Charan Singh Rayat

Statistical Methods in Medical Research

Springer

Statistical Methods in Medical Research

Charan Singh Rayat

Statistical Methods in Medical Research

Springer

Charan Singh Rayat
Department of Histopathology
Postgraduate Institute of Medical Education & Research
Chandigarh, India

ISBN 978-981-13-0826-0 ISBN 978-981-13-0827-7 (eBook)
https://doi.org/10.1007/978-981-13-0827-7

Library of Congress Control Number: 2018954730

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

It gives me great pleasure in writing the foreword to *Statistical Methods in Medical Research* by Dr. CS Rayat, who has more than 32 years' experience in applications of statistical methods in Biomedical Research. Dr. Rayat has more than 13 years postdoctoral research and teaching experience with great instinct for quantitative diagnostic pathology and has been on the editorial boards of reputed international medical journals.

The main aim of his book is to create interest of medical researchers and postgraduate students in the wonderful applications of statistical methods for analyzing the research findings as well as routine determinations of biochemical investigations for finding significance of outcome of research and "statistical quality control" of routine "medical laboratories" for accreditation and competence.

The book would be of great use for medical professionals, researchers, and students of biomedical sciences and other disciplines too. I wish Dr. Rayat great success in his endeavor and feel really happy for the release of the first edition of the book.

Pranab Dey MD. FRC Path.

Department of Cytology,
PGIMER
Chandigarh, India
March 2018

Preface

Null hypothesis (H_o) is the cardinal principle of equality in groups of objects or subjects under study. It is the foundation of Statistical Science. The extent of rejection of null hypothesis through statistical analysis validates our research or inference about group of subjects, objects, or standard operating procedures (SOPs) under study.

Statistical analysis plays a vital role in material science, business analysis, sports, management, and biomedical research. A variety of books are available on statistical analysis for students of mathematics, commerce, and management, but not a single book is available for biomedical researchers and basic medical scientists who have not studied mathematics at undergraduate level. I express my gratitude to Dr. Pranab Dey, Professor, Department of Cytology, PGIMER, Chandigarh, for encouraging me to write this book for biomedical researchers and basic medical scientists.

This book has been written with a focus on the requirements of students of various specialties of medical science, basic medical scientists, and health workers. Problems with solutions have been illustrated under various chapters for comprehensive understanding of the statistical methods and their applications. I am of the definite view that this book would meet the requirements of students of medicine and basic medical sciences. I adore the pioneers in the subject whose theories and methods have been cited in this book and dedicate the book to the galaxy of pioneers.

I would like to convey my thanks to my son Er. Harpreet Singh Rayat, a software engineer, and daughter Ms. Amandeep Kaur, M.Com., ICWA, for their support and critical comments on the manuscript. Microsoft Office 365 ProPlus, Version 1708, licensed to my son, has been used as source of Microsoft Excel for generating probability tables and data handling. I am very thankful to Ms. Jagjeet Kaur Saini and Dr. Naren Aggarwal of Springer India Pvt. Ltd, for taking interest in my manuscript and putting in sincere efforts for publishing this book.

I hope that this book would be liked by teachers and researchers in the field of Medicine and Basic Medical Sciences, and would lessen their dependency on others for statistical analysis. Any suggestions for improving this book would be highly acknowledged and appreciated.

Chandigarh
India
March 2018

C.S. Rayat

Contents

About the Author

Charan Singh Rayat completed his postgraduate studies and Ph.D. in Pathology at the Postgraduate Institute of Medical Education & Research (PGIMER), Chandigarh, India. He is a Lecturer in histopathology at PGIMER and has hands on experience in a variety of diagnostic and research studies in the field of immunology and pathology, along with more than 17 years of experience in ultrastructural pathology and morphometry. Dr. Rayat has published papers in numerous medical journals and is a peer reviewer for international journals. The Department of Immunopathology, PGIMER, Chandigarh, conferred awards on him in recognition of his diagnostic work in the field of immunopathology and histopathology and his contributions to patient care and clinical research. On his blogs, he also disseminates his knowledge for the benefit of medical students and patients.

1 Introduction to Statistics

1.1 Definitions of Statistics

1. Systematic evaluation and methodological numeric analysis of parametric data collected on a subject or population are called statistics.
2. The conclusive study of statistical methods and principles employed to understand the outcome of a research project is also termed as statistics.
3. Statistics means quantitative data analysis by statistical methods to elucidate the validity and accuracy of a scientific procedure or study.
4. Statistics could be defined as quantitative data affected to a certain extent by multiplicity of causes and evaluated through statistical methods.

1.2 Origin and Development of Statistics

Erstwhile kings used to collect information regarding population and wealth of their people. Collected data was analyzed to plan development of the state and to finance war. Statistics in those days was known as "science of kings." Later the data of diverse nature were obtained for general uses of the government. Students of game of chances also developed certain methods of statistical analysis. Biology and insurance as well as other natural sciences are bright fields for application of statistical methods.

1.3 Concept of Probability

The concept of probability is used in day-to-day life which stands for the probability of occurring or non-occurring of events. The notion of probability is used in social sciences, statistics, economics, industry, business, and engineering. The probability

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_1

is an element of uncertainty about the happening of an event. The following statements help in understanding the concept of probability:

1. There may be rain tomorrow.
2. Final examination may take place in the month of July.
3. The chance of winning a lottery is less than 1%.

The examples cited above are uncertain about the associated events. So, these statements are just conjectures. The concept of probability was very first used by Italian mathematician Eardan in his book entitled *The Book on Games of Chance* in 1963. The foundation of the "mathematical theory" of probability was laid by French mathematicians Blaise Pascal and Pierre de Fermat. Afterwards the Swiss mathematician James Bernoulli (1654–1705) contributed to the theory of probability; however, his concepts came to light after his death. Other pioneers associated with the probability are de Moivre (1667–1754), Thomas Bayes (1702–1761), Markov (1856–1922), and Kolmogorov.

The probability has a great role in our day-to-day life. Our personal life, social life, academic life, and even business life are deeply associated with probability. Managerial decisions in production and business are analyzed in the light of theories of probability to calculate risks and uncertainties.

1.4 Definition of Probability

The word probability denotes the likelihood of occurring of an event. It is an intelligent guess regarding happening of an event. The probability of occurring of an event could range between "0" and "1." If the event does not happen, the probability would be "0" (zero). On the other hand, when the event happens, the probability would become "1" (one). As per concept the event can happen in M ways or cannot happen in N ways. So, the total possibilities of probabilities are $M + N$, and the probability of happening can be calculated as

$$P = \frac{M}{M+N} = \frac{\text{Number of favorable cases}}{\text{Total number of likely events}}$$

Definition of Events The observed results of identical experiments are called events. An event may be elementary or composite. Events are denoted by capital letters A, B, C, etc. Each event is classified into success (p) or failure (q). Now, if success is zero (0), failure will have the value as "one" (1) or the other way. As per the rule of probability

$$\begin{aligned} p+q&=1\\ p&=1-q\\ q&=1-p \end{aligned}$$

1.5 Types of Events

1. *Mutually exclusive events*: All events, happening independent of the other, are called "mutually exclusive events." As in the case of tossing a coin, the head and tail cannot occur simultaneously. Only head or tail can occur. So, the occurrence of one event would exclude the occurrence of the other.
2. *Equally likely events*: Events are said to be "equally likely" if all the cases have an equal chance to occur. As we draw a playing card from a pack of cards, every card out of 52 cards will have the equal chance of being getting drawn. Similarly when a dice is thrown, 1, 2, 3, 4, 5, or 6 equally has a likely chance to occur. So, such events are called "equally likely events."
3. *Exhaustive events*: When all the cases of random experiment are included in the study, the events are called "exhaustive events." Possible outcomes in a dice are 1, 2, 3, 4, 5, or 6, and any of these may appear on the top.
4. *Simple and compound events*: "Simple events" are based on the natural law and without any calculations. An event may occur or may not occur. In "compound events" joint possibility of two or more events is considered. For example, if we have a bag containing five white, six red, and seven green balls, the possibility of drawing two balls at a time may be both white, both red, and both green. Possibility would also be of white and red, white and green, or red and green.
5. *Independent and dependent events*: The event is said to be independent if happening of this is not under the happening of any other event. For example: Drawing of a "King" from a pack of cards would always have a probability as $\frac{4}{52}$. A single card seen and put in the packet again would have probability as $\frac{1}{52}$.

   If the happening of the second event depends on the happening of the first event, we call it "dependent event." For example, if we try to draw a "King" from the pack of cards, its probability will be $\frac{4}{52}$. If we remove a card from the pack at random and try to draw a "King" again, its probability will be $\frac{4}{51}$. In the second case, if the drawn card was a "King" and we keep that aside and try to draw the second "King," then the probability of second King will be $\frac{3}{51}$.
6. *Complementary events*: Events are said to be "complementary events" if one event complements the other event for the failure or success. For example, in a throwing of "dice," if condition is applied that winning player will win if the score is odd number, it would mean that the player will win only if 1 or 3 or 5 comes at the top during the toss of dice.

1.6 Different Theories of Probability

The concept of probability is based on certain laws which may be modified or amended with passage of time. There are four theories which will help us understand the concept and applications of probability. These theories are:

1. Classical or priori probability
2. Relative theory of probability
3. Subjective approach
4. Axiomatic theory of probability

1.6.1 Classical or Priori Probability

The "classical or priori probability" is the earliest approach, and the same was developed by Laplace who also coined the definition of probability as "the ratio of the favorable case to the number of equally likely cases." If random experiment (A) results in N exclusive and equally likely events, out of which m are the favorable outcomes, then the probability (P) of occurring of favorable event is given by the following formula:

$$P(\mathrm{A}) = \frac{m}{N}$$

Example 1 Suppose you throw a dice. The probability of occurring of number "1" would be denoted as

$$P(1) = \frac{\text{No. of favorable cases}}{\text{Total No. of equally likely cases}} = \frac{1}{6}$$

So, classical probability has the following four properties:

(a) $0 \leq m \leq N$ is divided by N; we will have $0 < \frac{m}{N} < 1$, i.e., $P(\mathrm{A}) \leq 1$.

(b) If "A" is an impossible event, then $m = 0$. So, $P(\mathrm{A}) = \frac{0}{N} = 0$.

(c) If "A" is certain event, then $m = N$. So, $P(\mathrm{A}) = \frac{N}{N} = 1$.

(d) If occurrence of "A" implies to the occurrence of "B" and occurrence of "B" may or may not imply to the occurrence of "A," then $P(\mathrm{A}) \leq P(\mathrm{B})$.

1.6.2 Relative Theory of Probability

The “probability” of an event is personal or empirical according to “relative theory of probability.” It is proportionate to time under identical circumstances. The probability of happening of an event is determined based on past experiences. For example, if a teacher is asked to give questions for the coming examinations, then his probability would be based on past examinations. The “relative probability” of an event would be expressed as

$$P = \frac{\text{Relative Frequency}}{\text{No. of cases or items}}$$

Example 2 Suppose 10,000 products are manufactured on a machine and as per past experience 300 products were found to be defective items. In this case the probability of defective (D) items would be

$$P(\mathrm{D}) = \frac{300}{10000} = 0.03$$

1.6.3 Subjective Approach

Personalistic or “subjective concept of probability” measures that an individual has truth of a particular proposition. Frank Ramsey developed the subjective approach and published in his book entitled *The Foundations of Mathematics and Other Logical Essays* in 1926. This theory is used by everybody in day-to-day life for making decisions in business especially in those where one man dominates the show. Subjective concept of probability is also used in war where every personal approach varies due to individual instincts. This is the most flexible approach in comparison to other approaches. This requires careful analysis during its applications.

1.6.4 Axiomatic Theory of Probability

The word axiom represents the common saying in the society. The approach includes both “classical” and “empirical” probability to develop this theory. Russian mathematician Kolmogorov AN introduced this theory in 1993. According to this theory, the probability of an event ranges between “0” and “1,” and if the events are mutually exclusive, the probability of occurrence of either event “A” or event “B” would be denoted by

$$P(\mathrm{A\ or\ B}) = P(\mathrm{A}) + P(\mathrm{B})$$

1.7 Uses of Probability

Theories of probability have extensive applications in various fields in day-to-day life as enumerated below:

1. Law of statistical regularities.
2. Law of inertia of large numbers. Preparation of various estimates is based on probability of outcome.
3. Various parametric and non-parametric tests like *Z*-test, *t*-test, *F*-test, etc. are based on "theory of probability."
4. Probability is applied to "theories of games" for managerial decisions. The expected values are calculated through probability.
5. Sales managers and production managers apply various approaches of probability to take economic decisions in various situations of risks and market uncertainty.
6. The theory of "subjective probability" is always preferred if it is not possible to calculate each expectation.
7. Theoretical aspects of probability are applied to contest the practical significance of a project, experiment, game, or business.

1.8 Theorems of Probability

Basically, there are two theorems of probability: (1) addition theorem and (2) multiplication theorem.

Addition Theorem

It states that when two events are exclusive, the probability of occurring of event "A" or event "B" would be the sum of individual probability of each event as depicted below:

$$P(\mathrm{A\ \ or\ \ B}) = P(\mathrm{A}) + P(\mathrm{B})$$

Example 3 Suppose a card is drawn from a pack of cards. The probability of occurring of a King (K) or Queen (Q) would be

$$P(\mathrm{K\ \ or\ \ Q}) = P(\mathrm{K}) + P(\mathrm{Q}) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

It denotes that the King and Queen are mutually exclusive events, individuals, words, or outcomes. Mutually exclusive events have been shown in Fig. 1.1.

If the events are not mutually exclusive and the two events would overlap each other as depicted in the Fig. 1.2, the "addition theorem" as postulated for the King or Queen would not apply. The theorem is modified as below:

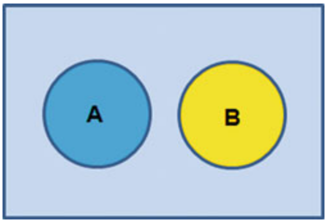


Fig. 1.1 Mutually exclusive events

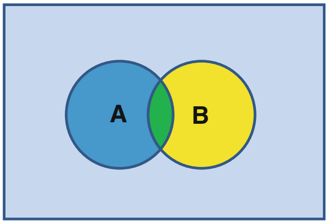


Fig. 1.2 Overlapping events (green area is the overlap)

$$P(\mathrm{A\ \ or\ \ B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{AB})$$

Example 4 There is a chronic ailment to a patient, and there are 75% (¾) chances that doctor "A" can treat it and 80% (⅘) that doctor "B" can treat it. What will be the probability of being cured if the patient gets treatment from doctor "A" and "B" simultaneously?

Doctor "A" treats the patient: $P(\mathrm{A}) = ¾$
Doctor "B" treats the patient: $P(\mathrm{B}) = ⅘$

Probability of A and B collectively:

$$\begin{aligned} P(\mathrm{A\ \ or\ \ B}) &= P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{AB}) \\ &= \frac{3}{4} + \frac{4}{5} - \frac{3}{4} \times \frac{4}{5} \\ &= \frac{3}{4} + \frac{4}{5} - \frac{3}{5} \\ &= \frac{15 + 16 - 12}{20} = \frac{19}{20} = 95\% \end{aligned}$$

So, there will be 95% chances that the patient gets treated.

If three events overlap each other as depicted in Fig. 1.3, then their probability would depend on the following theorem:

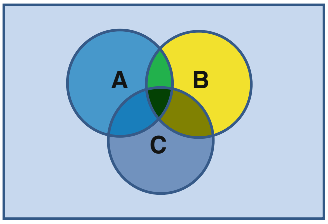


Fig. 1.3 Overlapping events (dark green area is the overlap of three events)

$$P(\mathrm{A\ \ or\ \ B\ \ or\ \ C}) = P(\mathrm{A}) + P(\mathrm{B}) + P(\mathrm{C}) - P(\mathrm{AB}) - P(\mathrm{AC}) - P(\mathrm{BC}) \\ - P(\mathrm{ABC})$$

Multiplication Theorem

The "multiplication theorem" states that if two events are independent, the occurrence of event "A" and event "B" would be the product of individual probabilities of both the events.

$$P(\mathrm{A\&B}) = P(\mathrm{A}) \times P(\mathrm{B})$$

If three independent events are there, then the probability of occurrence of events "A," "B," and "C" would be the product of probabilities of these three events.

$$P(\mathrm{A\&B\&C}) = P(\mathrm{A}) \times P(\mathrm{B}) \times P(\mathrm{C})$$

Example 5 Suppose we ask two persons to accomplish a task. The probability that person "A" can accomplish it is ¾, and probability that B can accomplish it is ⅔. What will be the probability if both the "A" and "B" are assigned the task?

$$\begin{aligned} P(\mathrm{A\&B}) &= P(\mathrm{A}) \times P(\mathrm{B}) \\ &= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2} \quad \text{(By Positive Approach)} \end{aligned}$$

Now if we apply "negative approach" in the same case:

"A" cannot accomplish the task = 1 − ¾=¼
"B" cannot accomplish the task = 1 − ⅔=⅓
A and B cannot accomplish the task = ¼×⅓=1⁄12
Probability that "A and B" can accomplish the task = 1 − 1⁄12=11⁄12

This shows that probability of accomplishing the task by "positive approach" is ½, whereas under "negative approach," it is 11⁄12. So, a modern statistician would suggest "negative approach" for solving a problem as it has improved probability.

Example 6 Let us find out the outcome of the "positive" and "negative" approach in the case of more than two options. Suppose we ask three persons to accomplish a task. The probability that "A" can accomplish it is ¾, that B can accomplish it is ⅔, and that "C" can accomplish it is ½. What will be the collective probability in this case?

Positive Approach

$$\begin{aligned} P(\mathrm{A\&B\&C}) &= P(\mathrm{A}) \times P(\mathrm{B}) \times P(\mathrm{C}) \\ &= \frac{3}{4} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{4} \end{aligned}$$

Negative Approach

"A" cannot accomplish the task = 1 − ¾=¼
"B" cannot accomplish the task = 1 − ⅔=⅓
"C" cannot accomplish the task = 1 − ½=½

Probability that "A and B and C" cannot accomplish the task = ¼×⅓×½=1/24.
Probability that "A and B and C" can accomplish the task = 1 − 1/24=23/24.
This shows that "negative approach" is better than "positive approach."

Conditional Probability
The "multiplication theorem" does not hold well when events are dependent. Two events "A" and "B" are said to be dependent when "B" can occur only when "A" occurs.

Suppose the boss says that he would attend your wedding only when there is no rain. His statement is conditional subject to the condition of "no rain." The "probability" (concern) to such event is termed as "conditional probability." When two events "A" and "B" are dependent, the conditional probability of "B" with given condition of "A" will be

$$\begin{aligned} P(\mathrm{B/A}) &= \frac{P(\mathrm{AB})}{P(\mathrm{A})} \\ P(\mathrm{B/A}) &= P(\mathrm{B}) \times P(\mathrm{A/B}) \\ P(\mathrm{A/B}) &= P(\mathrm{A}) \times P(\mathrm{B/A}) \end{aligned}$$

For three events:

$$P(\mathrm{ABC}) = P(\mathrm{A}) \times P(\mathrm{B/A}) \times P(\mathrm{C/AB})$$

Bayes' Theorem
If we do not know the concept of probability, then any task can be accomplished by daily inspection as per new inputs. Bayes' theorem is based on this concept of "revisiting probability" when new information is available. Reverend Thomas

Bayes, a British mathematician, developed the idea of "revisiting probability," and the same was published in 1763. This theorem is applied to ascertain the probability of event "B" when an associated event "A" had occurred. This is just like revising the probabilities based on additional information when the other event has already happened. Accordingly, the "posterior probability" of event "A" for a particular result of event "B" could be computed as

$$\begin{aligned} P(\mathrm{A/B}) &= \frac{P(\mathrm{A})P(\mathrm{B}\ )}{P\ (\mathrm{A})P(\mathrm{B}) + P(\mathrm{A})P(\mathrm{B}) + P(\mathrm{A})P(\mathrm{B})} \\ &= \frac{\text{Individual Joint Probability of X, Y or Z}}{\text{Sum of Joint Probabilities of X, Y and Z}} \end{aligned}$$

This formula for "Bayes' theorem" is based on conditional probability. This could be simplified and expressed in the form of a table as illustrated below as solution for Example 7.

Example 7 There are three salesmen in a store. Salesman "X" issues 40% bills daily, and out of these 1% are faulty; salesman "Y" issues 35% bills daily, and out these 2% are faulty; and salesman "Z" issues 25% bills daily, and out of these 3% are faulty. If a faulty bill is chosen at random, what would be the probability that the same was issued by salesman "X"?

Solution

Salesmen:	X, Y, and Z
Event A:	Issuing of bills
Event B:	Faulty bills
Condition:	Bill can be faulty (event B) only after it is issued (event A)

Posteriori probabilities have been computed in the Table 1.1.

After computation we found that probability of faulty bill issued by salesman X would be $\frac{0.004}{0.0185} = \frac{40}{185}$ or roughly $\frac{1}{5}$.

Table 1.1 Posteriori probabilities

1	2	3	4	5
Salesmen	Priori probability P(A)	Conditional probability event B given event A P(B/A)	Joint probability (2) × (3)	Posteriori probability $\frac{(4)}{P(\mathrm{B})}$
X	0.40	0.01	0.004	$\frac{0.004}{0.0185} = 0.216$
Y	0.35	0.02	0.007	$\frac{0.007}{0.0185} = 0.378$
Z	0.25	0.03	0.0075	$\frac{0.0075}{0.0185} = 0.405$
Total	**1.00**	**–**	**P(B) 0.0185**	**1.00**

Note Similar situation can be there in a hospital as illustrated below with the help of Example 8.

Example 8 Suppose renal transplant surgeries are done by three surgeons daily in a hospital. Surgeon "X" does 40% surgeries in a month, and out of these 1% of patients die within 3 months; surgeon "Y" does 35% surgeries in a month, and out these 2% of patients die within 3 months; and surgeon "Z" does 25% surgeries in a month, and out of these 3% patients die within 3 months. If at random a patient dies within 3 months after renal transplant surgery at this hospital, what would be the probability of surgeons X, Y, and Z individually that the patient may have been operated by him?

Solution

Surgeons:	X, Y, and Z
Event A:	Renal transplant surgery
Event B:	Death of patient within 3 months
Condition:	Death can occur (event B) only after transplant surgery (event A)

Posteriori probabilities have been computed in the Table 1.2.

After computation we found that probability of postoperative death could be linked to surgeons X, Y, and Z, respectively, as listed below:

1. Surgeon X $= \frac{0.004}{0.0185} = \frac{40}{185} = 21.6\%$
2. Surgeon Y $= \frac{0.007}{0.0185} = \frac{70}{185} = 37.8\%$
3. Surgeon Z $= \frac{0.0075}{0.0185} = \frac{75}{185} = 40.5\%$

Mathematical Expectations

We are always inclined to monetary gain or loss in daily life, may it be availing medical facilities or marketing some products. Mathematical expectations could be used to work out expected value of an even in different possibilities. Suppose "A" is a random variable which could occupy one value in A_1, $A_{2,}$ A_3, … A_n with

Table 1.2 Posteriori probabilities

1	2	3	4	5
Surgeon	Priori probability *P*(A)	Conditional probability event B given event A *P*(B/A)	Joint probability (2) × (3)	Posteriori I probability $\frac{(4)}{P(\mathrm{B})}$
X	0.40	0.01	0.004	$\frac{0.004}{0.0185} = 0.216$
Y	0.35	0.02	0.007	$\frac{0.007}{0.0185} = 0.378$
Z	0.25	0.03	0.0075	$\frac{0.0075}{0.0185} = 0.405$
Total	**1.00**	–	***P*(B) 0.0185**	**1.00**

corresponding probabilities of $P_1, P_2, P_3, \ldots P_n$. At this the mathematical expectation of "A" would be denoted by

$$E(\mathrm{A}) = \mathrm{P}_1\mathrm{A}_1 + \mathrm{P}_2\mathrm{A}_2 + \mathrm{P}_3\mathrm{A}_3 \ldots \mathrm{P}_n\mathrm{A}_n$$

So, the expected value would be the sum of the products $P_1A_1 + P_2A_2 + P_3A_3 \ldots P_nA_n$.

