c) None of the above

4. The weight of the parcel to be delivered at the post office is of _____data type.

a) Discrete
a) Continuous
b) Irrational
c) None of the above

5. Data of the quality of the food from the customers feedback (Good, Better, Best) is an example of____data.

a) Nominal
a) Ordinal
b) Interval
c) Ratio

6. In a cricket match, a player got 5 times six runs, 4 times 2 runs and 10 times 1 run. The frequency of the player to hit 2 runs is

a) 19
a) 5
b) 2
c) 4

7. Arranging the data as per the number of times it occurs (frequency) is known as

a) Frequency distribution
b) Distribution
c) Arranging data
d) Grouping data

8. The number of employees present in a meeting is a data of which type?

a) Ordinal
a) Interval
b) Ratio
c) Nominal

9. The list of the winners of a debate competition is a_____data.

a) Nominal
a) Ordinal
b) Interval
c) Ratio

10. The amount of rainfall held in the month of July is _____type of data.

    a) Discrete
    a) Continuous
    b) Qualitative
    c) Geographical

11. Population of a town in the last four months is a ____type of data.

    a) Spatial temporal
    a) Qualitative
    b) Chronological
    c) Continuous

12. Number of employees belonging to different parts of a country can be categorized as

    a) Qualitative
    a) Discrete
    b) Continuous
    c) Spatial data

13. For the given data, what is the total number of data values?

Items	A	B	C	D	E
Numbers	10	13	15	10	12

    a) 10
    a) 40
    b) 50
    c) 60

14. Priscar is organizing, summarizing, and displaying data of the recent games held in India. He is applying the concept of

    a) Descriptive statistics
    a) Inferential Statistics
    b) Time Series Analysis
    c) Indexing

15. John wants to find the average height of a graduate student in India. For this he has collected a sample of 500 graduate students from different parts of the country. He is using _____.

a) Descriptive statistics
a) Inferential Statistics
b) Time Series Analysis
c) Indexing

16. Consider this data to answer the related questions:

Number of persons in a house	1-3	3-5	5-8	8-11	11-14
Number of Families	10	14	8	9	7

I. What is the class width of the given data?

a) 1
a) 2
b) 3
c) 4

II. What is the total number of families taken for the collection of data?

a) 10
d) 20
e) 48
f) 54

III. How many families have more than 5 persons?

a) 10
g) 48
h) 24
i) 20

IV. What is the lower limit of the class interval with a frequency as 8?

a) 10
j) 14
k) 8
l) 5