Following points are kept in mind in mathematical expectations:

1. If various possibilities of the event are positive, then we calculate the sum of every expectation as $E(A) = P_1A_1 + P_2A_2 + P_3A_3 \ldots P_nA_n$.
2. If expected possibility carries positive as well as some negative expectations, then net difference of gain and loss is considered as given below:

$$E(\mathrm{A}) = \mathrm{P}_1\mathrm{A}_1 - \mathrm{P}_2\mathrm{A}_2 + \mathrm{P}_3\mathrm{A}_3 \text{ etc.}$$

3. Mathematical expectations are used to evaluate different possibilities under risks involved and select the comparatively better options. Suppose we want to set up a corporate hospital or diagnostic center. We would focus on the beneficial option between the two.

Let us solve an example to learn more about "mathematical expectation":

Example 9 Suppose a private hospital earns Rs. 80,000-00 daily on an average during rainy season and Rs. 95,000-00 daily on an average during rest of the year. What will be the expected earning daily if chances of rainy season are 20% in a year?

Solution

$$\begin{aligned} \mathrm{P}_1 &= 20\% = 0.20 \\ \mathrm{A}_1 &= 80,000\text{-}00 \\ \mathrm{P}_2 &= 80\% = 0.80 \\ \mathrm{A}_2 &= 95,000\text{-}00 \end{aligned}$$

$$\begin{aligned} E(\mathrm{A}) &= \mathrm{P}_1\mathrm{A}_1 + \mathrm{P}_2\mathrm{A}_2 \\ E(\mathrm{A}) &= 0.20 \times 80,000 + 0.80 \times 95,000 \\ &= 16,000 + 76,000 = 92,000\text{-}00 \\ &= \mathbf{Rs.92,000\text{-}00\ Ans}. \end{aligned}$$

2 Collecting Statistical Data

2.1 Methods of Collection of Data

1. *Enumeration method*: Enumerator visits the subjects, asks necessary questions, and records the reply in tabulated format for easy understanding.
2. *Questionnaire method*: The subjects are asked to answer the questions mentioned in the questionnaire and return the same to the investigator.
3. *Registration method*: The information is recorded in the register, or online submission forms are invited and used for analysis.

2.1.1 Procedure for Collection of Data

1. Planning the study
2. Preparation of questionnaire and the schedule to be followed
3. Criterion for including or excluding a case
4. Organizing the data
5. Preparation of finished tables or charts
6. Statistical analysis of findings

2.1.2 Planning the Study

While planning the study, the following points should be kept in mind:

1. All the previous studies conducted on the same topic.
2. Design the schedule in such a manner that the factors considered in the present enquiry may be comparable with the previous studies.
3. Only relevant factors should be considered in setting up an enquiry.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_2

4. A decision is to be taken: whether the complete enumeration is feasible or a sample study is to be conducted.
5. It should be decided whether the investigators are to be used or questionnaire are to be mailed. In case the investigators are engaged, place of enquiry should also be decided.

2.1.3 Devising the Questions and Making the Schedule

1. All factors related to the study should be decided.
2. Formulate these questions so that these could be answered readily and accurately. Each question should be simple and clear to understand. Ambiguous questions should be avoided to avoid wastage of time. Self-realized answers should be objective and placed in tabulated form. Questions and answers should be arranged in a planned manner with reference to the plan of study.
3. A scheduled form is drafted and printed as questionnaire.
4. Schedule should be supplemented by a sheet or booklet of information.

2.1.4 Population and Samples

1. *Population*: The word population in statistics not only refers to human beings but also to inanimate objects. It is an aggregate about which an investigator is trying to study either by complete enumeration or by drawing sample.
2. *The sampled population*: It is population of which our data are random sample. It is an aggregate process by which we obtain our sample by giving chance to every member of the aggregate to be included in the sample.
3. *Sample*: It is a fraction of population. It contains only some units of the population and not all. To study the whole population is not only too expensive but also time-consuming. Secondly it is not necessary to study all units of population to draw a meaningful conclusion. There are various ways in which a sample may be selected to represent the whole population. Such a sample is called a representative sample.
4. *Random sample*: If a sample is selected in such a way that each time an item is selected, each item of the population has an equal chance of being selected in the sample. Such sample is referred to as unrestricted or *simple random sample*.
5. *Systematic sample*: When a sample is obtained periodically, say every tenth item on a list or in a file is selected as sample, such sample is called *systematic sample*.
6. *Stratified sample*: When a population is heterogeneous and when the heterogeneity has a bearing on the characteristic being studied, the population may be divided into strata, and independent random samples of units are drawn from each stratum. *Note*: Stratified sampling cannot be used unless some information regarding the population and its strata is available. Stratification of the population should be based on the factor closely related to the topic under study. The strata may differ from each other, but there should be homogeneity within each stratum.

7. *Multistage sample*: In this category, the sampling is done in stages. For example, to draw a sample from a country divided into number of states (1st-stage units), a "random sample" of states will be drawn. Thereafter, within the selected states, number of districts (2nd-stage units) should be drawn by the same "simple random sampling method." This would complete "two-stage sampling." Further, districts selected may be divided into smaller units, and these units may be considered for drawing a "third-stage sample." This way a multistage sample can be completed.
8. *Other samples*: The above referred samples are known as "probability samples." There are some other sampling schemes whose reliability is in doubt due to human bias. However, these sampling schemes are deliberated here:
   (a) *Purposive sample*: In this case an investigator tries to make the sample to agree with the population in certain characteristics.
   (b) *Quota sample*: In this case the investigators are given the liberty to include anyone in the sample who fulfills the conditions laid by the investigator.
   (c) *The random point sample*: This method consists of locating many points at random at a map and enumerates a predetermined number of sampling units nearest to each point.
9. *Deciding on sampling plan*: The efficiency of the scheme is taken into consideration while deciding sampling plan. The efficiency of sampling plan refers to the reliability in relation to the cost of unit.

2.1.5 Using Schedules to Obtain the Information

1. When agents or enumerators take the schedules to the persons who are to furnish the information, the enumerators may explain the purpose of the investigation and solicit cooperation. Each question can be clearly explained as it is being asked.
2. Enumerators are required to study the schedule and printed instructions and then to take an examination.
3. Enumerators should be persons of unquestioned integrity and should also be polite and tactful.
4. Enumerator should plan his/her interviews to consume as little time as possible and should make every effort to get the desired information if it is feasible to do so.
5. The work of the enumerator may be facilitated if a letter of explanation regarding proposed visit by enumerator is sent in advance.
6. Sometimes enumerators conduct interviews and fill in the schedules afterward. This is done on the theory that people feel more free to talk if the remarks are not written at the time of interaction.
7. Even though an enumerator makes his request for the information as tactfully as possible, he/she may sometimes meet with a refusal. Another visitor with a different approach may have better luck. Sometimes it is considered to be a

good plan to have one especially qualified worker to follow up more difficult cases.

8. Sending schedules by mail rather than using enumerators is at the outset and less expensive method of collecting data. There is also an added advantage that the person supplying the information can fill out the forms at his convenience, instead of being disturbed by the enumerator perhaps at a busy or inconvenient time.

2.1.6 Editing the Schedules

1. *Tabulation*: After the filled-out schedules are received, these are recorded in tabulated form in a comprehensive way in a register or on computer.
2. *Coding*: Tabulation is frequently facilitated by coding. Coding should be done by editor using distinctive colors.
3. *Deciphering*: The handwriting of an enumerator or an informant may occasionally be difficult to read. Deciphering such copy is editors' task. He not only saves time of the tabulator but also insures accurate results. If the entries are literally unreadable, the schedules may be referred to the enumerator or the informant.
4. *Checking*: The editor may look over the schedules for inconsistencies.
5. *Verification for completeness*: The editor should also scrutinize the schedules to see if any entries are missing or incomplete. If the missing information is important, the schedule must be sent back to the enumerator or to the informant; otherwise the editor writes "N.R" (not reported).

2.1.7 Organizing the Data

1. *Hand sorting*: When the data is not large and schedules are in the form of cardboard or thick paper cards, these can be sorted out by hand sorting.
2. *Score or tally sheet*: Large data should be computed out through scoring procedure using tally marks.
3. *Mechanical tabulation*: When the data is too large, punching, verifying, sorting, and tabulation machines may be used.

2.1.8 Presentation and Analysis

1. After organizing the information collected on schedules by various means, the finished statistical tables and charts may be drawn up.
2. The tabulated data is analyzed statistically to draw conclusions for presentation and record.

3 Tabulated Presentation of Data

3.1 Examples Depicting Tabulated Presentation of Data

Example 1 Suppose you are provided with data on “live births” in the month of March 2016. There were 210 births, and out of these, 112 were males and 98 were females. This data has been classified in tabulated form depicted in Table 3.1.

Example 2 You are provided with a data of 95 deaths occurred in the month of May 2016 in a big hospital in its 9 departments/services, where 30 patients died within 48 h of stay and 65 patients died after 48 h of stay at the hospital. The data has been presented in tabulated form in Table 3.2.

Example 3 In a study conducted on a group of patients with solitary pulmonary nodules, 40 patients tested positive for “tuberculin sensitivity,” and 30 tested negative for “tuberculin sensitivity.” Patients were further subjected to “histoplasmin sensitivity,” and it was observed that out of 40 patients tested positive for “tuberculin sensitivity,” 22 were also positive for “histoplasmin sensitivity,” and out of 30 patients tested negative for “tuberculin sensitivity,” 25 tested positive for “histoplasmin sensitivity.” Arrange your data in tabulated form.

Solution
The data cited in Example 3 have been exhibited in the Table 3.3.

Table 3.1 Data of live births in the month of March

Sex of newborn	Number of live births
Male	112
Female	98
Total live births	**210**

© Springer Nature Singapore Pte Ltd. 2018
C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_3

Table 3.2 Distribution of deaths with reference to specialty department and time of stay in a hospital

Service department	Stay in hospital		Deaths
	<48 h	>48 h	
General medicine	11	22	33
General surgery	5	9	14
Pediatric medicine	6	11	17
Pediatric surgery	3	3	6
Newborn	3	2	5
CTU medicine	0	4	4
Surgery "others"	0	14	14
Obstetrics and gynecology	1	0	1
ENT	1	0	1
Total deaths	**30**	**65**	**95**

Table 3.3 Tuberculin and histoplasmin sensitivity in a group of patients with solitary pulmonary nodules

Tuberculin sensitivity	Histoplasmin sensitivity		Total
	Positive (+)	Negative (−)	
Positive (+)	22	18	40
Negative (−)	25	5	30
Total	47	23	**70**

Example 4 In a study conducted on 1000 normal adults for serum cholesterol levels, it was observed that 10 had cholesterol level between 120 and 139 mg/dl, 21 had 140–159 mg/dl, 30 had 160–179 mg/dl, 90 had 180–199 mg/dl, 120 had 200–219 mg/dl, 192 had 220–239 mg/dl, 180 had 240–259 mg/dl, 120 had 260–279 mg/dl, 90 had 280–299 mg/dl, 40 had 300–319 mg/dl, 30 had 320–339 mg/dl, 10 had 340–359 mg/dl, 61 had 360–379 mg/dl, 4 had 380–399 mg/dl, none had 400–419 mg/dl, 1 had 420–439 mg/dl, none had 440–459 mg/dl, and 1 had 460–479 mg/dl. Arrange your data in tabulated form, and calculate percentage of volunteers in each group.

Solution
The data mentioned in Example 4 have been depicted in Table 3.4.

Example 5 You are provided with a survey data on height in adult males born in Punjab, Haryana, and Himachal Pradesh. Represent the frequency distribution of height in adult males in tabulated form.

Solution
The frequency distribution of statures for adult males has been depicted in Table 3.5.

Table 3.4 Distribution of serum cholesterol levels in 1000 normal men aged between 30 and 60 years

Cholesterol levels mg/dl	No. of men	Percentage
120–139	10	1.0
140–159	21	2.1
160–179	30	3.0
180–199	90	9.0
200–219	120	12.0
220–239	192	19.2
240–259	180	18.0
260–279	120	12.0
280–299	90	9.0
300–319	40	4.0
320–339	30	3.0
340–359	10	1.0
360–379	61	6.1
380–399	4	0.4
400–419	0	0.0
420–439	1	0.1
440–459	0	0.0
460–479	1	0.1
Total	**1000**	**100**

Table 3.5 Distribution of height in adult males with reference to place of birth

	Number of men within the said limits of height		
Height without shoes (inches)	Punjab	Haryana	Himachal
57–58	1	0	1
58–59	3	1	0
59–60	12	0	1
60–61	39	2	0
61–62	70	2	9
62–63	128	9	30
63–64	320	19	48
64–65	524	47	83
65–66	740	109	108
66–67	881	139	145
67–68	918	210	128
68–69	886	210	72
69–70	753	218	52
70–71	473	115	33
71–72	254	102	21
72–73	117	69	6
73–74	48	26	2
74–75	16	15	1
75–76	9	3	1
76–77	1	4	0
77–78	1	1	0
Total	6194	1301	741

Diagrammatic Presentation of Data

4

4.1 Usefulness

1. The impression created by a diagram or a picture is likely to last longer in the mind than the effect created by a set of figures.
2. One has to tax his brain in understanding figures and drawing conclusions, but in case of diagrams, conclusions automatically follow.

4.2 Limitations

1. Diagrams only give an approximate view of data.
2. Diagrams are likely to be misused easily.

4.3 Characteristics of Diagrams and Rules for Drawing These

1. Diagrams are meant only to give a pictorial representation to the quantitative data with a view to make them comprehensive.
2. Diagrams do not approve or disapprove a particular fact.
3. Diagrams are not suitable for further analysis of data which could only be possible from tables with numeric values.
4. Data to be represented with diagrams should be homogenous and comparable.
5. Diagrams are not substitute for numeric figures.
6. Selection of a scale is a must for execution of a diagram.
7. If multiple diagrams are to be drawn, the same scale should be used for all.
8. Vertical scale is shown on the left-hand side and the horizontal scale at the bottom of the diagram.
9. Diagram should not be clumsy.
10. Diagram should be attractive.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_4

11. Various points could be emphasized in a diagram with different colors, line patterns, dots, and crossings.
12. Extreme care should be exercised in selection of a particular type of diagram capable of representing given set of figures of data.

4.4 Different Types of Diagrams

1. *Dimensional diagrams*:
   (a) One-dimensional diagrams: These are vertical or horizontal lines or bars. The lengths of the lines or bars are in proportion to the different figures these represent.
   (b) Two-dimensional diagrams: These are in the shape of rectangles, squares, or circles. The areas of squares, rectangles, or circles are in proportion to the size of items represented by these.
   (c) Three-dimensional diagrams: These are in the shape of cubes, blocks, or cylinders. Volumes of these are in proportion to the values being represented. Systolic blood pressure in nine patients has been depicted as scattergram in Fig. 4.1; and distribution of serum cholesterol levels in 1000 normal men aged 30–60 years has been depicted in Fig. 4.2 by bar graph.

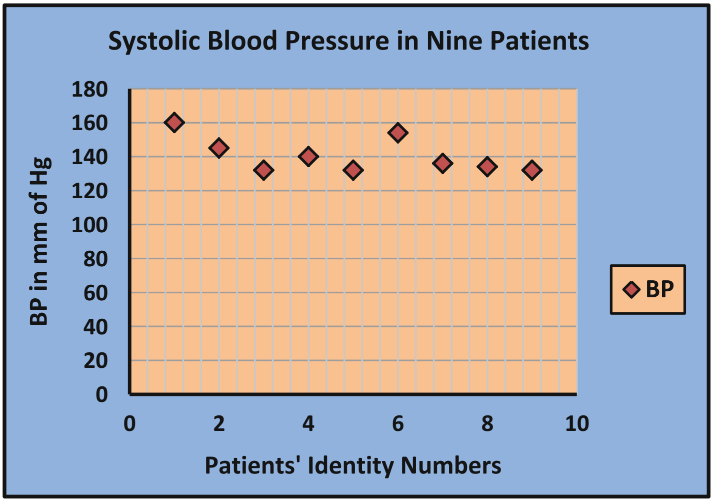


Fig. 4.1 Scattergram representing systolic blood pressure in nine patients

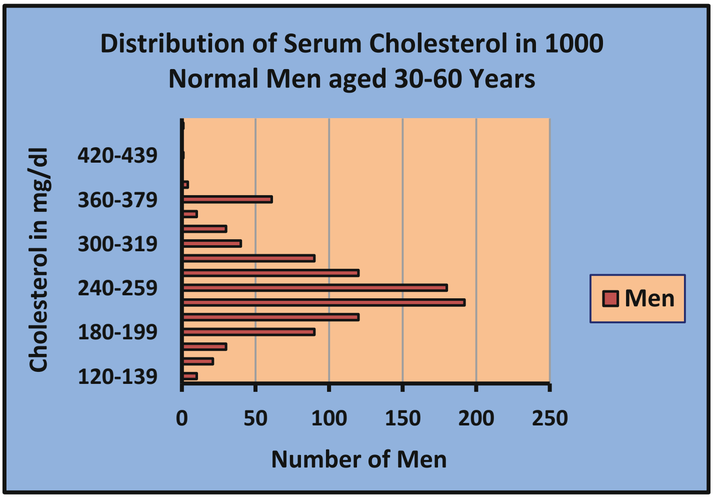


Fig. 4.2 Distribution of serum cholesterol levels in 1000 normal men

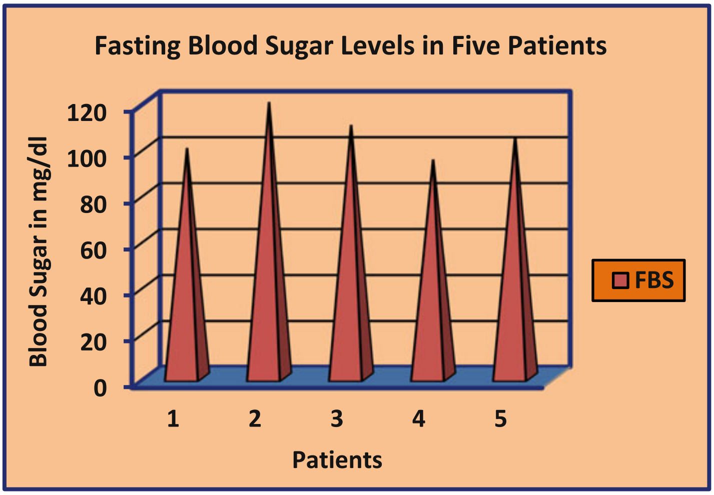


Fig. 4.3 Pictogram exhibiting fasting blood sugar in five patients

2. *Pictograms*: When figures are represented by pictures, these are called pictograms. Sizes and number of pictures are in proportion to the data figures. Data relating to the distribution in various ages is usually represented by "pyramids." Fasting blood sugar levels in five patients have been illustrated with pictogram in Fig. 4.3.

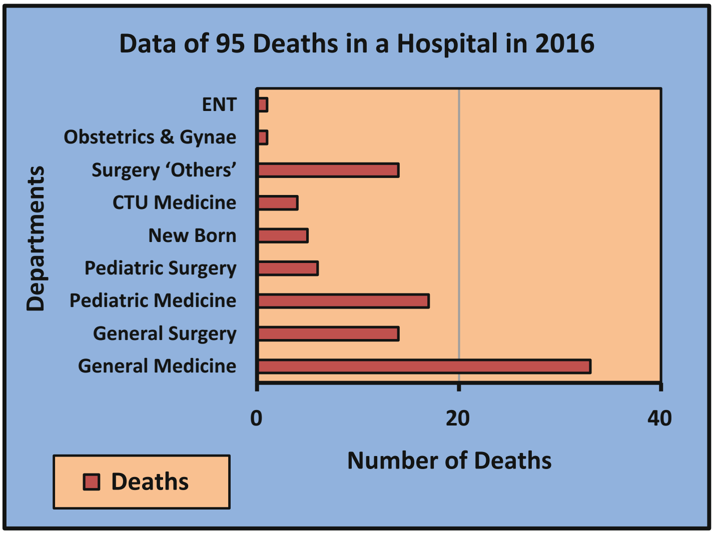


Fig. 4.4 Department-wise data of deaths in a hospital in the year 2016

3. *Cartograms*: Here maps are drawn, and the data figures representing the events/ phenomena are shown by signs or symbols.
4. *Graphs and curves*: Graphs can be in natural scale or ratio scale. In simple bar diagram, one bar would represent one figure, and as such, the number of bars will correspond to the number of observations. The thickness of bars is not taken into account. Width of bars should be uniform. Bars should be at equal distance. Bars can be drawn vertically or horizontally. If the data are not in the shape of time series, the observations may be arranged in ascending or descending order to accomplish comparison. Department-wise data regarding 95 deaths in a hospital in the year 2016 have been illustrated with bars in Fig. 4.4.
5. *Circles or pie diagrams*: Circles occupy a unique place among two-dimensional diagrams. The area of circles is directly proportional to the "square" of its "radius." For construction of circles, the "square roots" of various data figures are calculated first. Circles look more presentable than squares and are easy to be drawn. The aggregates can be presented by a big circle and the various components by sectors cut inside it. Such diagrams are known as "angular diagrams." Department-wise data regarding 95 deaths in a hospital in the year 2016 have been depicted with pie diagram in Fig. 4.5.

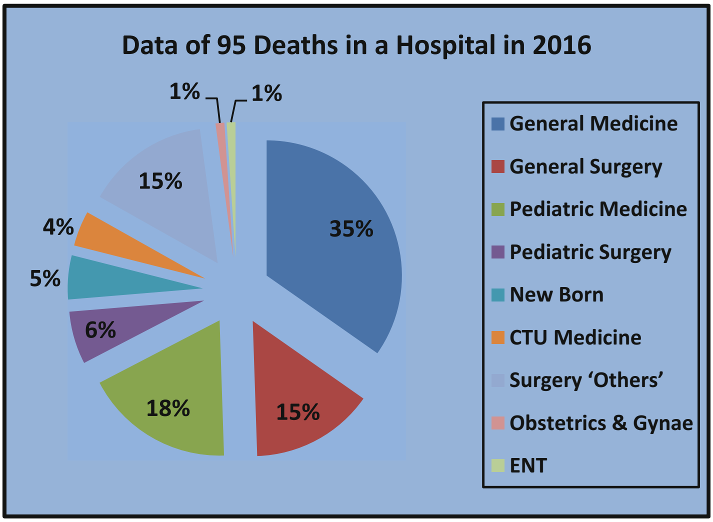


Fig. 4.5 Department-wise data of deaths in a hospital in the year 2016

5 Graphic Presentation of Data

5.1 Construction of Graph

1. In the construction of graph, two simple lines are first drawn, which cut each other at right angles. These are called axis. The horizontal line is called “abscissa” or x-axis, and the vertical line is called “ordinate” or y-axis. The point of intersection is called the point of origin. The basic construction of graph has been depicted in Fig. 5.1.
2. Draw a graph for practice keeping in mind the values of x and y as illustrated above.

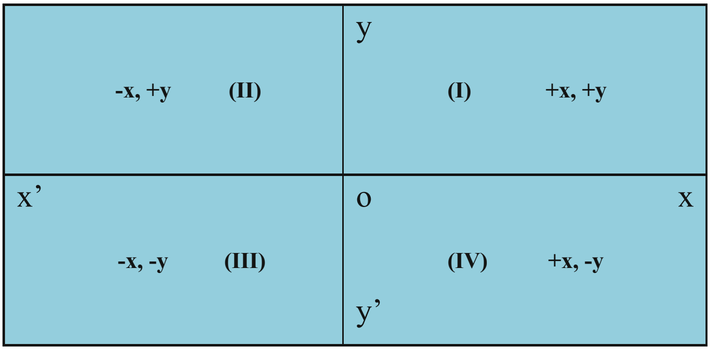


Fig. 5.1 In the above figure, x’ox is the abscissa and yoy’ is the ordinate. “o” is the point of origin. In the quadrant “I,” values of x and y are positive. In the quadrant “II,” values of x are negative, and values of y are positive. In the quadrant “III,” values of x and y are negative. In the quadrant “IV,” values of x are positive, and values of y are negative

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_5

5.2 Choice of Scale

1. The graph should be condensed in the space provided, choosing the appropriate scale for x-axis and y-axis. Examining the magnitude and sign of values, one can determine the positions of x-axis and y-axis.
2. Generally an independent variable is shown on the x-axis and a dependent variable on the y-axis. The horizontal scale need not begin with zero, but the vertical scale must begin with zero. If the fluctuations in the values of a variable presented on the y-axis are very small as compared to the size of the item, a false baseline is used.
3. In the natural scale, equal space represents the equal amount of magnitude. There is no definite relationship between the lengths of "abscissa" and the length of the "ordinate." It is a convention that x-axis is taken 1.5 times of the length of y-axis, but there is no strict rule.

5.3 Plotting of Data

1. After the scales have been decided and marked on the graph paper, the data can be plotted. If the variable is a continuous one, these points may be joined to give a smooth curve. If the data relates to the discrete variable, the points may be joined by straight lines.
2. Sometimes it is very difficult to smooth curves even in case of continuous variable, so the data may be shown by joining the points with straight lines. However, curves obtained by mathematical relationships must be smoothed, and these should not be shown by joining with straight lines.

5.4 Graphs of Time Series or Historigrams

1. Line charts are used to present the historical data due to their superiority over "bar charts." Line charts show the continuity of the variables, whereas "bar charts" indicate the discontinuity of the variables.
2. More time is consumed in depicting the data in the form of bars than in the form of "line chart." Line charts give quick understanding about the movement and the absolute magnitude of variable. Secondly, it is possible to interpolate a value from the graph, for the year for which the data was not available.
3. The graph shows the changes in the values of a variable. If the absolute values are taken into consideration, the graph drawn is known as "absolute historigram." If the graph is plotted from the index numbers, it is then called "index historigram."

5.5 Comparison of Time Series

1. If two or more variables, expressed in the same unit of measurement, are to be compared, these may be presented on the same graph choosing the scales suitable for all the variables. The same variable at two or more periods from the same place may also be represented on the same graph for comparison.
2. In order to compare the changes between different series, it is necessary to have a common base year. Equating the frequency of base year to 100 in each series, the percentage of other frequencies can be worked out. These percentages of each series may be plotted and the lines drawn to show the changes taking place. Thus, the relative changes of all the series can be compared.
3. Sometimes "mixed graphs" are prepared to study interrelated variables. In such graphs, one variable is usually shown in a "bar diagram" and the other in the shape of a "line diagram" or curve.

5.6 Semilogarithmic Scale

1. In case of natural scale, the absolute changes in a variable are studied. But to study relative changes of two or more variables or to study rate of change in a variable, a semilogarithmic scale is appropriate. In this case, y-axis is scaled logarithmically, whereas x-axis is in the linear scale. In the semilogarithmic scale (ratio scale), the length of the interval between two values is proportional to the ratio between these two values. Hence, equal spaces in the ratio scale represent equal logarithmic differences or equal ratios.
2. When to use semilogarithmic scale: (a) when comparison between series of widely different magnitudes is desired, (b) when comparison between series of different units is required, and (c) when relative change is to be studied or compared with various series.

5.7 Interpretations of Semilog Curves

1. If a semilog curve rises upward, it indicates that the growth is positive and the values are increasing.
2. If the curve is a straight line and is ascending, it indicates that the curve is increasing with a constant rate of growth.
3. If the curve is a straight line and is descending, it indicates that the curve is decreasing with a constant rate.
4. If the curve rises more strongly at one point and then at another, it indicates that the rate of growth is more in the former case than in the latter case.
5. If two curves are parallel to each other, their rate of increase or decrease remains the same.
6. If one curve is steeper than the other, its rate of change is higher than the other.

5.8 Properties of a Logarithmic Scale

1. Equal distances on the vertical scale represent equal proportionate changes.
2. Logarithmic scale does not begin with zero.
3. It is not possible to study an aggregate in its component parts.
4. Logarithmic scale cannot be used for studying absolute changes.

5.9 Normal Frequency Curve: Properties of Normal Frequency Curve Are Given Below

1. It is a unimodal, perfectly symmetrical bell-shaped curve.
2. Frequency decreases on either sides of the central value.
3. Frequency at the central value is the highest.
4. Mean, median, and mode coincide with the central value.
5. Frequencies of this distribution are in a definite mathematical relationship.
6. Total area under the normal frequency curve is equal to the total number of observations.
7. Number of observations falling between any two ordinates could be estimated.
8. Areas between "population mean" (μ) and "standard deviation" (σ) have been depicted in "tabulated" and "graphic" form in Table 5.1 and Fig. 5.2, respectively.
9. Near about the mean value, the curve is convex toward x-axis, whereas near the two tails, it is concave.
10. The points of inflection (the points where the change in curvature occurs) are at a distance of $\pm 1\sigma$, $\pm 2\sigma$, and $\pm\ 3\sigma$ on either side of the mean value (μ).
11. The values of all odd moments about the mean are zero, so the "skewness" is zero.

Table 5.1 Area between "population mean" (μ) and "standard deviation" (σ)

Area of "population mean ± SD"	Percentage (%) of observations
$\mu \pm \sigma$	68.27
$\mu \pm 1.96\ \sigma$	95.00
$\mu \pm 2\ \sigma$	95.45
$\mu \pm 2.576\ \sigma$	99.00
$\mu \pm 3\ \sigma$	99.73

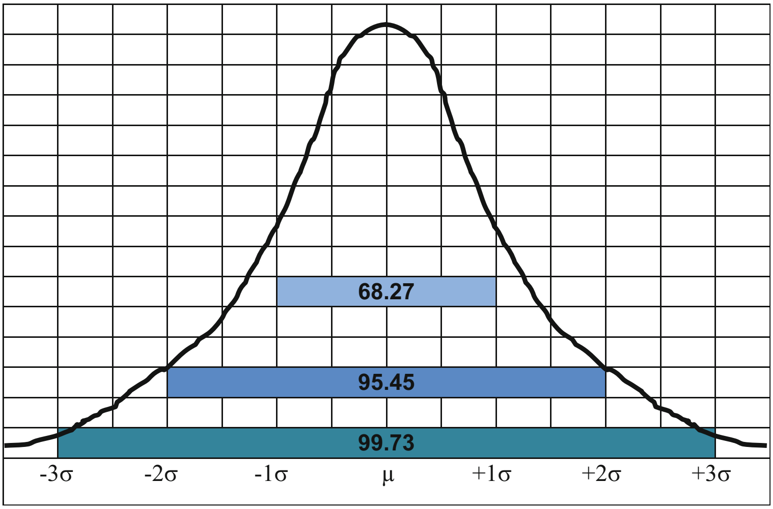


Fig. 5.2 Normal distribution graph showing percentage of observations on either side of "population mean" (μ) in relation to "SD" (σ)

5.10 Moderately Asymmetrical Curve

1. Curve is not symmetrical.
2. The distribution of frequencies in such a curve is not in a mathematical relationship.
3. Curve may either be "positively skewed" or "negatively skewed."

5.11 Extremely Asymmetrical Curve

1. The "skewness" is very high in this type curve.
2. The cluster of observations having maximum frequency is generally on one corner and not in the middle as in the case of symmetrical curve.

Measures of Central Tendency 6

6.1 Properties of Average

1. The average should be calculated for a set of homogenous observations.
2. It is called a measure of location and central tendency.
3. It reduces the complexity of data.
4. It is used as a representative of the whole set of observations and could be used for comparing the sets.

6.2 Characteristics of “Representative Average”

1. It should be easy calculation and simple to understand.
2. It should not be affected by fluctuations of sampling.
3. It should be based on all the observations.
4. It should be capable of “algebraic treatment.”
5. It should be rigidly defined.

6.3 Types of Averages

1. *Mean*: (a) arithmetic mean ($\bar{x}$), (b) geometric mean ($\bar{G}$), and (c) weighted mean ($\bar{\bar{x}}$)
2. *Median*: ($\tilde{x}$)
3. *Mode*: ($\widehat{x}$)

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_6

6.4 Arithmetic Mean ($\bar{x}$)

The arithmetic mean is the most common measure of "central tendency" in the case of continuous variables. It is an appropriate measure for symmetrical distributions. If the sample is assumed to have come from normal population, the arithmetic mean would be the most efficient measure of the population mean (μ). It is a widely used measure.

6.5 Arithmetic Mean ($\bar{x}$) of Ungrouped Data

The arithmetic mean is obtained by dividing the sum of all the observations by the number of observations. Let X_1, X_2, X_3, . . . X_n be the values of N observations.

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

Example Fasting "blood sugar levels" observed in five patients were 100 mg/dl, 120 mg/dl, 110 mg/dl, 95 mg/dl, and 105 mg/dl. Arrange the data in tabulated form and find out arithmetic mean.

Solution
The data has been tabulated in Table 6.1 and arithmetic mean has been worked out.

6.6 Arithmetic Mean ($\bar{x}$) of Grouped Data

When the observations/subjects are more, the process of summation becomes tedious. In such cases frequency distribution tables are prepared to simplify the calculations. Data is "grouped," and "midpoint" (x) of each group is ascertained. The number of "observations" falling in a particular group is called "frequency" (f). The product (xf) would give "sum" of observations in a group, and "sum" of all xf values

Table 6.1 Fasting blood sugar levels in five patients

Patients' ID number ($N = 5$)	Blood sugar level (mg/dl)
1.	100
2.	120
3.	110
4.	95
5	105
Total	**530**
$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$	**106**

Table 6.2 Total circulating albumin in g in 30 normal males as per body mass, aging 18–30 years

Circulating albumin	Midpoint	Frequency	Product
"g"	x	f	xf
99.5–105.5	104.5	3	313.5
109.5–119.5	114.5	6	687.0
119.5–129.5	124.5	5	622.5
129.5–139.5	134.5	8	1076.0
139.5–149.5	144.5	6	867.0
149.5–159.5	154.5	2	309.0
Total		**30**	**3875**
Mean			**129.1**

would give "sum of observations" (Σxf). The mean is calculated through dividing the "sum of observations" (Σxf) by "sum of frequencies" (Σf).

Example Absolute/total value of circulating "albumin" in grams (g) was computed as per "body mass" in 30 normal adults aged between 18 and 30 years.

Solution
The data has been grouped in six groups as shown in column 1 of Table 6.2. Midpoints (x), frequencies (f), and products (xf) have been exhibited in columns 2, 3, and 4, respectively.

$$\bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{3875}{30} = 129.1$$

The "mean" obtained from grouped data may not exactly tally with that obtained from ungrouped data. This is because the observations in each class do not exactly average to the "midpoint." This difference is slight if the observations are more.

6.7 Shortcut Method for Calculating "Mean" ($\bar{x}$)

The calculations of mean ($\bar{x}$) involve large numbers. This not only increases the labor but also chances of error. To minimize the labor, the following shortcut method is suggested.

The transformation method is used to convert x-values (observations) of the variable into t-values as a new variable. A "mean" (x_o) is assumed for transformation of data with reference to assumed class interval "c." This transformation is used in case the width of each class is constant and is equal to "c." However, if class interval is not the same in all the classes, "c" may be considered equal to 1. The transformation formulae could be as below:

When class intervals are equal as "c":

$$t = \frac{X - X\text{o}}{C}$$

When "c" = 1

$$t = \frac{X - X\text{o}}{1}$$

Example Let us focus our attention on the data given in Table 6.2 to learn the transformation method of calculating mean ($\bar{x}$). The class interval in this case is constant as 10 g; so, $c = 10$. Class midpoints represent "x." Assume x_o is 144.5. Making use of "transformation" $t = \frac{X - X_\text{o}}{C}$, we would get the values of new variable "t" corresponding to the class midpoint in the x-variable as illustrated in Table 6.3.

Charlier's Check

We can check the accuracy of calculations by applying Charlier's equation to the outcome of "transformation method."

Charlier's Equation: $\Sigma(t + 1)f = \Sigma tf + \Sigma f$

In the above case: Total of Col. '5' = Total of Col. '4' + Total of Col. '3'

$$-16 = -46 + 30 = -16$$

Therefore: L.H.S. = R.H.S., Hence calculations are accurate

$$\bar{t} = \frac{\Sigma tf}{n} = \frac{-46}{30} = -1.53$$
$$\bar{x} = X\text{o} + c\bar{t} = 144.5 + 10(-1.53) = 144.5 - 15.3 = \mathbf{129.2}$$

The mean ($\bar{x}$) obtained by the direct method was the same as obtained by shortcut method.

Table 6.3 Transformation method for calculating mean

(1)	(2)	(3)	(4)	(5)
Midpoint: X	$t = \frac{X - X_\text{o}}{C}$	Frequency: f	tf	$(t + 1)f$
104.5	−4	3	−12	−9
114.5	−3	6	−18	−12
124.5	−2	5	−10	−5
134.5	−1	8	−8	0
144.5	0	6	0	6
154.5	+1	2	+2	4
Total		**30**	−46	−16

6.8 Merits and Drawbacks of Mean ($\bar{x}$): As Listed in Table 6.4

Table 6.4 Merits and drawbacks of mean ($\bar{x}$)

	Merits	Drawback
1.	Easy for calculation and understanding	Mean $\bar{x}$ is not so easy to calculate as $\tilde{X}$ or $\widehat{X}$
2.	Least affected by fluctuations of sampling	Abnormal items considerably affect
3.	Based on all the observations	Greater importance is to larger item
4.	Algebraic treatment is applicable	Not possible to calculate if items are missing
5.	–	Mean $\bar{x}$ may not exist in the series

6.9 Geometric Mean ($\bar{G}$)

Geometric mean is calculated when the variable follows geometric progression or the law of exponential growth. Geometric mean is an appropriate measure of central tendency.

Let X_1, X_2, X_3, … X_n be the observations. Then the geometric mean $\bar{G}$ will be calculated as given below:

$$\bar{G} = \frac{n\sqrt{x_1.x_2.x_3 \ldots x_n.}}{1}$$

Taking logarithms on both sides, we get:

$$\text{Log}\,\bar{G} = \frac{1}{n}\log\,(X_1, X_2, X_3 \ldots X_n)$$

Therefore: $\bar{G} = \text{antilog}\,\frac{1}{n}\log(X_1, X_2, X_3 \ldots X_n)$

Logarithm of the “geometric mean” ($\bar{G}$) is equal to the “arithmetic mean” of the logarithms of the values OR “geometric mean” ($\bar{G}$) is the “antilog” of the “arithmetic mean of the logarithms” of the observed values of the variable.

6.10 Usefulness of Geometric Mean ($\bar{G}$)

The “geometric mean” ($\bar{G}$) is useful in many biologic applications of statistics. This is because many biological variables can be observed to be very closely distributed in terms of “logarithmic values.”

In the case of dilution assays, it is more appropriate to calculate the “geometric mean” than the “arithmetic mean.” The occurrence or nonoccurrence of a specified

Table 6.5 Doubling dilutions illustrated

Tube no. OR microwell no. followed by dilution factor									
1	2	3	4	5	6	7	8	9	10
1:d_1	1:d_1^2	1: d_1^3	1: d_1^4	1: d_1^5	1: d_1^6	1: d_1^7	1: d_1^8	1: d_1^9	1: d_1^{10}
1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	1:512	1:1024

Table 6.6 Widal test agglutination titer in 25 patients

Tube No.	Titer (x)	Frequency (f)	Log$_{10}$ x (y)	Log$_2$ x (z)
1.	2	0	0.301(0)	1
2.	4	0	0.602(0)	2
3.	8	0	0.903(0)	3
4.	16	2	1.204(2)	4
5.	32	6	1.505(6)	5
6.	64	3	1.806(3)	6
7.	128	5	2.107(5)	7
8.	256	4	2.408(4)	8
9.	512	3	2.709(3)	9
10.	1024	2	3.010(2)	10
Total		**25**	**51.17**	

reaction for each of the several tubes or microwells having geometrically increasing dilutions is noticed. If the initial dilution is 1:1 and the dilution factor is "d," then the first tube would contain 1: 1:d_1 (mix equal volume diluents and serum); later on, tubes will contain $1 : d_1^2, 1 : d_1^3, 1 : d_1^4 \ldots.$ and so on.

So in the case of doubling dilution assay (where $d = 2$), the dilutions will be as illustrated in Table 6.5.

Example Widal test for *Salmonella typhi* was done in 25 patients using doubling dilutions of sera of patients, and agglutination reactions were recorded after overnight incubation in a water bath at 37 °C. Frequency of occurrence of agglutination titer has been exhibited in Table 6.6.

Arithmetic mean in the above case comes out to be

$$\bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{5664}{25} = 226.56$$

Geometric mean in the above case would be

$$\bar{G} = \text{antilog}\ \tfrac{1}{n} \log\ (X_1, X_2, X_3 \ldots X_n) = \text{antilog}\ \tfrac{51.17}{25} = \text{antilog}\ 2.0468$$

$$\bar{G} = \text{antilog}\ 2.0468 = 111.4$$

6.11 Merits and Drawbacks of Geometric Mean ($\bar{G}$): As Listed in Table 6.7

Table 6.7 Merits and drawbacks of geometric mean ($\bar{G}$)

	Merits	Drawback
1.	Geometric mean is rigidly defined in its value as a precise figure	Geometric mean is neither easy to calculate nor easy to understand
2.	It is based on all the observations	If any value in a series is zero or negative, geometric mean can't be calculated
3.	Algebraic treatment is applicable	It may not exist in series
4.	It is not much affected by fluctuations of sampling	Smaller items can't be given more weightage

6.12 Weighted Mean

The weighted mean is calculated in the following cases:

1. When the value of a variable have different frequencies
2. When different multipliers (weightages) have been given to different items to express their importance
3. While solving complex statistical problem where combined average has to be worked from several means of different weightages

If X_1, X_2, … X_k are the values of a variable with weights W_1, W_2 … W_k, respectively, then the weighted mean can be calculated as follows:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} x_i} = \frac{\Sigma WX}{\Sigma W.}$$

If $\bar{X}_1\, \bar{X}_2 \ldots \bar{X}_k$ represents the means of the observations $n_{1,}\, n_2 \ldots n_k$, respectively, then "weighted mean" will be "sum of all the means" divided by number of means.

$$\bar{\bar{X}} = \frac{\Sigma\bar{X}}{k}$$

Table 6.8 Crude death rates of towns A and B

	Town A			Town B		
Age groups (in years)	Population	No. of deaths	Age sp. death rate	Population	No. of deaths	Age sp. death rate
0–10	5000	100	20	4000	80	20
10–40	15,000	150	10	10,000	110	11
40 plus	5000	100	20	4000	80	20
Total	25,000	350		18,000	270	
CDR	CDR = $\frac{350}{25{,}000} \times 1000$		**14**	CDR = $\frac{270}{18{,}000} \times 1000$		**15**

6.13 Crude Death Rate (CDR)

Crude death rate could be defined as the number of deaths occurred in a geographical region during a given time period *per thousand* in the midyear estimated population during the same period/year. It could be calculated as illustrated below:

$$\text{CDR} = \frac{\text{Number of deaths in a geographical region in a given year}}{\text{Mid year estimated population of that region during that year}} \times 1000$$

In the same way, CDR in different age groups can be determined using the same parameters. Crude death rates of towns A and B are illustrated in Table 6.8.

6.14 Standardized Death Rate (SDR)

The “total population” and its composition differ from locality to locality. So, “crude death rates” (CDRs) obtained are unequally weighted and are not comparable. To make them comparable, equal weightage should be given for both the death rates in each specific age group. Age composition of any one town or of the entire country may be taken as standard population to weigh the death rates in both the localities under comparison. This process is called “standardization,” and the death rates obtained by this process are called “standardized death rates.” Standardized death rates of towns A and B are illustrated in Table 6.9.

6.15 Median ($\tilde{x}$)

Median is the central value of a variable arranged in an array. It divides the data into two equal halves.

Table 6.9 Standardized death rates of towns A and B

	Town A			Town B		
Age groups (in years)	Standard population	Age sp. death rate	No. of deaths in standard population	Standard population	Age sp. death rate	No. of deaths in standard population
0–10	400	20	8	400	20	8
10–40	500	10	5	500	11	5.5
40 plus	100	20	2	100	20	2
Total	1000			1000		
Standardized death rate			**15**	**SDR**		**15.5**

6.16 Usefulness of Median ($\tilde{x}$)

Median is an appropriate measure of central tendency, when:

1. The distribution of variable is asymmetrical or skewed.
2. The variable is qualitative.
3. The end classes are open.
4. The class intervals are unequal.

6.17 The Computation of Median ($\tilde{x}$) of Unclassified Data

Let $X_1, X_2, \ldots X_n$ be a set of n observations.

Let $X_1^*, X_2^*, \ldots X_n^*$ be the observations of the same set arranged in an ascending order of magnitude. Then

$$\text{Median}\left(\tilde{x}\right) = x^* \frac{(n+1)}{2} \left(?^{\text{th}}\text{item}\right) \ldots \text{when } n \text{ is odd, and Median}\left(\tilde{x}\right)$$

$$= \frac{1}{2}\left[x^* \frac{n}{2} + \left(x^* \frac{n}{2} + 1\right)\right] \left(?^{\text{th}} \text{ item}\right) \ldots \text{when } n \text{ is even.}$$

Suppose the total number of observations in the data, arranged in ascending order, are 21 (odd), then median ($\tilde{x}$) value will be $\left(\frac{21+1}{2}\right)$th observation, i.e., 11th observation in this case.

Now suppose, if the total of observations in the data, arranged in ascending order, are 30 (even), then median ($\tilde{x}$) value will be $\frac{1}{2}\left[x^* \frac{n}{2} + \left(x^* \frac{n}{2} + 1\right)\right]$, i.e., mean of 15th and 16th observation.

6.18 The Computation of Median ($\tilde{x}$) of Classified Data

While calculating the median ($\tilde{x}$) from a frequency table, it is assumed that all values within a class interval are evenly placed. Find the illustration in Table 6.10, where "total circulating albumin" in grams in males of age group of 20–29 years has been depicted. There are six groups of range of "total circulating albumin."

Their frequencies are arranged in a cumulative frequency table. The middle item, i.e., $\frac{n}{2}$ th falls in the class 129.5–139.5.

For less than 129.5, there are 14 observations.

For less than 139.5, there are 21 observations.

Corresponding to 15th item (middle item), we have to find a value indicating the median, assuming that the observations are equally distributed in the range starting from 129.5 to 139.5. By applying the theory of proportionate, we get

$$\frac{\tilde{x} - 129.5}{139.5 - 129.5} = \frac{\mathbf{15} - 14}{21 - 14}$$

$$\begin{aligned}\text{Therefore, } \tilde{x} &= 129.5 + \frac{\mathbf{139.5} - 129.5\,(15 - 14)}{21 - 14} = 129.5 + \frac{10}{7}\\ &= 129.5 + 1.43 = 130.93\,\text{g}.\end{aligned}$$

6.19 Merits and Drawbacks of Median ($\tilde{x}$)

	Merits	Drawbacks
1.	Rigidly defined	Not suitable for algebraic treatment
2.	Can be calculated easily and simple to understand	Method of interpolation has to be applied to calculate median ($\tilde{x}$) in continuous series
3.	It is not affected by values of extreme items	If large or small items are to be given importance, then median ($\tilde{x}$) would be unsuitable
4.	Not necessary to know the distribution in entire data	More likely to be affected by sampling fluctuations
5.	It is appropriate measure for data on linear scale	Arranging the data in an array is sometimes very tedious

6.20 Mode ($\hat{x}$)

Mode is the value of a variable that occurs more frequently in a data. It is an observation having high frequency. When one is interested to study the "most frequent cause of death" or "most frequent symptom of disease," the modal ($\hat{x}$) value is being calculated. Mode ($\hat{x}$) may be in the middle, in the beginning, or toward the end of set of observations.

Table 6.10 Total circulating albumin in males of 20–29 years

Total circulating albumin (g)	Frequency (f)	Cumulative frequency below upper limit of interval
99.5–109.5	2	2
109.5–119.5	6	8
119.5–129.5 (a)	6	14
129.5–139.5 (b)	7	21
139.5–149.5	8	29
149.5–159.5	1	30
Total	30	

6.21 Calculating Mode ($\widehat{x}$) in Ungrouped Data

Let the values of the observations be 15, 12, 13, 15, 15, 11, 14, and 16. Here, 15 have occurred thrice, whereas the other values occurred only once. Hence $\widehat{x} = 15$.

6.22 Calculating Mode ($\widehat{x}$) in Grouped Data

Let us consider Table 6.10 for the grouped data. The highest frequency "8" lies in the class 139.5–149.5. Therefore, modal value lies in this class. But to determine the value of mode ($\widehat{x}$), three classes have to be considered.

1. Highest frequency class (139.5–149.5)
2. Class preceding to highest frequency class (129.5–139.5)
3. Class succeeding to highest frequency class (149.5–159.5)

Let us tabulate the required classes for convenience:

Class (total circulating albumin)	Frequency (f)	New symbol	Transformation
129.5–139.5	7	f_0	
139.5–149.5	8	f_1	$\Delta_1 = f_1 – f_0 = 1$
149.5–159.5	1	f_2	$\Delta_2 = f_1 – f_2 = 7$
Class interval (C) = 10			

Modal value ($\widehat{x}$) can be obtained by the following formula:

$$\begin{aligned} \widehat{x} &= 139.5 + \frac{\Delta 1}{\Delta 1 + \Delta 2} \times C = 139.5 + \frac{1}{1+7} \times 10 \\ &= 139.5 + \frac{10}{8} = 139.5 + 1.25 = 140.75 \end{aligned}$$

6.23 Properties of Mode ($\hat{x}$)

1. It represents most typical value.
2. It is not affected by extreme values.

6.24 Comparison of Mean ($\bar{x}$), Median ($\tilde{x}$), and Mode ($\hat{x}$)

1. Mean, median, and mode are easy to be calculated and simple to understand.
2. Mean and median are more stable and rigid in comparison to mode.
3. Mean is calculated by considering all observations. But median and mode are calculated from few values only.
4. Mean is affected by extreme values but median and mode are not.
5. Mean is capable of algebraic treatment, whereas median and mode are not.
6. Mean and median are affected by fluctuations of sampling, whereas mode is not.
7. Mean has definite advantage over median and mode as it possesses certain sampling properties. In some cases median and mode have preference over mean value.
8. Median is used when variable is more qualitative than quantitative as well as also when the end classes are open.
9. Mean and median have minimal properties, whereas mode has maximal property.
10. In the case of symmetrical distributions, mean, median, and mode coincide.
11. In the case of moderately skew distributions:

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

6.25 When to Use Mean ($\bar{x}$), Median ($\tilde{x}$), and Mode ($\hat{x}$)?

6.25.1 Use of Mean ($\bar{x}$)

1. When the values are distributed symmetrically around the central point or the distribution is moderately skew, mean ($\bar{x}$) is calculated. Mean ($\bar{x}$) determines the center of gravity of the distribution.
2. When greatest stability is required.
3. When other statistical parameters are to be computed.

6.25.2 Use of Median ($\tilde{x}$)

1. When exact midpoint of distribution is required
2. When extreme values markedly affect the mean

6.25.3 Use of Mode ($\hat{x}$)

1. When a quick and approximate measure of central tendency is required.
2. When the measure of central tendency should be the most typical value. For example, most frequent style of dress, most frequent cause of death, most frequent symptom of a disease, etc.

6.26 Quartiles, Deciles, and Percentiles

Quartiles

The median ($\tilde{x}$) divides the array into two parts, each part having equal number of observations. We may choose three points to divide the array into four parts, each part containing equal number of observations. These three points are called Q_1, Q_2, and Q_3. The 25% of observations will be below Q_1 and 25% beyond Q_3. Q_2 divides the observations into two equal halves: 50% below this and 50% above this. From the above deliberations, it is clear that $\tilde{x} = Q_2$.

Now consider a cumulative frequency in Table 6.11, wherein Q_1 would correspond to $\frac{n}{4}$th item. Q_3 would correspond to $\frac{3n}{4}$th item. To find out the values of Q_1 and Q_3, the same formula is used as for finding out the median ($\tilde{x}$) is applied.

$$\begin{aligned} Q_1 &= a + \frac{b-a}{f(b)-f(a)} = 109.5 + \frac{129.5 - 119.5}{14-8} \\ &= 109.5 + \frac{10}{6} = 109.5 + 1.67 = 111.17 \\ Q_3 &= c + \frac{d-c}{f(d)-f(c)} = 139.5 + \frac{149.5 - 139.5}{29-21} \\ &= 139.5 + \frac{10}{8} = 139.5 + 1.25 = 140.75 \end{aligned}$$

Deciles

We may choose nine points to divide the array into ten parts, each part having equal number of observations. These nine points are called deciles. These may be denoted

Table 6.11 Total circulating albumin in males of 20–29 years

Total circulating albumin (g)	Frequency (f)	Cumulative frequency below upper limit of interval
99.5–109.5	2	2
109.5–119.5 (a)	6	8 $f(a)$
119.5–129.5 (b)	6	14 $f(b)$
129.5–139.5 (c)	7	21 $f(c)$
139.5–149.5 (d)	8	29 $f(d)$
149.5–159.5	1	30

by $D_1, D_2 \ldots D_9$. But $D_5 = Q_2 = \tilde{x}$. The same formula can be used to find the values of $D_1, D_2 \ldots D_9$ but corresponding to $\frac{n}{10}$th item, $\frac{2n}{10}$th item $\ldots$ $\frac{9n}{10}$th item, respectively.

Percentiles

We may choose 99 points to divide the array into 100 parts, each part containing equal number of observations. These are called percentiles. These are denoted by P_1, $P_2 \ldots P_{99}$. But $P_{50} = D_5 = Q_2 = \tilde{x}$. The same formula can be used to find out the values of P_1, $P_2 \ldots P_{99}$ corresponding to $\frac{n}{100}$th item, $\frac{2n}{100}$th item $\ldots$ $\frac{99n}{10}$th item, respectively.

Measures of Dispersion

7

7.1 Definition

The dispersion may be defined as "the extent to which the magnitudes or qualities of the items differ," that is, the degree of diversity. If the object to be described is a single distribution, the absolute dispersion may be computed. But to compare two or more distributions, relative dispersion is necessary. Absolute dispersions are always expressed in the units the measurements are made. But the relative dispersion is a pure number expressed as a ratio or percentage. Dispersion is also expressed by fluctuation, spread, scatter, or variation (Fig. 7.1).

7.2 Various Measures of Dispersion

1. Range
2. Quartile deviation
3. Mean deviation
4. Standard deviation

7.2.1 Range

Range is obtained by calculating difference between the highest and the lowest observation of the set. It is the simplest measure based on only two extreme items.

$$\text{Range} = \text{Highest value} - \text{Lowest value}.$$

If the distributions to be compared are expressed in the same unit of measure, their dispersions can be compared with reference to their range. The greater the range, the greater is the variation in values of the group. If the distributions to be

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_7

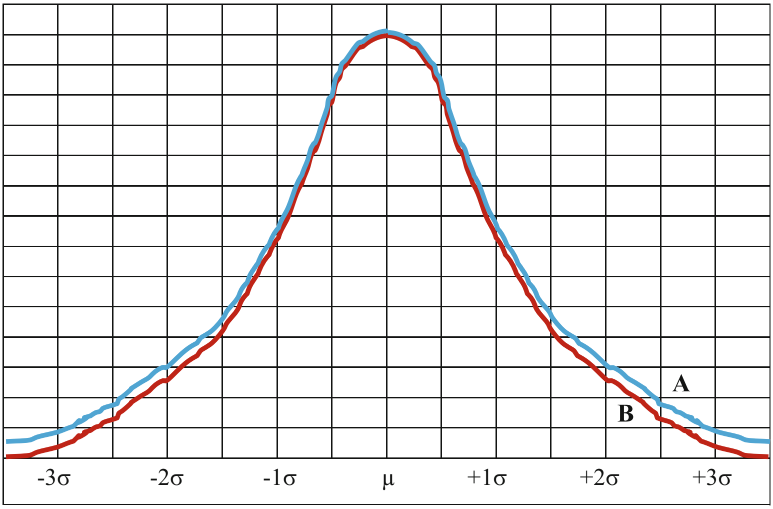


Fig. 7.1 Both distributions have been exhibiting same measure of central tendency

compared are expressed in different units of measure, it would not be possible to compare these by absolute dispersion. In that case relative dispersions are to be calculated. Relative dispersion may be obtained by dividing the range by the sum of extreme items.

$$\text{Relative Dispersion} = \frac{\text{Difference of extreme items}}{\text{Sum of extreme items}}$$

As "range" is determined only by two extreme values, it does not take cognizance of the manner in which the values are distributed within the range. Hence, the use of "range" is restricted to few fields as quality control charts, daily temperature (maximum and minimum), stock exchange quotations, amount of rainfall, etc.

7.2.2 Quartile Deviation

The difference between the third quartile (Q_3) and the first quartile (Q_1) is known as "interquartile range" (Q_3 – Q_1). This "range" indicates the middle 50% of observations. The larger the "interquartile range" ($Q_3 - Q_1$), the larger is the absolute variability. The "interquartile range" ($Q_3 - Q_1$) when divided by two, we get "semi-quartile range" $\frac{1}{2}(Q_3 - Q_1)$. Semi-quartile range is known as "quartile deviation" (QD).

$$\text{Thus QD} = \frac{1}{2}(Q_3 - Q_1)$$

In the case of symmetrical distribution, the "quartile deviation" (QD) on either side of median ($\tilde{x}$) is indicated by the value of Q_1 and Q_3.

$$\tilde{x} - \text{QD} = Q_1 \text{ and}$$
$$\tilde{x} + \text{QD} = Q_3$$

Therefore, 50% of observations lie between $\tilde{x} \underline{+}$ QD.

If the distributions to be compared are in different units of measure, it is advisable to convert the absolute dispersion into relative dispersion. This can be done by choosing an appropriate standard measure in the denominator. In such a case, median ($\tilde{x}$) is used as preferred denominator to calculate "relative dispersion." Relative dispersion is known as "coefficient of dispersion."

$$\text{Relative Dispersion} = \frac{\text{QD}}{\tilde{x}}$$

Another standard which is commonly used in the denominator is the "average" of the two quartiles.

$$\text{Relative Dispersion} = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

7.2.3 Mean Deviation

The "mean deviation" is based on total number of observations and is free from the defects of variability of observations. It is better than "quartile deviation" and "range" which are based on few observations and suffer from variability of observations. Table 7.1 represents the inadequacy of "range" and "quartiles" as measures of "dispersion."

Range = 74–65 = 9 inches in both the cases

Mean deviation also known as "average deviation" is the average of the absolute deviations from the measures of central tendency. The measure of central tendency employed in this case is the "mean" or "median."

Let X_1, X_2 … X_n be "n" observations. Then $\bar{x} = \frac{x_{1+}x_{2+}\ldots x_n}{n}$ and "absolute deviations" from mean are $x_1 - \bar{x}, x_2 - \bar{x} \ldots x_n - \bar{x}$. Taking average, we get "mean deviation" (MD) about $\bar{x}$.

Table 7.1 Inadequacy of range and quartiles

Group A		Group B	
Order	Height in inches	Order	Height in inches
1.	65	1.	65
2.	68	2.	66
3.	68	3.	67
4.	68 (Q_1)	4.	68 (Q_1)
5.	68	5.	68
6.	70	6.	69
7.	70	7.	69
8.	70 $(\tilde{x})$	8.	70 $(\tilde{x})$
9.	70	9.	70
10.	70	10.	71
11.	70	11.	71
12.	72 (Q_3)	12.	72 (Q_3)
13.	72	13.	73
14.	72	14.	73
15.	74	15.	74

$$\mathrm{MD} = \frac{1}{n}\sum\nolimits_{i=1}^{K} X_i - \bar{X} = \frac{1}{n}\sum\nolimits_{i=1}^{K} x_i$$

where $X_i = X_i - \bar{X}$ and $X_i - \bar{X}$ indicate the "absolute difference" between X_i and $\bar{X}$ values (i.e., the minus sign has not been taken into consideration while taking difference).

Similarly, MD (about $\tilde{X}$) $= \frac{1}{n}\sum_{i=1}^{K} X_i - \tilde{X}$.

In the case of classified data

$$\mathrm{MD}\ \left(\text{about } \tilde{X}\right) = \frac{1}{n}\sum\nolimits_{i=1}^{K} F_i \quad \left|X_i - \tilde{X}\right| = \frac{1}{n}\sum\nolimits_{i=1}^{K} F_i\left|\tilde{X}\right.$$

$$\mathrm{MD}\ \left(\text{about } \bar{X}\right) = \frac{1}{n}\sum\nolimits_{i=1}^{K} F_i \quad |X_i - \bar{X}|$$

Computations of Mean Deviation

Let us consider Group-B in the Table 7.1 to compute "mean" ($\bar{X}$) from unclassified data.

Total height $= \Sigma x = 1046$ and $N = 15$.

Therefore $\bar{X} = \frac{\Sigma x}{N} = \frac{1046}{15} = 69.7$

From the Table 7.1, it is evident that $\tilde{X} = 70$.

Table 7.2 Computation of mean deviation for unclassified data of Group-B from the previous table

Order	Height in inches	$\lvert X - \bar{X} \rvert$	$\lvert X - \tilde{X} \rvert$
1.	65	65–69.7 = 4.7	65–70 = 5
2.	66	66–69.7 = 3.7	66–70 = 4
3.	67	67–69.7 = 2.7	67–70 = 3
4.	68	68–69.7 = 1.7	68–70 = 2
5.	68	68–69.7 = 1.7	68–70 = 2
6	69	69–69.7 = 0.7	69–70 = 1
7.	69	69–69.7 = 0.7	69–70 = 1
8.	70	70–69.7 = 0.3	70–70 = 0
9.	70	70–69.7 = 0.3	70–70 = 0
10.	71	71–69.7 = 1.3	71–70 = 1
11.	71	71–69.7 = 1.3	71–70 = 1
12.	72	72–69.7 = 2.3	72–70 = 2
13	73	73–69.7 = 3.3	73–70 = 3
14.	73	73–69.7 = 3.3	73–70 = 3
15.	74	74–69.7 = 4.3	74–70 = 4
Total	1046	32.3	32
MD		$\frac{32.3}{15} = 2.15$	$\frac{32}{15} = 2.13$

Table 7.3 Computation of mean deviation after classifying data (given $\bar{X} = 70$ and $\tilde{X} = 69.8$)

Height (inches)	Class frequency (f)	Class midpoint (X)	$\lvert X - \bar{X} \rvert$	$(X - \bar{X})f$	$\lvert X - \tilde{X} \rvert$	$(X - \tilde{X})f$
1	2	3	4	5 = (2 × 4)	6	7 = (2 × 6)
65–66	2	65.5	4.5	9.0	4.3	3.6
67–68	3	67.5	2.5	7.5	2.3	6.9
69–70	4	69.5	0.5	2.0	0.3	1.2
71–72	3	71.5	1.5	4.5	1.7	5.1
73–74	3	73.5	3.5	10.5	3.7	11.1
Total	**15**			**33.5**		**32.9**

Let us consider Table 7.2 for learning computations of “mean deviation” from unclassified data and Table 7.3 for learning computations of “mean deviation” from classified data. For classified data values of “mean” ($\bar{X}$) and “median” MD ($\tilde{X}$) have been given as 70 and 69.8, respectively.

Mean deviation (MD) about the median ($\tilde{X}$) is least.

$$\text{MD (about } \bar{X}) = \frac{(X - \bar{X})f}{15} = \frac{33.5}{15} = 2.23$$

$$\text{MD (about } \tilde{X}) = \frac{(X - \tilde{X})f}{15} = \frac{32.9}{15} = 2.19$$

The “mean deviation” about the median is least in above illustration. Hence, it would be more appropriate to calculate the MD about median rather than the MD about arithmetic mean.

Relative measure of “mean deviation” is given by $\frac{\text{MD}}{\tilde{X}} \times 100$. This is also known as “coefficient of variation” (CV) obtained from “mean deviation.” For example, for calculating CV, refer to the data given in Table 7.3:

$$\text{CV} = \frac{\text{MD}}{\tilde{X}} \times 100 = \frac{2.19}{69.8} \times 100 = 3.14$$

7.2.4 Standard Deviation

While finding out “mean deviation,” the signs of the deviations are ignored. The numerical values of the deviations are considered for finding out average regardless of their signs.

The other way of ignoring the signs is by “squaring.” The average derived from squared deviations is called “variance” of the set of observations. The “variance” is also known as “mean square deviation.” If the deviations are taken around the “arithmetic mean,” then the “variance” can be symbolically represented as

For Ungrouped Data

$$V = \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i)^2$$

Where

V: denotes variance
N: number of observations
i: taking values from 1, 2, . . . N
$\bar{X}$: is the arithmetic mean of the set of observations and
$x_i = X_i - \bar{X}$

In the case of “classified data,” the formula for calculating “variance” will be as given below:

$$V = \frac{1}{N} \sum_{i=1}^{K} \left(X_i - \bar{X}\right)^2 f_i = \frac{\sum_{i=1}^{K} (X_i)^2 f_i}{\sum_{i=1}^{K} f_i} = \frac{\sum Xf}{\sum f}$$

Where

$$N = \sum_{i=1}^{K} f_i$$

f_i indicates the frequency of the class for i = 1, 2, . . . k.
k indicates the number of classes in the distribution.

Variance is the second moment about the mean. Other measures of dispersion discussed so far are of the same dimension as the observations, whereas the "variance" is a different measure.

Therefore, "square root" is extracted to compute the "variance" from the sum of squared deviations. This measure of dispersion is known as "standard deviation." It is also known as "root mean square deviation." It can be defined as the square root of the arithmetic mean of the squared deviations of the observations from their arithmetic mean.

7.3 Symbols Representing the Standard Deviation

The "standard deviation" obtained from large number of observations or population data is denoted by sigma (σ), and the same obtained from the sample is denoted by "small-ess" (s).

7.4 Symbols Used in Statistical Analysis

Symbol	Description
σ	Population standard deviation
s	Sample standard deviation
N	Population total observations
n	Sample total observations
f_i	Class frequency for $i = 1, 2, \ldots K$
K	Number of classes
Σ	Summation sign
μ	Population mean
$\bar{x}$	Sample mean
$\sqrt{}$	Square root sign

7.5 Formulae for Calculating Standard Deviation from Ungrouped Data

Formulae are tabulated below:

	Population	Sample
1.	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$
2.	$\sigma = \sqrt{\frac{1}{N}\left(\Sigma X^2 - N\mu^2\right)}$	$s = \sqrt{\frac{1}{n-1}\left(\Sigma X^2 - n\bar{X}^2\right)}$
3.	$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \frac{(\Sigma X)^2}{N}}$	$s = \sqrt{\frac{\Sigma X^2}{n-1} - \frac{(\Sigma X)^2}{n(n-1)}}$
4.	$\sigma = \frac{1}{N}\sqrt{\Sigma X^2 - (\Sigma X)^2}$	$s = \sqrt{\frac{n\Sigma X^2 \quad -(\Sigma X)^2}{n(n-1)}}$

The abovementioned formulae exhibit that in the case of population data analysis, the "variance" is obtained by dividing the "sum of squared deviations" by the number of observations (N). But in the case of a "sample," the "variance" is obtained by dividing the "sum of squared deviations" by the degrees of freedom, i.e., "one less than the number of observations" ($n-1$). By taking the "square root" (√) of "variance," we get "standard deviation."

The estimate obtained from the "sample" by this adjustment gives us an unbiased estimate of population parameter (σ). When "n" is large, division by "n" or "$n-1$" makes negligible difference.

7.6 Computation of Measures of Dispersion

7.6.1 From Ungrouped Data

Example 1 Refer to Table 7.4 below for computation of "mean" and "measures of dispersion" for the age distribution in a hypothetical family.

Mean: $\bar{X} = \frac{125}{5} = 25$

Measures of Dispersion

1. Range = 60–2 = 58
2. Mean Deviation about mean: MD $= \frac{\sum|d|}{N} = \frac{110}{5} = 22$
3. Variance: $V = \frac{\Sigma d^2}{N} = \frac{2668}{5} = 533.6$
4. Standard Deviation: $s = \sqrt{V} = \sqrt{533.6} = 23.1$

Table 7.4 Age distribution in a hypothetical family

Members of family		Age (X)	Mean age ($\bar{X}$)	Deviations from mean: $d = \left(X - \bar{X}\right)$	Square of deviations from mean: $\left(X - \bar{X}\right)^2 = d^2$
1		2	3	4 = (2–3)	5 = 4 × 4
Father		60	25	35	1225
Mother		45	25	20	400
Siblings	(i)	10	25	−15	225
	(ii)	8	25	−17	289
	(iii)	2	25	−23	529
Sum		125		0	2668

Table 7.5 Total circulating albumin in grams

Subjects	Albumin: X	X^2
1.	106	11,236
2.	106	11,236
3.	114	12,996
4.	116	13,456
5.	116	13,456
6.	118	13,924
7.	118	13,924
8.	119	14,161
9.	120	14,400
10.	122	14,884
Sum	1155	1,33,673

Example 2 Refer to Table 7.5 below for computation of "mean" and "standard deviation" of data for "total circulating albumin" in grams in 10 subjects.

Mean: $\bar{X} = \frac{\Sigma X}{N} = \frac{1155}{10} = 115.5$

$$\text{Standard Deviation :} \quad s = \sqrt{\frac{\Sigma X^2 - n\left(\bar{X}\right)^2}{n-1}} = \sqrt{\frac{133,673 - 10(115.5)^2}{10-1}}$$
$$= \sqrt{30.06}$$
$$= 5.48$$

7.7 Formulae for Calculating Standard Deviation from Grouped Data

Formulae are tabulated below:

	Population	Sample
1.	$\sigma = \sqrt{\frac{\sum_{i=1}^{K} f_i\,(Xi-\mu)^2}{\sum_1^K f_i}}$	$s = \sqrt{\frac{1}{n-1}\sum_1^K f_i\left(X_i-\bar{X}\right)^2}$
2.	$\sigma = \sqrt{\frac{1}{N}\left[\Sigma X^2 f - N\mu^2\right]}$	$s = \sqrt{\frac{1}{n-1}\left[\Sigma X^2 f - n\bar{X}^2\right]}$
3.	$\sigma = \sqrt{\frac{\Sigma X^2 f}{N} - \frac{(\Sigma Xf)^2}{N^2}}$	$s = \sqrt{\frac{\Sigma X^2 f}{n-1} - \frac{(\Sigma Xf)\ ^{2}}{n(n-1)}}$
4.	$\sigma = \frac{1}{N}\sqrt{N\left[\Sigma X^2 f - (\Sigma Xf)^2\right]}$	$s = \sqrt{\frac{n\left(\Sigma X^2 f\right)-(\Sigma Xf)^2}{n(n-1)}}$

7.8 Computing Standard Deviation from Classified Data

Example Refer to Table 7.6 below for computation of “mean,” variance, and “standard deviation” of data for “total circulating albumin” in grams in 30 normal males aged between 20 and 29 years.

Mean: $\bar{X} = \frac{\Sigma Xf}{\Sigma f} = \frac{3895}{30} = 129.8$ g.

$$\text{Standard Deviation}: s = \sqrt{\frac{n\left(\Sigma X^2 f\right) - (\Sigma Xf)^2}{n(n-1)}}$$

$$= \sqrt{\frac{30 \times 511{,}047.5 - (3895)^2}{30(30-1)}}\ s$$

$$= \sqrt{\frac{15{,}331{,}425 - 15{,}171{,}025}{870}} = \sqrt{\frac{160{,}400}{870}}$$

$$= \sqrt{\frac{160{,}400}{870}} = \sqrt{184.37} = 13.6$$

Variance: ‘V’ $= s^2 = (\sqrt{184.37})^2 = 184.37$

Table 7.6 Total circulating albumin in males aged between 20 and 29 years

Total circulating albumin (g)	Frequency (f)	Class midpoint (X)	Xf	X^2f
1	2	3	$4 = 2 \times 3$	$5 = 3 \times 4$
99.5–109.5	2	104.5	209.0	21,840.50
109.5–119.5	6	114.5	687.0	78,661.50
119.5–129.5	6	124.5	747.0	93,001.5
129.5–139.5	7	134.5	941.5	126,631.75
139.5–149.5	8	144.5	1156.0	167,042.00
149.5–159.5	1	154.5	154.5	23,870.25
Total	30		3895.0	511,047.50

7.9 Shortcut Method for Computing "Standard Deviation" from Classified Data

Steps

1. Transform the variable "*X*" into new variable "*t*" by following the procedure mentioned below if the "class intervals" are equal magnitude.

$$t = \frac{X - X_o}{c}$$

Where

t is a new variable.
X is the original variable.
X_o is an assumed variable.
C is the class interval.
If the class intervals are not of equal size, then "*c*" may be equated to one (1).

2. Then calculate $\bar{t}$ in place of $\bar{X}$ or μ, by the formula

$$\bar{\mathbf{t}} = \frac{\mathbf{\Sigma tf}}{\mathbf{\Sigma f}}$$

3. Calculate variance of new variable "*t*" by using the formula

Table 7.7 Asymptomatic persons in various age groups investigated for "intestinal parasitism"

Age groups X	Persons F	t	tf	t^2f	$(t+1)^2f$
1	2	3	4 (2 × 3)	5 (3 × 4)	6
1–11	73	−3	−234	702	312
11–21	177	−2	−354	708	177
21–31	347	−1	−347	347	0
31–41	206	0	0	0	206
41–51	112	+1	112	112	448
51–61	52	+2	104	208	468
61–71	24	+3	72	216	384
71–81	4	+4	16	64	100
Total	1000		−631	2357	2095

$$V_t = \frac{1}{N}\left[\frac{\Sigma t^2 f}{1} - \frac{(\Sigma tf)^2}{N}\right] \text{ – in case of population.}$$

$$V_t = \frac{1}{n-1}\left[\frac{\Sigma t^2 f}{1} - \frac{(\Sigma tf)^2}{n}\right] \text{ – in case of sample.}$$

4. Find the "square root" (√) of the "variance" (V_t) to get the "standard deviation" of the new variable "t," to be denoted by σ_t in the case of "population" and by s_t in the case of case of "sample" study.
5. Retransform the "standard deviation" of "t" variable into that of "X" variable.
   $\sigma_x = c.\sigma_t$ (i.e., product of c and σ_t)
   $s_x = c.s_t$ (i.e., product of c and s_t)

where "c" is the class interval.

Example Refer to Table 7.7 for computation of "mean," variance, and "standard deviation" by shortcut method in classified data of 1000 asymptomatic persons investigated for "intestinal parasitism" in India.

Charlier's Check

$$\Sigma(t+1)^2 f = \Sigma t^2 f + 2\Sigma tf + \Sigma f$$
$$2095 = 2357 + 2(-631) + 1000$$
$$2095 = 2095$$

LHS = RHS, so calculations in the columns of the table are correct.

$$\bar{t} = \frac{\Sigma tf}{\Sigma f} = \frac{-631}{1000} = -0.631$$
$$\mu = X_o + c\bar{t} = 36 + 10(-631) = 36 - 6.31 = 29.69.$$

Variance with Reference to "*t*" (V_t)

$$V_t = \frac{1}{N}\left[\sum\nolimits_1^K t^2 f - \frac{(\Sigma tf)^2}{N}\right]$$
$$= \frac{1}{1000}\left[2357 - \frac{(631)^2}{1000}\right]$$
$$= \frac{1}{1000}[2357 - 398.161] = 1.959$$

Hence "standard deviation" in terms of "*t*" (σt):

$$\boldsymbol{\sigma t} = \sqrt{1.959} = 1.4$$

But the "standard deviation" with reference to the variable "*X*" (σx) would be the product of "class interval," "*c*," and σt as computed below:

$$\boldsymbol{\sigma x} = c\sigma t = 10 \times 1.4 = 14$$

Coefficient of Variation (CV)
It is a ratio of standard deviation (σx) to the mean (μ) expressed as percent:

$$\text{CV} = \frac{\text{SD}}{\mu} \times 100 = \frac{14}{29.69} \times 100 = 47.15\%$$

7.10 Merits and Demerits of Various Measures of Dispersion

1. Once the choice of particular average is made, the choice of dispersion is circumscribed. For example, (a) SD with mean ($\bar{X}$) and (b) MD or QD with median ($\tilde{X}$).
2. Range is calculated in two extreme values. QD is calculated on Q_1 and Q_3. Range and QD remain the same, no matter how other values are distributed in the series.
3. MD and SD are obtained by making use of all the observations. Hence they are affected by extreme values. This effect is greater in the case of SD.
4. To determine the quartiles with any degree of accuracy, the sample size should be very large. Otherwise quartiles will be erratic subject to wide fluctuations.
5. SD is useful for both small and large samples.
6. SD possesses excellent sampling properties especially for samples obtained from normal populations. Hence its use is preferable.

7.11 When to Use Various Measures of Dispersion?

Range

1. When the data are much too scattered
2. When the knowledge of extreme scores or of total spread is only required

Quartile Deviation (QD)

1. When $\tilde{X}$ is the measure of central tendency
2. Where the extreme scores influence SD disproportionately
3. When the concentration around the median is of primary importance

Median Deviation (MD)

1. When it is desired to weigh all deviations from the mean or median, according to size
2. When extreme deviations would influence QD unduly

Standard Deviation (SD)

1. When the statistics having the greatest stability is required
2. When extreme deviations should exercise proportionately greater effect than the variability
3. When "coefficient of correlation" (r) and other statistics are to be computed

Correlation 8

8.1 Causation and Correlation

Suppose we find direct correlation between two variables. But it does not mean that the change in variable "*Y*" is a direct cause of a change in variable "*X*." If at all the change in "*Y*" is directly associated with a change in the variable "*X*," then it would be certain that *X* and *Y* are correlated. The existence of correlation may be due to any one of the following:

8.2 One Variable Being a Cause of Another

The cause variable is taken as an independent variable (*X*), and the effect variable is considered as a dependent one (*Y*). Suppose "age" and "height" are correlated. Age is an independent variable which is a cause for change in height, the dependent variable.

8.2.1 Both Variables Being the Result of a Common Cause

Women in a group were followed, after a given operation. The duration of survival and number of children born to a woman were recorded. These factors were related, and it was found that there was a high degree of "positive correlation." It would be interesting to interpret this data in either of the following two ways:

1. Prolonged life of a woman tends to bear more children.
2. Bearing of children tends to prolong the life of a woman.

Note Both these interpretations are absurd. Neither prolonged life has any effect on bearing of children nor bearing of children increases the life span of a woman. One

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_8

can therefore think of some other factors such as age and state of health at the time of operation, which could tend to affect both the survival time and bearing of children.

8.2.2 Chance

Rainfall of some place in north may find high degree of correlation with per acre yield of rice in the south. It would be meaningless to think that the rainfall recorded in the north has any effect on the yield of rice in the south. Such correlations are called spurious or chance correlations. Hence, one must reasonably think of any likelihood relationship existing between the two variables under study.

So, one should be very careful in interpreting the relationship when correlation between the two variables exists.

8.3 Methods of Studying Correlation

1. The scatter diagram
2. Pearson's coefficient of correlation
3. The regression line

8.3.1 The Scatter Diagram

Usually the scale on Y-axis starts from zero though the scale on X-axis need not start from zero. But in cases of "scatter diagram," this restriction on the side of Y-axis is also removed. Both X- and Y-axes may be stared at the minimum values of the respective variables.

8.3.2 Pearson Coefficient of Correlation for Ungrouped Data

Pearson's coefficient of correlation is a measure of the degree of relationship between the two variables. It is denoted by "r" in the case of the sample estimate and by "ρ" in case of the correlation obtained from the whole population. This is also known as the product moment component of correlation. The computation formulae for the both have been illustrated in Table 8.1.

The formulae may also be written in different forms for the sake of convenience in calculations. These are as below:

$$\mathbf{r} = \frac{\sum_1^n \left(Xi - \bar{X}\right) \cdot \left(Yi - \bar{Y}\right)}{\sqrt{\sum_1^n \left(Xi - \bar{X}\right)^2 \cdot \sum_1^n \left(Yi - \bar{Y}\right)^2}}$$

Table 8.1 Pearson's coefficient of correlation formulae

Obtained from the sample	Obtained from the whole population
$\mathbf{r} = \frac{\Sigma\frac{x}{sx}\cdot\Sigma\frac{y}{sy}}{n-1}$	$\boldsymbol{\rho} = \frac{\Sigma\frac{x}{\sigma x}\cdot\Sigma\frac{y}{\sigma y}}{N}$
where	where
$x = X-\bar{X}$	$x = X-\mu X$
$y = Y-\bar{Y}$	$y = Y-\mu Y$
$n =$ no. of pairs of items	$N =$ no. of pairs of items
$sx =$ SD of X-variables	$\sigma x =$ SD of X-variables
$sy =$ SD of Y-variables	$\sigma y =$ SD of Y-variables

$$\mathbf{r} = \frac{\sum_1^n XiYi - n\bar{X}\bar{Y}}{\sqrt{\left[\sum_1^n Xi^2 - n\bar{X}^2\right]\cdot\left[\sum_1^n Yi^2 - n\bar{Y}^2\right].}}$$

$$\mathbf{r} = \frac{\sum_1^n XiYi - \frac{\sum_i^n Xi\cdot\sum_i^n Yi}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_i^n Xi^2}{1} - \frac{\left(\sum_{\mathrm{i}}^n Xi\right)^2}{n}\right]\cdot\left[\frac{\sum_i^n Yi^2}{1} - \frac{\left(\sum_i^n Yi\right)^2}{n}\right]}}$$

$$\mathbf{r} = \frac{\sum_1^n uv - \frac{\sum_i^n u\cdot\sum_i^n v}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_i^n u^2}{1} - \frac{\left(\sum_i^n u\right)^2}{n}\right]\cdot\left[\frac{\sum_i^n v^2}{1} - \frac{\left(\sum_i^n v\right)^2}{n}\right]}}$$

In the above formula, "u" and "v" are the new variables used to simplify the computation: $u = X-X_{\mathrm{o}}$ and $v = Y-Y_{\mathrm{o}}$, where X_{o} and Y_{o} are the assumed means.

Pearson's coefficient of correlation (**r**) can also be computed by the "difference formula" as given below:

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n d^2}{2\sqrt{\sum_1^n x^2\cdot\sum_1^n y^2}}$$

In which $\sum_1^n d^2 = \sum_1^n (x-y)^2$ and $x = X-\bar{X}$; $y = Y-\overline{Y.}$

The above equation can also be modified as below:

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n (x-y)^2 - 2\left(\sum_1^n x\right)\left(\sum_1^n y\right)}{2\sqrt{\left[n\sum_1^n x^2 - \left(\sum_1^n x\right)^2\right]\left[n\sum_1^n y^2 - \left(\sum_1^n y\right)^2\right]}}$$

8.3.2.1 Examples Illustrating the Computations of *r*-Test

Example 1 Computations of "*r*" when deviations are taken from their means. Data of height and weight of five students have been tabulated in Table 8.2.

$$\mathbf{r} = \frac{\sum_1^n \left(Xi - \bar{X}\right) \cdot \left(Yi - \bar{Y}\right)}{\sqrt{\sum_1^n \left(Xi - \bar{X}\right)^2 \cdot \sum_1^n \left(Yi - \bar{Y}\right)^2}}$$
$$= \frac{\sum_1^n xy}{\sqrt{\sum_1^n (x)^2 \cdot \sum_1^n (y)^2}} = \frac{55}{\sqrt{20 \times 750}} = \mathbf{0.449}$$

$\mathrm{df} = n{-}2 = 5{-}2 = 3$
$\mathbf{r}_{0.05} = \mathbf{0.878}$

Decision

The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 2 Computations of "*r*" when deviations are taken from the assumed means. Data of height and weight of five students has been tabulated in Table 8.3.

Table 8.2 Data of height and weight for *r*-test

Student	Height "inches" *X*	Weight "kg" *Y*	$x = X–\bar{X}$	$y = Y–\bar{Y}$	x^2	y^2	*xy*
1	72	70	+3	0	9	0	0
2	69	65	0	−5	0	25	0
3	66	50	−3	−20	9	400	60
4	70	80	+1	+10	1	100	10
5	68	85	−1	+15	1	225	−15
Sum	**345**	**350**			**20**	**750**	**55**
Mean	**69**	**70**					

Table 8.3 Data of height and weight for *r*-test

			$u = X–Xo$	$v = Y–Yo$			
Student	Height "inches" *X*	Weight "kg" *Y*	$X_o = 70$	$Y_o = 70$	u^2	v^2	*uv*
1	72	70	+2	0	4	0	0
2	69	65	−1	−5	1	25	5
3	66	50	−4	−20	16	400	80
4	70	80	0	+10	0	100	0
5	68	85	−2	+15	4	225	−30
Sum			**−5**	**0**	**25**	**750**	**+55**

$$\mathbf{r}=\frac{\sum_{1}^{n}uv-\frac{\sum_{\mathrm{i}}^{n}u\cdot\sum_{\mathrm{i}}^{n}v}{\mathrm{n}}}{\sqrt{\left[\frac{\sum_{\mathrm{i}}^{n}u^{2}}{1}-\frac{\left(\sum_{\mathrm{i}}^{n}u\right)^{2}}{n}\right]\cdot\left[\frac{\sum_{\mathrm{i}}^{n}v^{2}}{1}-\frac{\left(\sum_{\mathrm{i}}^{n}v\right)^{2}}{n}\right]}}$$

$$=\frac{55-\frac{(-5)\cdot(0)}{5}}{\sqrt{\left[25--\frac{(5)^{2}}{5}\right]\cdot\left[750--\frac{(0)^{2}}{5}\right]}}=\frac{55}{\sqrt{20\times750}}=\mathbf{0.449}$$

$\mathrm{df}=\mathrm{n}-2=5-2=3$

$\mathbf{r}_{0.05}=\mathbf{0.878}$

Decision

The computed value of **r** = 0.449 is less than the "table value" of $\mathbf{r}_{0.05}=0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 3 Computations of "*r*" from observed data without taking deviations. Data of height and weight of five students has been tabulated in Table 8.4.

$$\mathbf{r}=\frac{\sum_{1}^{n}XiYi-\frac{\sum_{i}^{n}Xi\cdot\sum_{i}^{n}Yi}{n}}{\sqrt{\left[\frac{\sum_{i}^{n}Xi^{2}}{1}-\frac{\left(\sum_{i}^{n}Xi\right)^{2}}{n}\right]\cdot\left[\frac{\sum_{i}^{n}Yi^{2}}{1}-\frac{\left(\sum_{i}^{n}Yi\right)^{2}}{n}\right]}}$$

$$=\frac{24205-\frac{345\times350}{5}}{\sqrt{\left[23825-\frac{(345)^{2}}{5}\right]\cdot\left[25250-\frac{(350)^{2}}{5}\right]}}$$

$$=\frac{24205-24150}{\sqrt{[23825-23805]\cdot[25250-24500]}}=\frac{55}{\sqrt{20\times750}}=\frac{55}{122.47}=\mathbf{0.449}$$

$\mathrm{df}=n-2=5-2=3$

$\mathbf{r}_{0.05}=\mathbf{0.878}$

Decision

The computed value of **r** = 0.449 is less than the "table value" of $\mathbf{r}_{0.05}=0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Table 8.4 Data of height and weight for *r*-test

Student	Height "inches" X	Weight "kg" Y	X^2	Y^2	XY
1	72	70	5184	4900	5040
2	69	65	4761	4225	4485
3	66	50	4356	2500	3300
4	70	80	4900	6400	5600
5	68	85	4624	7225	5780
Sum	**345**	**350**	**23,825**	**25,250**	**24,205**

Table 8.5 Data of height and weight for *r*-test

			$x = X–\bar{X}$	$y = Y–\bar{Y}$			$(x–y)$
Student	Height "inches" X	Weight "kg" Y	$(\bar{X} = 69)$	$(\bar{Y} = 70)$	x^2	y^2	2
1	72	70	+3	0	9	0	9
2	69	65	0	−5	0	25	25
3	66	50	−3	−20	9	400	289
4	70	80	+1	+10	1	100	81
5	68	85	−1	+15	1	225	256
Sum	**345**	**350**			**20**	**750**	**660**
Mean	**69**	**70**					

Example 4 Computations of "*r*" by the "difference formula." Data of height and weight of five students has been tabulated in Table 8.5.

$$\mathbf{r} = \frac{\sum_1^n x^2 + \sum_1^n y^2 - \sum_1^n d^2}{2\sqrt{\sum_1^n x^2 \cdot \sum_1^n y^2}} = \frac{20 + 750 - 660}{2\sqrt{20 \times 750}}$$

$$= \frac{110}{2\sqrt{20 \times 750}} = \frac{55}{\sqrt{20 \times 750}} = 0.449$$

$\mathrm{df} = n–2 = 5–2 = 3$
$\mathbf{r}_{0.05} = 0.878$

Decision

The computed value of $\mathbf{r} = 0.449$ is less than the "table value" of $\mathbf{r}_{0.05} = 0.878$. So, "null hypothesis" (H_o) is accepted. Hence, there is no significant correlation between the height and weight of students.

Example 5 The body weights of five chicks were 180, 170,170, 190, and 190 grams, respectively, and their comb weights were found to be 50, 40, 20, 60, and 60 grams, respectively. Find out if there is any correlation between the body weight and comb weight of chicks.

Table 8.6 Body weights and comb weights of chicks

Chick no	Body weight (x)	Comb weight (y)	$u{-}170$	$v = y{-}40$
1	180	50	10	10
2	170	40	0	0
3	170	20	0	−20
4	190	60	20	20
5	190	60	20	20
Total			**50**	**30**

Solution

Data has been transformed by subtracting 170 from the body weights and 40 from the comb weights as shown in Table 8.6.

$$\begin{aligned}\Sigma u^2 &= 100 + 400 + 400 = 900\\ \Sigma v^2 &= 100 + 400 + 400 + 400 = 1300\\ \Sigma uv &= 100 + 400 + 400 = 900\end{aligned}$$

$$\begin{aligned}\mathbf{r} &= \frac{\Sigma uv - \dfrac{\Sigma u \cdot \Sigma v}{n}}{\sqrt{\left(\Sigma u^2 - \dfrac{(\Sigma u)^2}{n}\right)\left(\Sigma v^2 - \dfrac{(\Sigma v)^2}{n}\right)}}\\ &= \frac{900 - \dfrac{50 \times 30}{5}}{\sqrt{\left(900 - \dfrac{50 \times 50}{5}\right)\left(1300 - \dfrac{30 \times 30}{5}\right)}}\\ &= \frac{900 - 300}{\sqrt{(900 - 500)(1300 - 180)}}\\ &= \frac{600}{\sqrt{400 \times 1120}} = \frac{600}{20\sqrt{1120}} = \frac{30}{\sqrt{1120}} = \frac{30}{33.5} = 0.895 = +\mathbf{0.895}\end{aligned}$$

$\mathbf{df} = \mathrm{n} - 2 = 5 - 2 = 3$; $\mathbf{r_{0.05}} = 0.878$

Decision

The calculated $\mathbf{r} = +0.895$ is greater than $\mathbf{r_{0.05}} = 0.878$. So, "null hypothesis" (H_o) is rejected ($p < 0.05$). Hence, there is direct correlation between the "body weights" and "comb weights" of chicks.

8.3.3 Regression Line

To determine the amount of change that normally takes place in the Y-variable for a unit change in the X-variable, a line is fitted to the points plotted on the scatter diagram. This line is described as $Y = a + bx$ and is said to be line of regression of Y on X. Here, "a" and "b" are the two constants: $a = Y$-intercept and $b =$ slope of the regression line. The "b" may also be written as "by" = regression coefficient of Y on X. It is also possible to find $bXY =$ "regression coefficient" of X on Y. There is a definite relationship between "**r**" and these two "regression coefficients" bYX and bXY. The "**r**" is the geometric mean of bXY and bYX.

Therefore:

$$\mathbf{r} = \sqrt{bYX \cdot bXY}; \text{But } bYX = \frac{\Sigma\left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\left(X - \bar{X}\right)^2}$$

whereas:

$$\mathbf{r} = \frac{\sum_1^n \left(Xi - \bar{X}\right) \cdot \left(Yi - \bar{Y}\right)}{\sqrt{\sum_1^n \left(Xi - \bar{X}\right)^2 \cdot \sum_1^n \left(Yi - \bar{Y}\right)^2}}$$

Hence it can be proved that $bYX = r\sqrt{\frac{\left(Y - \bar{Y}\right)^2}{\left(X - \bar{X}\right)^2}} = \frac{sy}{sx}$

8.4 Proportions of "*r*"

1. The "r" values range from -1.00 through 0.00 to +1.00.
2. It is a pure number, independent of the units of measurement of the variables X and Y.
3. If $r = -1$, a perfect inverse linear relationship exists between the variables (e.g., volume $\propto \frac{1}{\text{Power}}$).
4. If $r = -0$, linear relationship between the two variables X and Y does not exist (e.g., number of births registered vs number of cars registered).
5. If $r = +1$, there is a perfect direct linear relationship (e.g. diameter vs circumference).
6. If $r = -0.7$ or + 0.7 in a large set, the degree of relationship between the two variables seems to be high.
7. If $r = +0.6$, it does not mean that 60% of the values are related.
8. The computation of r is valid only if the variables are approximately normally distributed.
9. The r^2 is known as coefficient of determination. If $r^2 = 0.756$, it means that approximately 75.6% of the variation in Y is only due to the linear regression of Y on X.

Chi-Square Test (χ^2 – Test) 9

9.1 Degrees of Freedom

The term "degrees of freedom" refers to the "independent constrains" in a set of data. Let us consider a 2 × 2 contingency table: Table 9.1.

If A and B are independent, the probability of the occurrence of A and B is given by $P(AB) = P(A) \cdot P(B) = \frac{a}{n} \cdot \frac{c}{n}$.

Hence under the hypothesis of independence, the expected frequency of the cell AB is given by $n \cdot P(AB) = n \cdot \left(\frac{a}{n} \cdot \frac{c}{n}\right) = \frac{ac}{n}$. Once the cell frequency of a cell is determined, the frequencies of the other cells could be automatically filled, keeping the marginal frequencies unaltered. It is evident from this that only one frequency can independently be determined. Hence for 2 × 2 contingency table, the "degrees of freedom" (df) is only one. However, it can be determined by the following formula:

$$\mathrm{df} = (c-1)(r-1)$$

Where

c = number of columns in a contingency table.
r = number of rows in a contingency table.

If the data is not in the form of contingency table but in the shape of a series, then the degrees of freedom is determined as given in Table 9.2.

There are seven classes of frequencies in the data given above. We can independently determine the expected frequencies for six classes, keeping the total unaltered. So, there are six degrees of freedom (df). This makes it clear that if there are "n" classes of frequencies, "$n-1$" class frequencies can be determined and that means "$n-1$" degrees of freedom.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_9

Table 9.1 Understanding the data distribution for χ^2 – test

Variables	$\boldsymbol{A}$	$\boldsymbol{\bar{A}}$	**End total**
$\boldsymbol{B}$	ab	$\bar{a}b$	$\boldsymbol{c}$
$\boldsymbol{\bar{B}}$	$a\bar{b}$	$\bar{a}\bar{b}$	$\boldsymbol{d}$
End total	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{n}$

Table 9.2 Data arranged as series

No. of heads	Frequency
0	2
1	10
2	38
3	105
4	188
5	257
6	226
Total	827

9.2 Levels of Significance

The divergence of theory and fact is always tested in terms of probabilities. The probabilities indicate the extent of confidence laid on the conclusions drawn. Table values of χ^2 are available at various probability levels. These levels of significance are considered at 5% ($p = 0.05$), 1% ($p = 0.01$), and 0.1% ($p = 0.001$). If the observed value of χ^2 is less than given at $p = 0.05$ in χ^2 table, then we accept that the difference must be due to chance fluctuations. If the χ^2 value is significant at 5% level of significance, it is said to be probably significant ($p < 0.05$). If the χ^2 is greater than 1% level of significance, we declare the result to be significant ($p < 0.01$). If the χ^2 value is greater than even at 0.1% of significance, we declare that the result is highly significant ($p < 0.001$).

9.3 Applications of χ^2 – Test

9.3.1 Double Variable Data

Example 1 In a study it was observed that out of 50 non-vaccinated children 20 got measles and out of 50 vaccinated children 10 got measles. Test significance of effectiveness of vaccination for immunity against measles. Distribution of data has been exhibited in the Table 9.3.

Table 9.3 Distribution of data for application of χ^2 – test

Vaccination status	No measles	Got measles	End total
Not vaccinated	30	20	**50**
	a	*b*	$\boldsymbol{m_3}$
Vaccinated	40	10	**50**
	d	*c*	$\boldsymbol{m_4}$
End total	**70**	**30**	**100**
	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	***N***

Solution

$$\chi^2 = \frac{N(ac - bd)^2}{m_1 \times m_2 \times m_3 \times m_4}$$

$$\chi^2 = \frac{100(30 \times 10 - 20 \times 40)^2}{70 \times 30 \times 50 \times 50}$$

$$\chi^2 = \frac{100(30 \times 10 - 20 \times 40)^2}{70 \times 30 \times 50 \times 50} = \frac{100}{21} = 4.8$$

$$\text{df} = (2-1)(2-1) = 1$$

Value of χ^2 corresponding to df = 1 for $p = 0.05$ is 3.841.

Conclusion $p < 0.05$, significant at 5% level. The null hypothesis (H_o) stands rejected. So, vaccination has potential to protect us from disease.

9.3.2 Goodness of Fit Test

Goodness of fit is contested to test the null hypothesis that there is no significant difference between the distribution of the observed frequencies and expected frequencies.

Example 2 Total acid content of stomach was measured in a group of patients following stimulating dose of histamine as tabulated below in the Table 9.4. Apply the goodness of fit test to this data to test the validity of null hypothesis.

$$\chi^2 = \sum\nolimits_{i=1}^{k} \frac{(O-E)^2}{E} = \sum \frac{(O-E)^2}{E} = 2.4413$$

df = (No. of classes of frequencies − 1) = 5–1 = 4
From the table we find that

χ^2 critical value at p 0.05 is 9.488, and at p 0.01 it is 13.277 corresponding to df = 4. The computed value in this case is 2.4413, which is less than 9.488. So, $p > 0.05$.

Conclusion As p-value is nonsignificant, hence null hypothesis is accepted. That means distribution of observed frequencies is consistent with the distribution of expected frequencies. There is no evidence of divergence, and differences are only due to chance fluctuations.

Example 3 A study was conducted on 1500 persons. It was observed that 400 of them were pipe smokers (PS) and the rest were non-smokers. Due to a certain disease, 50 pipe smokers and 150 non-smokers died. Test a null hypothesis (H_o) when it is expected that $^1/_3$ of the persons under study were pipe smokers and the ratio born by persons died (D) to alive (A) was 1:7 (D/A::1:7).

Solution

Observed and expected values have been displayed in Table 9.5.

Given: (i) One third (⅓) of persons under study were pipe smokers and (ii) the ratio born by persons died (D) to alive (A) was 1/7 (D/A::1:7).

So, pipe smokers = 500 and non-smokers = 1000

Now work out the "expected" values of the persons "died" and "alive" in both the groups as per given ratio, and enter in table as "expected" values. Then calculate $\frac{(O-E)^2}{E}$ for all groups.

Table 9.4 Distribution of patients with gastric ulcer, diagnosed according to total acid content of stomach after stimulating dose of histamine

Acid level units	Observed frequency (O)	Expected frequency (E)	($O-E$)	$\frac{(O-E)^2}{E}$
50–59	6	6	0	0.0000
60–69	12	9	+3	1.0000
70–79	17	18	−1	0.0556
80–89	11	10	+1	0.1000
90–99	4	7	−3	1.2857
Total	**50**	**50**	**0**	**2.4413**

Table 9.5 Distribution of data for χ^2 for goodness of fit test

Smoking status	**Died**	**Alive**	**End total**
Pipe smokers	50	350	**400**
	$E = \frac{1\times500}{8}$	$E = \frac{7\times500}{8}$	$\boldsymbol{m_3}$
Non-smokers	150	950	**1100**
	$E = \frac{1\times1000}{8}$	$E = \frac{7\times1000}{8}$	$\boldsymbol{m_4}$
End total	**200**	**1300**	**1500**
	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{N}$

Pipe Smokers

Expected values of 'Died' $= \frac{1\times 500}{8}$,

$$\frac{(\mathbf{O}-\mathbf{E})^{\mathbf{2}}}{\mathbf{E}} = \frac{\left(50-\frac{500}{8}\right)^2}{\frac{500}{8}} = \frac{8\left(\frac{400-500}{8}\right)^2}{500} = \frac{5}{2} = 2.5$$

Expected values of 'Alive' $= \frac{7\times 500}{8}$

$$\frac{(\mathbf{O}-\mathbf{E})^{\mathbf{2}}}{\mathbf{E}} = \frac{\left(350-\frac{7\times 500}{8}\right)^2}{\frac{500}{8}} = \frac{8\left(\frac{2800-3500}{8}\right)^2}{500} = \frac{35}{2} = 17.5$$

Non-smokers

Expected values of 'Died' $= \frac{1\times 1000}{8}$

$$\frac{(\mathbf{O}-\mathbf{E})^{2}}{\mathbf{E}} = \frac{\left(150-\frac{1000}{8}\right)^2}{\frac{1000}{8}} = \frac{8\left(\frac{1200-1000}{8}\right)^2}{1000} = 5$$

Expected values of 'Alive' $= \frac{7\times 1000}{8}$

$$\frac{(\mathbf{O}-\mathbf{E})^{2}}{\mathbf{E}} = \frac{\left(950-\frac{7000}{8}\right)^2}{\frac{1000}{8}} = \frac{8\left(\frac{7600-7000}{8}\right)^2}{1000} = \frac{45}{7} = 6.4$$

$$\chi^2 = \sum\nolimits_{i=1}^{k} \frac{(O-E)^2}{E} = 2.5 + 17.5 + 5.0 + 6.4 = 31.4$$
$$\chi^2 = 31.4$$
$$\mathrm{df} = 4 - 3 = 1$$
$$p < 0.01$$

As $p < 0.01$ signifies that observed values differ from expected values (H_o is rejected).

Conclusion Observed values do not fit well with expected values.

Example 4 Out of 680 patients examined at a Cancer Hospital had malignant neoplasia of cervix uterus ($n = 270$), oral cavity ($n = 120$), breast ($n = 100$), larynx ($n = 100$), and lungs ($n = 90$), respectively. Apply chi-square test to find out the fitness of observed values with expected values if expected value of malignant neoplasia could be 40%, 20%, 15%, 15%, and 10%, respectively, of these organs.

Solution

The data has been arranged in the Table 9.6, and chi square for goodness of fit test has been applied.

Table 9.6 Cases of malignant neoplasia

Organ involved	Observed no. (O)	Expected no. (E)	$\frac{(O-E)^2}{E}$
Cervix uterus	270	272	0.014
Oral cavity	120	136	1.900
Breast	100	102	0.400
Larynx	100	102	0.400
Lungs	90	68	7.118
Total	680	680	9.102

Table 9.7 Hemorrhage in 224 cases due to premature separation of placenta in various stages of gestation

	Stages of gestation			
Total blood loss (ml)	Immature	Premature	Full term	Total
<500	23 ($E = 17.8$)	47 ($E = 53.5$)	51 ($E = 49.7$)	121
500–1000	4 ($E = 7.7$)	29 ($E = 23.0$)	19 ($E = 21.3$)	52
>1000	6 ($E = 7.5$)	23 ($E = 22.5$)	22 ($E = 21.0$)	51
Total	**33**	**99**	**92**	**224**

$$\chi^2 = \sum \frac{(O - E)^2}{E} = 9.10$$
$$\text{df} = 5 - 1 = 4$$

Table value of $\chi^2 = 9.49$ (at 5% level of significance).

Decision

The computed value of $\chi^2 = 9.10$ is less than "table value" of $\chi^2 = 9.49$ (at 5% level of significance: $p > 0.05$). Hence, null hypothesis (H_o) is accepted. The observed values of malignant neoplasia fit well with expected values.

9.4 Coefficient of Contingency

Example 1 A study was conducted on blood loss due to postpartum hemorrhage with reference to gestational period. As per null hypothesis, if there is no association between blood loss and gestational period, then the blood loss should be equal in all the gestational groups. Expected frequencies are calculated on this assumption. Data regarding hemorrhage in premature separation of placenta in terms of volume of blood loss with reference to gestational period has been exhibited in Table 9.7.

Solution

Expected frequencies with reference to Table 9.7 were worked out as illustrated below:

	Stages of gestation			
Total blood loss (ml)	immature	Premature	Full term	Total
<500	$E = \frac{33\times121}{224} = 17.8$	$E = \frac{99\times121}{224} = 53.5$	$E = \frac{92\times121}{224} = 49.7$	121
500–1000	$E = \frac{33\times52}{224} = 7.7$	$E = \frac{99\times52}{224} = 23.0$	$E = \frac{92\times52}{224} = 21.3$	52
>1000	$E = \frac{33\times51}{224} = 7.5$	$E = \frac{99\times51}{224} = 22.5$	$E = \frac{92\times51}{224} = 21.0$	51
Total	**33**	**99**	**92**	**224**

Applying the "observed" and "expected" frequencies, the value of coefficient of contingency (χ^2) is computed as given below:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} = 1.5 + 0.79 + 0.03 + 1.78$$
$$+1.57 + 0.25 + 0.3 + 0.01 + 0.04 = 6.27\boldsymbol{\chi^2} = \mathbf{6.27}$$

The χ^2 value for the example solved is 6.27. Degrees of freedom are computed as $(c-1)(r-1) = (3–1)(3–1) = 2 \times 2 = 4$. P value comes out to be $p > 0.05$. That is not significant at 5% level of significance. So, null hypothesis (H_o) holds good. Hence there is no association between blood loss and period of gestation.

Example 2 Suppose it is believed that the survival is closely associated with "treatment A" rather than with "treatment B." To prove or reject this notion, a null hypothesis (H_o) is framed.

H_o = survival or death has no association with type of treatment.

Statement

Out of 257 patients getting "treatment A," 41 died and 216 survived of a particular disease. Whereas out of 244 patients being treated with treatment "B" for the same disease, 64 died and 180 survived. Analyze the data for null hypothesis (H_o).

Solution

Observed and expected values have been displayed in Table 9.8.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} = 3.09 + 0.82 + 3.25 + 0.86 = 8.02$$
$$\chi^2 = 8.02$$

Degrees of freedom = $(c-1)(r-1) = 1 \times 1 = 1$.

Now, table values of χ^2 at 5%, 1%, and 0.1% levels of significance are 3.841, 6.635, and 10.827, respectively. Obtained χ^2 value 8.02 falls between 1% and 0.1% levels of significance. So, $p < 0.01$. Therefore, the result is significant at 1% level of significance, and null hypothesis (H_o) is rejected. Hence, "treatment A" is more effective in the survival of patients than the "treatment B."

Example 3 A survey study was done in a slum area to assess the cleanliness of children with reference to condition of house. Children were graded as clean, partially clean, and dirty, and houses were graded as clean or not clean. Analyze the data inhibited in Table 9.9 to suggest that the condition of house has a bearing on the condition of the children.

Table 9.8 Association of two different treatments with outcome

Treatment	**Outcome**		**End total**
	Died	**Survived**	
A	**41**	**216**	**257**
	$E = \frac{105\times257}{501} = 53.9$	$E = \frac{396\times257}{501} = 203.1$	m_3
B	**64**	**180**	**244**
	$E = \frac{105\times244}{501} = 51.1$	$E = \frac{396\times244}{501} = 192.9$	m_4
End total	**105**	**396**	**501**
	m_1	m_2	N

Table 9.9 Survey data of children and houses in a slum area

Condition of children	**Condition of house**		**End total**
	Clean	**Not clean**	
Clean	75	40	**115**
	$E = \frac{140\times115}{245} = 65.7$	$E = \frac{105\times115}{245} = 49.3$	m_3
Partially clean	40	15	**55**
	$E = \frac{140\times55}{245} = 31.4$	$E = \frac{105\times555}{245} = 23.6$	m_4
Dirty	25	50	**75**
	$E = \frac{140\times75}{245} = 42.9$	$E = \frac{105\times75}{245} = 32.1$	m_5
End total	**140**	**105**	**245**
	m_1	m_2	N

Solution
Extension of Table 9.9 for χ^2 Computation

Observed frequency (O)	Expected frequency (E)	Difference ($O-E$)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
75	65.7	9.3	86.49	1.316
40	31.4	8.6	73.96	2.355
25	42.9	−17.9	320.41	7.469
40	49.3	−9.3	86.49	1.754
15	23.6	−8.6	73.96	3.134
50	32.1	17.9	320.41	9.982
Sum				**26.01**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E} = 26.01$$
$$V = \mathrm{df} = (c-1)(r-1) = (2-1)(3-1) = 1 \times 2 = 2;$$
$$\chi^2 \mathbf{0.05} = \mathbf{5.99},$$

Conclusion Calculated value of χ^2 is very much greater than table value. So, null hypothesis is rejected. Hence, conditions at home has a bearing on the cleanliness of children.

9.5 Yates' Correction

The χ^2 test is applied to discrete variables to ascertain the association between these. Yates, F. in 1934 has given a method to be applied when expected frequency/value in any cell of "two-by-two" (2 × 2) contingency table is less than 5. Correction is done in the formula for total number of observations. Half of the number of observations ($\frac{N}{2}$) is subtracted from the absolute value of "*ac−bd*" before taking square and multiplying with N as illustrated here under with reference to data for vaccinated and non-vaccinated people arranged in Table 9.6 as Example 1 for Yates' correction.

Formula without Yates' correction:

$$\chi^2 = \frac{N(ac-bd)^2}{m_1 \cdot m_2 \cdot m_3 \cdot m_4}$$

Formula with Yates' correction:

$$\chi^2 = \frac{N\left[|ac-bd| - \frac{N}{2}\right]^2}{m_1 \cdot m_2 \cdot m_3 \cdot m_4}$$

Table 9.10 Distribution of data of 24 patients with tuberculosis

Vaccination status	Died of TB	Survived	End total
Vaccinated (BCG)	2	10	**12**
	$(E = 2.5)\ a$	$(E = 9.5)\ b$	$\boldsymbol{m_3}$
Not vaccinated	8	4	**12**
	$(E = 7.5)\ d$	$(E = 4.5)\ c$	$\boldsymbol{m_4}$
End total	**10**	**14**	**24**
	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{N}$

Example 1 In a study it was observed that out of 24 cases suffering from tuberculosis (TB), half were vaccinated, and half were not vaccinated with BCG. Out of 12 non-vaccinated cases, 8 died, and 4 survived after treatment, whereas in vaccinated group of patients, 2 died, and 10 survived. Analyze statistically for role of BCG vaccination in protection from tuberculosis. The data has been arranged in Table 9.10.

$$\chi^2 = \frac{N\,(ac - bd)^2}{m_1 \cdot m_2 \cdot m_3 \cdot m_4} = \frac{24\,(8 - 80)^2}{10 \times 14 \times 12 \times 12} = \frac{24(72)^2}{20,160} = \frac{124,416}{20,160} = 6.171$$

$$\chi^2 = \frac{N\left[|ac - bd| - \frac{N}{2}\right]^2}{m_1 \cdot m_2 \cdot m_3 \cdot m_4} = \frac{24\left[|8 - 80|\frac{24}{2}\right]^2}{10 \times 14 \times 12 \times 12} = \frac{24[60]^2}{20,160} = \frac{86,400}{20,160} = 4.286$$

Now, without Yates' correction, $\chi^2 = 6.1714$, and with Yates' correction, $\chi^2 = 4.2857$.

χ^2 critical value at 5% level of significance for 1 df is 3.841 and at 1% level of significance, it is 6.635.

Conclusion χ^2 values with and without Yates' correction in the study cited above are greater than $\chi^2 = 3.841$ (at 5% level). So, $p < 0.05$. Hence, BCG vaccination plays a protective role against tuberculosis.

9.6 Application of Yates' Correction to Multivariable Data

Formula without Yates' correction:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

Formula with Yates' correction:

$$\chi^2 = \sum_{i=1}^{k} \frac{(|O - E| - 0.5)^2}{E}$$

9.7 Inference

1. If both χ^2 and χ^2 (corrected) are greater than χ^2 (critical value), then a significant difference is established. Null hypothesis (H_o) stands rejected.
2. If both χ^2 and χ^2 (corrected) are less than χ^2 (critical value), then no significant difference is established. Null hypothesis ($\mathrm{H_o}$) is accepted.
3. If $\chi^2 > \chi^2$ (critical value) but χ^2 (corrected) $< \chi^2$ (critical value), then more observations are required to determine the result.

Normal Curve and Sampling Distribution 10

10.1 Normal Distribution

Normal distribution was first described by Abraham De Moivre and then developed by Laplace and Gauss. It is the most important of the theoretical distributions. It is a continuous probability distribution in which the random variable "*X*" can assume either a finite set of value or numerous infinite set of values. Distribution plays a vital role where inferences are made regarding the value of population mean (μ). A continuous variable "*X*" is said to be normally distributed if it has the probability density function represented by the equation:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\pi = 3.1416$ and $e = 2.7183$.

We know that $\frac{X-\mu}{\sigma} = Z$ (standard normal variate):

$$\text{So, } P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z)^2}{2\sigma^2}}$$

10.2 Properties of Normal Curve

1. Normal distribution is a continuous probability distribution having two parameters: mean ($\bar{X}$) and standard deviation (σ).
2. Normal curve is a bell-shaped curve.
3. It is a symmetrical curve.
4. In this mean, median, and mode coincide, i.e., mean ($\bar{X}$) = median ($\tilde{X}$) = mode ($\widehat{X}$).
5. Median is at equal distance from Q_1 and Q_3, i.e., $\left(Q_3 - \tilde{X}\right) = \left(\tilde{X} - Q_1\right)$.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_10

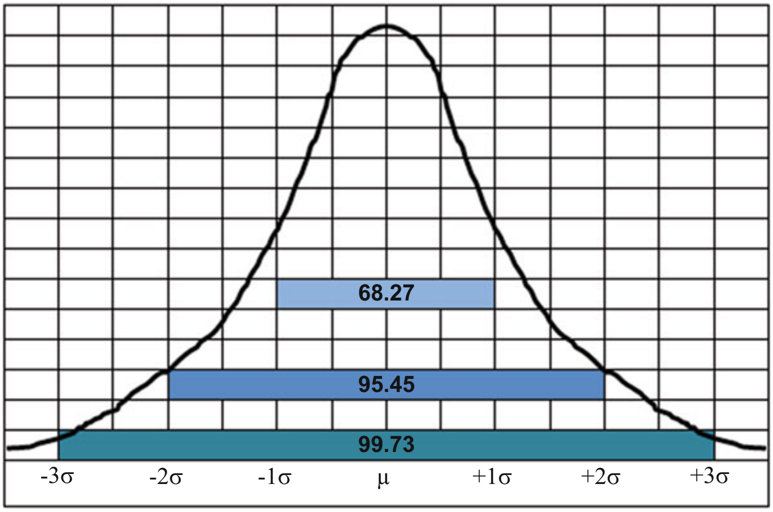


Fig. 10.1 Normal curve

6. It is unimodal class.
7. Its points of inflection are always at "one standard deviation" from the mean ($\bar{X}$).
8. The curve is symmetrical and does not touch the baseline.
9. The normal curve has maximum height at mean value.
10. The ordinate divides the curve into two equal parts.
11. The curve has "permanent areas relationships" as exhibited in Fig. 10.1.

10.3 Applications of Normal Curve

If X is normally distributed with mean μ and variance σ^2, the "standard normal variate" "Z" is normally distributed with mean "0" and variance 1. Area contained under the normal curve at any point "Z" could be noted down from the table for "Z values." Making use of this table, we can determine the area contained between any two ordinates in the normal curve.

10.4 Sampling Variations

Assume that an unbiased random sample is drawn from a population. The sample mean obtained may not exactly tally with the population mean. The error between the sample mean and the universe's mean is subjected to only chance fluctuations. This error is called "sampling error."

If repeated samples of same size are drawn randomly from the same universe, not only the samples' means differ with the population mean, but also the sample means differ among each other. But there is a scope to prove that the distribution of these samples' means follows a normal distribution with reference to population mean (μ), and "mean variance" $\sigma^{2}{}_{/n}$ (population variance divided by sample size), if the size of sample "n" is large ($n > 30$). If "n" is small, the sampling distribution of $\bar{X}$ follows the pattern of Student's "t"-distribution. This measure of variation (${}^{2\sigma}{}_{/\sqrt{n}}$) differs from individual σ and is termed as "standard error (SE) of mean." SE is applied to calculate "t" value. For sampling distribution of means of large samples, the properties of normal distribution hold good.

Because:

$\mu \pm {}^{\sigma}{}_{/\sqrt{n}}$ contains 68% of the means.
$\mu \pm {}^{2\sigma}{}_{/\sqrt{n}}$ contains 95% of the means.
$\mu \pm {}^{3\sigma}{}_{/\sqrt{n}}$ contains 99.7% of the means.

10.5 Effect of Sample Size on Standard Error

It has been observed that with the increase in sample size, the reliability of the sample mean ($\bar{X}$) increases. In other words, the population mean (μ) can be estimated with greater confidence or lesser scope of error as the size of sample increases.

In a population study, standard deviation (σ) was 3.00 and mean ((μ) was 27.00. This gave out "variance" (σ^2) as 9.00. When six samples of sample size (n) of 4, 9, 16, 36, 100, and 400 were studied from the same population, the standard error (${}^{\sigma}{}_{/\sqrt{n}}$) came out to be 1.50, 1.00, 0.75, 0.50, 0.30, and 0.15, respectively, thus reducing the sample observations' limits to be included in 68.27% distribution. The same has been illustrated in Table 10.1.

10.6 Assumptions

We always assume that:

1. Sample is random.
2. Sample has been drawn from the normal population.
3. If the population is normal, then the sampling distribution of means is also normal.
4. Even if the population is not much deviating from normality, the sampling distribution approaches normality with increasing size of sample.

If the population is markedly deviated from normality, then the sampling distribution will not follow the normal distribution. So, investigators should have knowledge of the structure of the population from which the sample is drawn. However, the assumption of normality could be often made without serious error in the absence of knowledge of the population.

Table 10.1 Effect of sample size on standard deviation on sampling distributions of means ($\mu = 27.00$ and $\sigma = 3.00$)

Sample size "*n*"	Standard error ($^{\sigma}/_{\sqrt{n}}$)	Limits to be included in 68.27 distribution: $\mu \pm (^{\sigma}/_{\sqrt{n}})$
4	1.50	25.50–28.50
9	1.00	26.00–28.00
16	0.75	26.25–27.75
36	0.50	26.50–27.50
100	0.30	26.70–27.30
400	0.15	26.85–27.15

In actual experience, the population is not exactly known. So, the mean and standard deviation of population cannot be known. Hence, the estimates obtained from a single sample are used to determine the mean and standard error of the sampling distribution of mean. The estimates from the sample, mean ($\bar{X}$) and standard deviation (s), are unbiased estimates of the population parameters μ and σ. Therefore, the "standard error of the mean" for a sample of size "*n*" is determined by $s/_{\sqrt{n}}$.

Example 1 Suppose a random sample of size 100 is selected and excretion of urea in urine is detected in every individual. The mean urea excretion is found to be 8.000 g with a standard deviation of 1.600 g. When we workout standard error from these estimates, that would be $s/_{\sqrt{n}} = \frac{1.6}{\sqrt{100}} = 0.16$. The confidence limits for the population mean can be determined as given below:

$$95\%\ \text{Confidence limits} : \bar{X} \pm 2\ \text{SE} = 8.0 + 2 \times 0.16 = 8.0 \pm 0.32$$

This means the range 7.68–8.32 g would contain the population mean with 95% confidence. In other words, if 100 repeated samples are drawn of the same population, 95% individuals would have the value of excreted urea between 7.68 and 8.32 g.

Tests of Significance 11

11.1 Statistical Decision

When we draw an inference after applying statistical tests to observations of single sample from population, such a decision is termed as "statistical decision."

11.2 Statistical Hypothesis (Null Hypothesis)

To take a statistical decision, we must make certain assumptions about the population involved. Such assumptions may prove to be true or false. These assumptions are called hypotheses.

We usually frame the null hypothesis with an intension of rejecting it. Suppose there are two treatments "A" and "B." To prove one treatment is more effective than the other, we frame the null hypothesis (H_o) that there is no difference in the effects of two treatments. Under H_o the observed differences may only be due to sample fluctuations. Any hypothesis which differs from H_o is called an alternative or unconventional hypothesis.

11.3 Test of Hypothesis and Significance

After framing a null hypothesis, we would like to try to determine the difference in the random sample, from the population value under the null hypothesis. If this difference is markedly high on the basis of chance fluctuations with respect to theory of sampling distributions, we reject the assumed hypothesis and declare that the difference is significant.

The procedures laid to decide whether to accept or reject the hypothesis or to determine whether observed sample differs significantly from expected results are called the tests of hypothesis, tests of significance, or rules of statistical decision.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_11

11.4 Type I and Type II Errors

If we deliberately reject H_o when the other H_o is true, we commit an error. This error will be called "Type I" error. On the other hand, if we accept H_o when it must be rejected, again we commit an error. This type of error is called "Type II" error.

We should minimize the errors for a good statistical decision. It is a difficult task to minimize errors as sometimes minimizing one type of error would increase the other type of error. Practically one type of error may be more serious than the other type. Hence under compromise one can try to limit the more serious error for a good statistical inference. There should be deliberate effort to increase the sample size to limit the both types of errors.

11.5 Level of Significance

In testing a hypothesis, the maximum probability with which we would like to restrict "Type I" error is called level of significance of the test. This probability is often denoted by $\propto$ and is generally specified before a sample is drawn.

In practice the level of significance chosen is 0.05 (5% level of significance). In designing a test of significance, there would be only 5 chances in 100 for accepting the H_o. This denotes that we are 95% confident of rejecting the null hypothesis.

11.6 Tests Involving Normal Distribution

If observations (X) are normally distributed with mean $\boldsymbol{\mu}$ and variance (σ^2), then the "standard normal variate" $Z = \frac{x-\mu}{\sigma}$ is also normally distributed with mean "0" (of variate) and variance "1."

Under the null hypothesis (H_o), the "Z"-scores of the sample statistic will be between -1.96 and 1.96 in 95 samples out of 100 samples. Hence, we are 95% confident that the "Z"-scores will lie in this region if the hypothesis is true.

However, if a single sample is drawn at random and has a "Z"-score lying outside the region of -1.96 to 1.96, we would conclude that such an event can happen only in 5 cases out of 100 cases, if the given hypothesis is true. We would then say Z-score differed significantly from expected under the hypothesis. Hence the null hypothesis is rejected. The graphic illustration has been shown in Fig. 11.1 exhibiting shaded areas beyond -1.96 to 1.96.

The shaded area (5%) is the level of significance of the test. It represents the probability of our being wrong in rejecting the H_o. If the Z-scores fall in the shaded area, the critical region, we reject the null hypothesis at 5% level of significance. It would mean that Z-scores of a given sample statistics are significant at 0.05 level of significance ($p < 0.05$).

11.7 Rules of Statistical Decision

1. Reject the null hypothesis at 5% level of significance ($p > 0.05$) if Z-score of the sample statistic lies outside the range –1.96 to 1.96.
2. Accept the null hypothesis otherwise.
3. Other levels of significance could be 1% ($p < 0.01$) or 0.1% ($p < 0.001$).

11.8 One-Tailed or Two-Tailed Tests

Figure 11.1 represents critical region of rejection of the null hypothesis on both the extreme tails. Due to this type of presentation, such tests are called *two-tailed tests*. Sometimes we may be interested in the extreme values on one side of the mean (i.e., in one tail of distribution). For example, if we are interested in testing the hypothesis that one process is better than the other "rather than testing one process is better or worse than the other," we apply *one-tailed tests*. In such cases, critical region of rejection of the null hypothesis (H_o) is only on one side of the distribution.

Critical values of "standard normal variate" "Z" both for "one-tailed" and "two-tailed" tests at 5% ($p < 0.05$), 1% ($P < 0.01$), and 0.01% ($p < 0.001$) levels of significance have been depicted in Table 11.1.

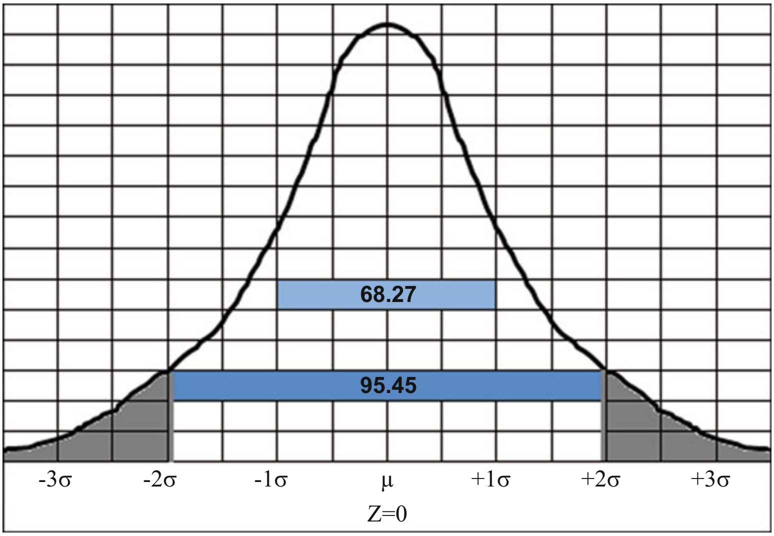


Fig. 11.1 Normal distribution curve showing shaded areas for variate beyond −1.96 to 1.96 in a two-tailed test

Table 11.1 Critical values of standard normal variate “*Z*”

	Level of significance		
Critical value of variate “*Z*”	5%	1%	0.01%
One-tailed test	-1.645	-2.33	-3.08
	or 1.645	or 2.33	or 3.08
Two-tailed test	-1.96	-2.85	-3.27
	and 1.96	and 2.85	and 3.27

11.9 Student’s “*t*”-Distribution

When samples are large ($n > 30$), the sampling distribution is approximately normal. In case of small samples ($n \leq 30$), the distribution cannot be considered normal. The sample distribution may swing to the left or right with decrease in sample size. The “small sampling theory” is applicable to both the small and large samples. It is also branded as “exact sampling theory.”

$$\text{Statistics} : t = \frac{|\bar{x} - \mu|}{{}^{s}\!/\!_{\sqrt{n}}}$$

Suppose samples of size “n” are drawn from a normal population or approximately normal population with population mean “μ” and standard deviation σ. For each sample if we compute “t” using the sample mean “$\bar{X}$” and standard deviation “s,” we would get a sampling distribution for “t.” This distribution is given by the formula

$$Y = \frac{Y_{\mathrm{o}}}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}$$

wherein, Y_{o} is a constant depending on “n” such that the total area under the curve is 1.

This distribution is known as Student’s “t”-distribution. This was postulated and published by Gosset under the pseudonym “Student” during the early part of the twentieth century.

For large sample size ($n \geq 30$), this curve closely approximates to the “standard normal curve.”

$$\mathrm{Y} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(t)^2}{2}}.$$

11.10 Confidence Intervals

Using the table of *t*-distribution, we can define 95% and 99% confidence intervals. By doing so, we will be able to estimate the population parameter (μ) within specified limits.

If $-t_{0.\ 025}$ and $t_{0.025}$ are the values of "*t*"-distribution for which 2.5% of the area lies in each tail of the "*t*"-distribution curve/graph, then 95% confidence interval for *t* is

$$-t_{0.025} < \frac{|\bar{x} - \mu|}{s/\sqrt{n}} < t_{0.025}$$

Therefore:

$$-t_{0.025} \times \frac{s}{\sqrt{n}} < (|\bar{X} - \mu|) < t_{0.025} \times \frac{s}{\sqrt{n}} \quad \text{or}$$

$$-t_{0.025} \times \frac{s}{\sqrt{n}} < \mu < \bar{X} < t_{0.025} \times \frac{s}{\sqrt{n}}$$

Hence, μ lies in the interval given above with 95% confidence (i.e., probability 0.95). The $\mathrm{t}_{0.025}$ represents 97.5th percentile value, while $-t_{0.\ 025}$ represents 2.5th percentile value. In general, we can represent confidence limits for population means by

$$\bar{X} \pm t_{\mathrm{c}} \times \frac{s}{\sqrt{n}}$$

where, t_c = critical value or confidence coefficient.

This depends on the level of confidence desired and the sample size.

11.11 Applications of Student's *t*-Test

Student's *t*-test is applied in a variety of ways for statistical analysis as listed below:

1. Comparison of sample with population
2. Comparison of sample with sample
3. Comparison of sample with sample by "*t*-paired test"

11.11.1 Comparison of Sample with Population

Example 1 Ten individuals are chosen at random from a population, and their heights are measured in inches and noted down as 65, 66, 66, 67, 68, 69, 70, 70, 71, and 71. In the light of this data, discuss the suggestion that mean height in the population is 67 inches.

The following conditions are fulfilled:

1. The population distribution of height is approximately normally distributed.
2. The random sample has been drawn from the population.

Suggestions
Null hypothesis (H_o): $\mu = 67''$
Alternate hypothesis (H_i): $\mu \neq 67''$

Solution

$$\bar{X} = \frac{\Sigma X}{N} = \frac{683}{10} = 68.3''$$

$$s = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{44.1}{9}} = \sqrt{4.9} = 2.214$$

$$t = \frac{|\bar{x} - \mu|}{{}^{s}\!/\!\sqrt{n} = \frac{68.3-67}{2.214} \times \sqrt{10} = \frac{1.3}{2.214} \times \sqrt{10} = 1.855}$$

Degrees of freedom (df) = 10−1 = 9
Referring to the table values of t-distribution for 9 df, we get $t_{0.05} = 2.262$.
Here, $t < t_{0.05}$; therefore $p > 0.05$. Hence, it is not significant at 5% level of significance.

Conclusion
We fail to reject the null hypothesis (H_o). Hence, we accept with 95% confidence that the mean height of population may be equal to 67″ from which this random sample has been drawn.

11.11.2 Comparison of Sample with Sample

Requirements for comparison of two samples

	Sample I	Sample II
Mean	$\bar{X}_1 = \frac{\Sigma X}{n_1}$	$\bar{X}_2 = \frac{\Sigma X}{n_2}$
Standard deviation	$s_1 = \sqrt{\frac{\Sigma\left(X - \bar{X}_1\right)^2}{n_1 - 1}}$	$s_2 = \sqrt{\frac{\Sigma\left(X - \bar{X}_2\right)^2}{n_2 - 1}}$

If both the samples are drawn from the same population, then the estimates s_1 and s_2 may be pooled, to get a better estimate of the population's "standard deviation" (s_p). The pooled estimate is worked out by the following formula:

$$s_p = \sqrt{\frac{\Sigma\left(X-\bar{X}_1\right)^2 + \Sigma\left(X-\bar{X}_2\right)^2}{n_1+n_2-2}}$$

$$\left(s_p\right)^2 = \frac{\Sigma\left(X-\bar{X}_1\right)^2 + \Sigma\left(X-\bar{X}_2\right)^2}{n_1+n_2-2}$$

Degrees of freedom (df) $= n_1 + n_2 - 2$

$$\text{Now} : t = \frac{|\bar{X}_1 - \bar{X}_2|}{{}^{s_p}/_{\sqrt{n}}} \quad \text{or}$$

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1 2}{n_1} + \frac{s_2 2}{n_2}}}$$

We have to test whether this value of "t" is $>t_{0.05}$ or $<t_{0.05}$ for the said degrees of freedom (df). If $t > t_{0.05}$, then we reject the null hypothesis (H_o). If $t < t_{0.05}$, then we accept the null hypothesis.

Example 2 An experiment was conducted for assessing the effect of vitamin A-deficient diet. Out of the 20 inbred rats, 10 rats were fed on normal diet, and the other 10 were fed on vitamin A-deficient diet. The amount of vitamin A in the serum of rats of both the groups was determined, and mean and standard deviation was worked out as shown in Table 11.2.

(a) Find out whether the mean value ($\bar{X}_2$) of rats fed on vitamin A-deficient diet is the same as the mean value ($\bar{X}_1$) of those fed on normal diet.
(b) If difference is there, prove that this difference is due to sampling variation or due to the deficiency of vitamin A.

Solution

(a) The absolute difference between means of groups: $|\bar{X}_1 - \bar{X}_2| =$ 3375–2570 = 805 IU.

Table 11.2 Mean and standard deviation of vitamin A levels in two groups of inbred rats

	Normal diet (vitamin A, IU)	Vitamin A-deficient diet (vitamin A, IU)
Mean	$\bar{X}_1 = 3375$	$\bar{X}_2 = 2570$
Standard deviation	$s_1 = 626$	$s_2 = 533$

(b) Now 805 IU is the absolute difference of means of two groups ($|\bar{X}_1 - \bar{X}_2|$) of samples n_1 and n_2 from the normal populations.

Under the null hypothesis, the statistical value of "*t*" for the sampling distributions is worked out as

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1 2}{n_1} + \frac{s_2 2}{n_2}}} = \frac{805}{260} = 3.096$$

df = 10 + 10–2 = 18

We have to find out whether this value of "*t*" is $>t_{0.05}$ or $<t_{0.05}$ at the said degrees of freedom. If value of "*t*" is $<t_{0.05}$, then we reject the null hypothesis, and if it is $>t_{0.05}$, we accept the null hypothesis. When the value of "*t*" is $>t_{0.05}$ at df = 18, we conclude that $p < 0.05$.

In the example cited above, the value of "*t*" is 3.096 and is even $>t_{0.01}$. So, $p < 0.01$. The difference is due to the deficiency of vitamin A. Hence, null hypothesis is rejected.

11.11.3 Comparison of Sample with Sample by "*t*-Paired Test"

The "*t*-paired test" is applied when two sets of observations are to be compared from the same subjects/patients. The data could be sets of clinical/biological or biochemical investigations on a group of patients.

Example 3 In a group of nine hypertensive patients, systolic blood pressure was recorded in mm of Hg, before and after treatment as exhibited in Table 11.3. Test the significance of treatment on patients or accept the null hypothesis (H_o).

Table 11.3 Systolic blood pressure (BP) in nine patients before and after the treatment

Subjects	BP before treatment (A)	BP after treatment (B)	Difference *d* = A–B	d^2
1	160	130	+30	900
2	145	123	+17	289
3	132	132	0	0
4	140	130	+10	100
5	132	120	+12	144
6	154	125	+29	841
7	136	125	+11	121
8	134	136	−2	4
9	132	136	−4	16
Sum	**1265**	**1162**	**+103**	**2415**

Solution

Formula for *t*-paired test:

$$t = \frac{|\bar{d}|}{\text{S.E.}} \text{ OR } t = \frac{|\bar{d}|}{\frac{\text{SD}}{\sqrt{n}}} \text{ OR } t = \frac{|\bar{d}|}{\sqrt{\frac{s^2}{n}}}$$

Mean difference $(\bar{d}) = \frac{103}{9} = 11.4.$

Square of standard deviation: $s^2 = \frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1} = \frac{2415 - 1178.8}{8} = \frac{1236.2}{8}$
$= 154.5.$

Standard deviation: $s = \sqrt{154.5} = 12.43.$

Now: $t = \frac{|\bar{d}|}{\frac{\text{SD}}{\sqrt{n}}} = \frac{|\bar{d}|}{s} \times \sqrt{n} = \frac{11.4}{12.4} \times \sqrt{9} = \frac{11.4}{12.4} \times 3 = 2.76.$

Degrees of freedom (df) = 9–1 = 8.
Table values of "*t*" at df = 8 are $t_{0.05} = 2.306$ and $t_{0.01} = 3.355$.

Since the computed *t*-value (2.76) for the given data is $>t_{0.05}$ (2.306), which translates to $p < 0.05$, so we reject the null hypothesis (H_o). In this case the *t* (2.76) is $<t_{0.01}$ (3.355). Hence, we conclude that the difference due to treatment with a certain drug is significant at 5% level ($p < 0.05$).

Variance-Ratio Test and Analysis of Variance (ANOVA)

12

12.1 Applications of the Variance-Ratio Test (*F*-Test)

F-test is applied when only two samples are to be compared. The following assumptions are taken into consideration:

1. Samples should have been drawn at random.
2. The observations in each group should be normally distributed.
3. Samples should be independent.
4. Since *F*-distribution is a ratio of "squared values" (s_1^2 and s_2^2), it would never be negative.
5. Null hypothesis (H_o) prevails if variance of all the populations is equal.

Let us practice the applications of *F*-test with solved examples.

Example 1 In an ultrastructural morphometric study on renal biopsies, "glomerular basement membrane thickness" (GBMT) was measured in ten cases affected with "diabetic nephropathy" (DN) as well as in equal number of normal controls. Mean GBMT in DN cases was 560 nm with "standard deviation" (s) of 45 nm, and mean GBMT in controls was 325 nm with "standard deviation" of 18 nm. Apply *F*-test to ascertain the significance of the study.

Solution

Here: $s_1 = 45$ and $s_2 = 18$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$.

$$F = \frac{s_1^2}{s_2^2} = \frac{45^2}{18^2} = \frac{2025}{324} = 6.25$$

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_12

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 6.25$ is greater than 3.18 ($p > 0.05$). Hence, the difference in variances is significant. So, both the populations are heterogenous.

Example 2 In a clinical study, "systolic blood pressure" (SBP) was recorded in ten smokers and ten non-smoker adults. In the smokers, the mean SBP was 155 mm of Hg with "standard deviation" (s) of 19 mm of Hg, and in non-smokers the mean SBP was 130 mm of Hg with "standard deviation" (s) of 10 mm of Hg. Apply F-test to verify the null hypothesis (H_o).

Solution

Here: $s_1 = 19$ and $s_2 = 10$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$

$$F = \frac{s_1^2}{s_2^2} = \frac{19^2}{10^2} = \frac{361}{100} = 3.61$$

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 3.61$ is greater than 3.18 ($p > 0.05$). Hence, the null hypothesis is rejected as the difference in variances is significant. So, smokers and non-smokers are distinct groups of population.

Example 3 In an ultrastructural morphometric study on renal biopsies, "glomerular basement membrane thickness" (GBMT) was measured in ten cases affected with "minimal change disease" (MCD) as well as in equal number of normal controls. Mean GBMT in MCD cases was 320 nm with "standard deviation" (s) of 24 nm, and mean GBMT in controls was 316 nm with "standard deviation" of 21 nm. Apply F-test to ascertain the significance of the study.

Solution

Here: $s_1 = 24$ and $s_2 = 21$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$

$$F = \frac{s_1^2}{s_2^2} = \frac{24^2}{21^2} = \frac{576}{441} = 1.306$$

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 1.306$ is less than 3.18 ($p > 0.05$). Hence, the difference in variances is not significant. So, both the populations are equal as these have the same variance.

Example 4 In an ultrastructural morphometric study on renal biopsies, "glomerular basement membrane thickness" (GBMT) was measured in ten cases affected with "membranous glomerulonephritis" (MGN) as well as in equal number of normal controls. Mean GBMT in MGN cases was 1050 nm with "standard deviation" (s) of 258 nm, and mean GBMT in controls was 335 nm with "standard deviation" of 38 nm. Apply F-test to ascertain the significance of the study.

Solution

Here: $s_1 = 258$ and $s_2 = 38$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$

$$F = \frac{s_1^2}{s_2^2} = \frac{258^2}{38^2} = \frac{66{,}564}{1444} = 46.096$$

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 46.096$ is very much greater than 3.18 ($p < 0.05$). Hence, the difference in variances is highly significant. So, both the samples belong to different populations.

Example 5 In an ultrastructural morphometric study on renal biopsies, “glomerular basement membrane thickness” (GBMT) was measured in ten cases affected with “thin basement membrane disease” (TBMD) as well as in equal number of normal controls. Mean GBMT in TBMD cases was 218 nm with “standard deviation” (*s*) of 40 nm, and mean GBMT in controls was 316 nm with “standard deviation” of 17 nm. Apply *F*-test to ascertain the significance of the study.

Solution

Here: $s_1 = 40$ and $s_2 = 17$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$

$$F = \frac{s_1^2}{s_2^2} = \frac{40^2}{17^2} = \frac{1600}{289} = 5.536$$

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 5.536$ is greater than 3.18 ($p < 0.05$). Hence, the difference in variances is significant. So, both the populations are distinct.

Example 6 In an ultrastructural morphometric study on renal biopsies, “glomerular basement membrane thickness” (GBMT) was measured in ten cases diagnosed as “Alport’s syndrome” (AS) as well as in equal number of normal controls. Mean GBMT in AS cases was 366 nm with “standard deviation” (*s*) of 70 nm, and mean GBMT in controls was 319 nm with “standard deviation” of 22 nm. Apply *F*-test to ascertain the significance of the study.

Solution

Here: $s_1 = 70$ and $s_2 = 22$

$v_1 = 10 - 1 = 9$ and $v_2 = 10 - 1 = 9$

$$F = \frac{s_1^2}{s_2^2} = \frac{70^2}{22^2} = \frac{4900}{484} = 10.123$$

Table value of $F_{0.05} = 3.18$ (at $v_1 = 9$ and $v_2 = 9$). Observed value of $F = 10.123$ is very much greater than 3.18 ($p < 0.05$). Hence, the difference in variances is significant. So, both the populations are distinct.

12.1.1 Analysis of Variance (ANOVA)

Analysis of variance splits the variance into two components when more than two samples are to be compared for variance equality as per null hypothesis (H_o). These splits of variance are:

1. Variance within the samples
2. Variance between the samples

We have learnt that F-test was applied when there were only two samples. ANOVA is used to test the H_o among more than two samples. We can apply "direct method," "shortcut method," or "coding method" may it be "one-way classification" or "two-way classification."

Let us learn the applications of all the three methods under *one-way classification*.

12.1.1.1 Direct Method

Example 7 In an ultrastructural study on renal biopsies, "glomerular basement membrane thickness" (GBMT) was measured, and four samples of ten adult individuals each were drawn out considering them having normal renal morphology. Apply ANOVA to test the null hypothesis (H_o) to ascertain that these samples belong to the normal population. Observations of GBMT have been tabulated in Table 12.1.

Solution

The null hypothesis states that there is no difference in the variance of four samples (groups), i.e., $s_1^2 = s_2^2 = s_3^2 = s_4^2$.

Table 12.1 GBM thickness in four normal groups

Sr. no.	Sample I	Sample II	Sample III	Sample IV
1.	338	299	338	297
2.	297	293	308	338
3.	299	309	297	299
4.	297	307	339	302
5.	320	286	299	316
6.	293	403	297	320
7.	309	326	294	362
8.	307	363	307	297
9.	298	307	309	320
10.	358	358	326	293
Mean	**312**	**325**	**311**	**314**
SD	**21**	**38**	**17**	**22**

ANOVA is completed in seven steps:

1. Calculate the mean ($\bar{x}$) of all the four samples.
2. Calculate the grand mean ($\bar{\bar{x}}$), that is, the mean of four means.
3. Calculate the sum squares between the samples after calculating squares of differences of the sample mean from the grand mean multiplied by the number of observations in each sample. It is termed as SSC (sum of squares between columns or samples).
4. Calculate the sum of squares within samples after calculating squares of differences of observations from the sample mean and then taking sum of these in the bottom row in each column. It is termed as SSW or SSE (sum of squares within samples or the sum of squares of error from the mean).
5. Calculate the "total sum of squares" (SST), i.e., the sum of SSC and SSE (SSC + SSE).
6. Put the above information in the ANOVA table, and work out the "mean sum of squares of columns" (MSC) and "mean sum of squares of error" (MSE) with reference to v_1 ($C-1$) and v_2 ($n-1$).
7. Decision is taken upon the F value.

Let us apply these steps to the data illustrated in Table 12.1.

Step I: The mean of each sample has been calculated and mentioned in the table as $\bar{x}_1$ $=312$, $\bar{x}_2=325$, $\bar{x}_3=311$, and $\bar{x}_4=314$.

Step II: Grand mean $(\bar{\bar{x}}) = \dfrac{\bar{x}_1+\bar{x}_2+\bar{x}_3+\bar{x}_4}{4} = \dfrac{312+325+311+314}{4} = 315.5$

Step III: Calculation of SSC:

$$\begin{aligned}\mathbf{SSC} &= n_1\left(\bar{x}_1-\bar{\bar{x}}\right)^2 + n_2\left(\bar{x}_2-\bar{\bar{x}}\right)^2 + n_3\left(\bar{x}_3-\bar{\bar{x}}\right)^2 + n_4\left(\bar{x}_4-\bar{\bar{x}}\right)^2\\ &= 10(-3.5)^2 + 10(9.5)^2 + 10(-4.5)^2 + 10(-1.5)^2\\ &= 10\times 12.25 + 10\times 90.25 + 10\times 20.25 + 10\times 2.25\\ &= \ 122.5 + 902.5 + 202.5 + 22.5 = \mathbf{1250}\end{aligned}$$

Step IV: Calculation of the sum of squares of differences from the mean within samples and sum of all these sums in the last row of all the columns. It is also known as SSW or SSE as shown in Table 12.2.

$$\mathbf{SSE} = 4066 + 12,743 + 2592 + 4344 = 23,745$$

Step V: Calculation of the "total sum of squares between the samples and within the samples" (SST):

Table 12.2 Calculation of SSE of four samples

Sr. No.	X_1	$(x_1 - \bar{x}_1)^2$	X_2	$(x_2 - \bar{x}_2)^2$	X_3	$(x_3 - \bar{x}_3)^2$	X_4	$(x_4 - \bar{x}_4)^2$
1.	338	676	299	676	338	729	297	289
2.	297	225	293	1024	308	9	338	576
3.	299	169	309	256	297	196	299	225
4.	297	225	307	324	339	784	302	144
5.	320	64	286	1521	299	144	316	4
6.	293	361	403	6084	297	196	320	36
7.	309	9	326	1	294	289	362	2304
8.	307	25	363	1444	307	16	297	289
9.	298	196	307	324	309	4	320	36
1.0	358	2116	358	1089	326	225	293	441
Mean	**312**		**325**		**311**		**314**	
SUM		**4066**		**12,743**		**2592**		**4344**

$$\textbf{SST} = \text{SSC} + \text{SSE} = 1250 + 23,745 = 24,995$$

Step VI: Plot the above values in the ANOVA table.

ANOVA table

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean square**	**Tests**
Between the columns (between samples)	"SSC"	$V_1 = C{-}1$	**MSC** $= \frac{\text{SSC}}{C-1}$	$\boldsymbol{F} = \frac{\textbf{MSC}}{\textbf{MSE}}$
	1250	4–1 = 3	$\frac{1250}{3} = 416.66$	$= \frac{\textbf{416.66}}{\textbf{659.58}}$ $= \mathbf{0.632}$
Within the columns (within samples)	"SSE"	$V_2 = n{-}c$	**MSE** $= \frac{\text{SSC}}{n-c}$	
	23,745	40–4 = 36	$\frac{23745}{36} = 659.58$	
Total	**SST = 24,995**	$\boldsymbol{V = n-1}$		
		40–1 = 39		

Finding table value of "*F*" in "Fisher's *F*-table"

Degrees of freedom: $V_1 = c-1 = 4–1 = 3$, $V_2 = 40–4 = 36$

$V_1 = 3$ (numerator)

$V_2 = 36$ (denominator)

$\boldsymbol{F_{0.05} = 2.88}$

Step VII: *Decision*

Calculated value of $F = 0.632$ is less than table value $F_{0.05} = 2.88$ at 5% level. So, the difference in variances is not significant. The null hypothesis (H_o) is accepted. Hence, all the four samples come from the equal variance population.

Table 12.3 Calculation of squares of observations

$N = 10$	X_1	x_1^2	X_2	x_2^2	X_3	x_3^2	X_4	x_4^2
1.	338	114,244	299	89,401	338	114,244	297	88,209
2.	297	88,209	293	85,849	308	94,864	338	114,244
3.	299	89,401	309	95,481	297	88,209	299	89,401
4.	297	88,209	307	94,249	339	114,921	302	91,204
5.	320	102,400	286	81,796	299	89,401	316	99,856
6.	293	85,849	403	162,409	297	88,209	320	102,400
7.	309	95,481	326	106,276	294	86,436	362	131,044
8.	307	94,249	363	131,769	307	94,249	297	88,209
9.	298	88,804	307	94,249	309	95,481	320	102,400
10.	358	128,164	358	128,164	326	106,276	293	85,849
Sum	**3116**		**3251**		**3114**		**3144**	
Sum		**975,010**		**1,069,643**		**972,290**		**992,816**

12.1.2 Shortcut Method

Let us apply the "shortcut method" to solve "Example 7." The data has been arranged in Table 12.3, and squares of observations of all the four samples have been calculated and tabulated.

Solution

Step I: Calculate the total of all the items as:

$$\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4 = 3116 + 3251 + 3114 + 3144 = 12,625\ (T)$$

Step II: Calculate the "correction factor" as $\frac{T^2}{N} = \frac{12{,}625 \times 12{,}625}{40} = 3{,}984{,}765$.
Step III: Calculate the "sum of squares of all the observations" (SST):

$$\begin{aligned}&\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2 - \frac{T^2}{N}\\&= 975,010 + 1,069,643 + 972,290 + 992,816 - 3,984,765 = 24,995\\&\mathrm{SST} = 24,995\end{aligned}$$

Step IV: Calculate the "sum of squares of columns" (SSC):

$$
\begin{aligned}
\frac{(\sum x_1)^2}{N_1}+\frac{(\sum x_2)^2}{N_2}+\frac{(\sum x_3)^2}{N_3}+\frac{(\sum x_4)^2}{N_4}-\frac{T^2}{N} &= \frac{(3116)^2}{10} \\
&+\frac{(3251)^2}{10}+\frac{(3114)^2}{10}+\frac{(3144)^2}{10}-\frac{12625^2}{40} \\
&= 970,945 + 1,056,900 + 969,699 + 988,473 - 3,984,765 \\
&= 1252 \mathrm{SSC} = 1252
\end{aligned}
$$

Step V: Calculate the "sum of squares within samples" (SSE):

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSC} = 24,995 - 1252 = 23,743$$

Step VI: Plot the above values in the ANOVA table.

ANOVA table

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean square**	**Tests**
Between the columns (between samples)	"SSC"	$V_1 = C-1$	**MSC** $= \frac{\mathrm{SSC}}{C-1}$	$\boldsymbol{F} = \frac{\mathbf{MSC}}{\mathbf{MSE}}$
	1252	4–1 = 3	$\frac{1252}{3} = 417.33$	$= \frac{\mathbf{417.33}}{\mathbf{659.52}}$ $=\mathbf{0.632}$
Within the columns (within samples)	"SSE"	$V_2 = n-c$	**MSE** $= \frac{\mathrm{SSC}}{n-\mathrm{c}}$	
	23,743	40–4 = 36	$\frac{23743}{36} = 659.52$	
Total	**SST = 24,995**	$\boldsymbol{V = n-1}$		
		40–1 = 39		

Finding table value of "F" in "Fisher's F-table"
Degrees of freedom: $V_1 = c-1 = 4–1 = 3$, $V_2 = 40–4 = 36$
$V_1 = 3$ (numerator)
$V_2 = 36$ (denominator)
$\boldsymbol{F_{0.05} = 2.88}$

Step VII: *Decision*

Calculated value of $F = 0.632$ is less than table value $F_{0.05} = 2.88$ at 5% level. So, the difference in variances is not significant. The null hypothesis (H_o) is accepted. Hence, all the four samples come from the equal variance population.

Table 12.4 Calculation of squares after coding

$N = 10$	X_1	x_1^2	X_2	x_2^2	X_3	x_3^2	X_4	x_4^2
1	38	1444	−1	1	38	1444	−3	9
2	−3	9	−7	49	8	64	38	1444
3	−1	1	9	81	−3	9	−1	1
4	−3	9	7	49	39	1521	2	4
5	20	400	−14	196	−1	1	16	256
6	−7	49	103	10,609	−3	9	20	400
7	9	81	26	676	−6	36	62	3844
8	7	49	63	3969	7	49	−3	9
9	−2	4	7	49	9	81	20	400
10	58	3364	58	3364	26	676	−7	49
SUM	**116**		**251**		**114**		**144**	
SUM		**5410**		**19,043**		**3890**		**6416**

12.1.3 Coding Method

The "coding method" is considered the best method as we can compute the result more quickly. Let us apply it to the problem illustrated as Example 7. Let us deduct 300 from each item (observation) in all the samples and tabulate in Table 12.4.

Solution

Step I: Calculate the total of all the items as:

$$\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4 = 116 + 251 + 114 + 144 = 625\ (T)$$

Step II: Calculate the "correction factor" as $\frac{T^2}{N} = \frac{625 \times 625}{40} = 9765.$
Step III: Calculate the "sum of squares of all the observations" (SST):

$$\begin{aligned}&\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2 - \frac{T^2}{N}\\&= 5410 + 19{,}043 + 3890 + 6416 - 9765 = 24{,}994\\&\text{SST} = 24{,}994\end{aligned}$$

Step IV: Calculate the "sum of squares of columns" (SSC):

$$\frac{\left(\sum x_1\right)^2}{N_1}+\frac{\left(\sum x_2\right)^2}{N_2}+\frac{\left(\sum x_3\right)^2}{N_3}+\frac{\left(\sum x_4\right)^2}{N_4}-\frac{T^2}{N}=\frac{(116)^2}{10}$$
$$+\frac{(251)^2}{10}+\frac{(114)^2}{10}+\frac{(144)^2}{10}-\frac{625^2}{40}=1346$$
$$+6300+1299+2073-9765=1253\text{SSC}=1253$$

Step V: Calculate the "sum of squares within samples" (SSE):

$$\text{SSE}=\text{SST}-\text{SSC}=24{,}994-1253=23{,}741$$

Step VI: Plot the above values in the ANOVA table.

ANOVA table

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean square**	**Tests**
Between the columns (between samples)	"SSC"	$V_1 = C{-}1$	**MSC** $= \frac{\text{SSC}}{C-1}$	$\boldsymbol{F} = \frac{\textbf{MSC}}{\textbf{MSE}}$
	1253	4–1 = 3	$\frac{1253}{3} = 417.66$	$= \frac{\textbf{417.66}}{\textbf{659.47}}$ $=$**0.633**
Within the columns (within samples)	"SSE"	$V_2 = n{-}c$	**MSE** $= \frac{\text{SSC}}{n-\text{c}}$	
	23,741	40–4 = 36	$\frac{23741}{36} = 659.47$	
Total	**SST = 24,994**	$\boldsymbol{V = n{-}1}$		
		40–1 = 39		

Finding table value of "F" in "Fisher's F-table"
Degrees of freedom: $V_1 = c-1 = 4–1 = 3$, $V_2 = 40–4 = 36$
$V_1 = 3$ (numerator)
$V_2 = 36$ (denominator)
$\boldsymbol{F_{0.05} = 2.88}$

Step VII: *Decision*

Calculated value of $F = 0.633$ is less than table value $F_{0.05} = 2.88$ at 5% level. So, the difference in variances is not significant. The null hypothesis (H_o) is accepted. Hence, all the four samples come from the equal variance population.

Table 12.5 Calculation of squares after coding

$N = 10$	X_1	x_1^2	X_2	x_2^2	X_3	x_3^2	X_4	x_4^2
1	3	9	4	16	13	169	8	64
2	13	169	2	4	12	144	4	16
3	5	25	0	0	2	4	−3	9
4	9	81	2	4	12	144	9	81
5	11	121	5	25	2	4	3	9
6	15	225	1	1	14	196	17	289
7	13	169	2	4	12	144	12	144
8	7	49	−4	16	14	196	9	81
9	6	36	2	4	20	400	10	100
10	8	64	6	36	6	36	9	81
Sum	**90**		**20**		**107**		**78**	
Sum		**948**		**110**		**1437**		**874**

12.1.4 Coding Method

Example 8 The percentage volume occupied by glomerular capillary space was measured through interactive morphometric method in ten cases each of MCD, MGN, TBMD, and AS. The data have been tabulated in Table 12.5 after deducting 30 from each item. Test the "null hypothesis" (H_o) by "one-way ANOVA" through coding method.

Solution

Step I: Calculate the total of all the items as:

$$\sum x_1 + \sum x_2 + \sum x_3 + \sum x_4 = 90 + 20 + 107 + 78 = 295\ (T)$$

Step II: Calculate the "correction factor" as $\frac{T^2}{N} = \frac{295\times 295}{40} = 2175$.
Step III: Calculate the "sum of squares of all the observations" (SST):

$$\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2 - \frac{T^2}{N}$$
$$= 948 + 110 + 1437 + 874 - 2175 = 1194$$
$$\text{SST} = 1194$$

Step IV: Calculate "sum of squares of columns" (SSC):

$$\begin{aligned}
&\frac{\left(\sum x_1\right)^2}{N_1}+\frac{\left(\sum x_2\right)^2}{N_2}+\frac{\left(\sum x_3\right)^2}{N_3}+\frac{\left(\sum x_4\right)^2}{N_4}-\frac{T^2}{N}\\
&=\frac{(90)^2}{10}+\frac{(20)^2}{10}+\frac{(107)^2}{10}+\frac{(78)^2}{10}-\frac{295^2}{40}\\
&=810+40+1145+608-2175=428\\
&\mathrm{SSC}=428
\end{aligned}$$

Step V: Calculate the "sum of squares within samples" (SSE):

$$\mathrm{SSE}=\mathrm{SST}-\mathrm{SSC}=1194-428=766$$

Step VI: Plot the above values in the ANOVA table.

ANOVA table

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean square**	**Tests**
Between the columns (between samples)	"SSC"	$V_1 = C-1$	**MSC** $= \frac{\mathrm{SSC}}{\mathrm{C}-1}$	$\boldsymbol{F} = \frac{\mathbf{MSC}}{\mathbf{MSE}}$
	428	4–1 = 3	$\frac{428}{3} = 142.67$	$= \frac{\mathbf{142.67}}{\mathbf{21.28}}$ =**6.70**
Within the columns (within samples)	"SSE"	$V_2 = n-c$	**MSE** $= \frac{\mathrm{SSC}}{n-c}$	
	766	40–4 = 36	$\frac{766}{36} = 21.28$	
Total	**SST = 1194**	$\boldsymbol{V = n-1}$		
		40–1 = 39		

Finding table value of "*F*" in "Fisher's *F*-table"

Degrees of freedom: $V_1 = c-1 = 4\text{–}1 = 3$, $V_2 = 40\text{–}4 = 36$

$V_1 = 3$ (numerator)

$V_2 = 36$ (denominator)

$\boldsymbol{F_{0.05} = 2.88}$

Step VII: Decision

Calculated value of $F = 6.70$ is very much greater than table value $F_{0.05} = 2.88$ at 5% level. So, the difference in variances is highly significant. So, the null hypothesis (H_o) is rejected. Hence, these renal diseases cause pathological variations in glomerular capillary space volume fraction.

Let us learn the applications of the ANOVA method under *two-way classification*.

Assumptions The variance of two factors is compared in the "two-way ANOVA." The variance in one factor is studied in columns, and the other is studied in rows. The

Table 12.6 Biochemical investigations done on four spectrophotometers

Biochemists	SPM-I	SPM-II	SPM-III	SPM-IV
1	30	34	40	30
2	35	24	35	22
3	27	28	30	26
4	38	30	40	34
5	34	25	38	28

following "sum of squares" are computed and incorporated in the "two-way ANOVA table":

1. SSC = sum of squares of columns.
2. SSR = sum of squares of rows.
3. SSE = sum of squares of residual error.
4. SSC, SSR, and SSC values are incorporated in the two-way ANOVA table which is illustrated below.

Two-way ANOVA table

Source of variation	**Sum of squares**	**Degrees of freedom**	**Mean squares**	**Tests**
Between column	SSC	$V_1 = C-1$	$\text{MSC} = \frac{\text{SSC}}{C-1}$	$F_1 = \frac{\text{MSC}}{\text{MSE}}$
Between rows	SSR	$V_2 = R-1$	$\text{MSR} = \frac{\text{SSR}}{R-1}$	$F_2 = \frac{\text{MSR}}{\text{MSE}}$
Residual error	SSE	$V_3 = (C-1)(R-1)$	$\text{MSE} = \frac{\text{SSE}}{(C-1)(R-1)}$	
Total	SST	$V = N-1$		

Example 9 Five biochemists were assigned a task of performing biochemical investigations using four different spectrophotometers for a day. The data of the number of investigations performed by them has been tabulated below in Table 12.6.

Test

(a) Whether mean productivity of spectrophotometers differs or is the same?
(b) Whether the productivity of biochemists differs or is the same?

Solution

All the three methods applied for "one-way ANOVA" can be applicable for "two-way ANOVA" also, but we would use the most convenient method, the coding method in this case. Let us test the hypothesis that the "mean productivity" of spectrophotometers or biochemists does not differ.

Deducting 30 (i.e., coding from 30) from each item, we get the values as tabulated in Table 12.7.

Step I: Compute the total (T) of rows + columns = 28

Step II: Compute the correction factor $\frac{T^2}{N} = \frac{28^2}{20} = 39.2$.

Table 12.7 Biochemical investigations done on four spectrophotometers (after coding from 30)

Biochemists	SPM-A	SPM-B	SPM-C	SPM-D	Total
1	0	4	10	0	14
2	5	−6	5	−8	−4
3	−3	−2	0	−4	−9
4	8	0	10	4	22
5	4	−5	8	−2	5
Total	**14**	**−9**	**33**	**−10**	**28**

Step III: Compute SSC as:

$$\frac{(\sum x_1)^2}{N_1}+\frac{(\sum x_2)^2}{N_2}+\frac{(\sum x_3)^2}{N_3}+\frac{(\sum x_4)^2}{N_4}-\frac{T^2}{N}$$
$$=\frac{(14)^2}{5}+\frac{(-9)^2}{5}+\frac{(33)^2}{5}+\frac{(-10)^2}{5}-\frac{28^2}{20}$$
$$=39.2+16.2+217.8+20-39.2=254\ (\text{SSC})$$

Step IV: Compute the sum of squares of rows (SSR) as:

$$\frac{(14)^2}{4}+\frac{(-4)^2}{4}+\frac{(-9)^2}{4}+\frac{(22)^2}{4}+\frac{(5)^2}{4}-\frac{28^2}{20}$$
$$=49+4+20.25+121+6.25-39.2=161.3(\text{SSR})$$

Step V: Compute the sum of square of total (SST) square of all the 20 items as:

$$0+25+9+64+16+16+36+4+0+25$$
$$+100+25+0+100+64+0+64+16+16+4-39.2$$
$$=114+81+289+100-39.2=544.8(\text{SST})$$

Step VI: Compute SSE as:

$$\text{SSE}=\text{SST}-(\text{SSC}+\text{SSR})=544.8-(254+161.3)=129.5$$

Two-way ANOVA table				
Variation Sources	**Sum of squares**	**Degrees of freedom**	**Mean squares**	**Tests**
Between column	SSC	$V_1 = C-1$	MSC $= \frac{254}{3}$	$F_1 = \frac{84.66}{10.79}$
	(254)	=4–1 = 3	=84.66	=7.85
Between rows	SSR	$V_2 = R-1$	MSR $= \frac{161.3}{4}$	$F_2 = \frac{40.32}{10.79}$
	(161.3)	=5–1 = 4	=40.32	3.74
Residual error	SSE	$V_3 = (C-1)(R-1)$	MSE $= \frac{129.5}{12}$	
	(129.5)	=3×4 = 12	=10.79	
Total	**SST**	$\boldsymbol{V = N-1}$		
	(544.8)	**=20–1 = 19**		

Step VII

(a) *Decision Regarding Spectrophotometers*

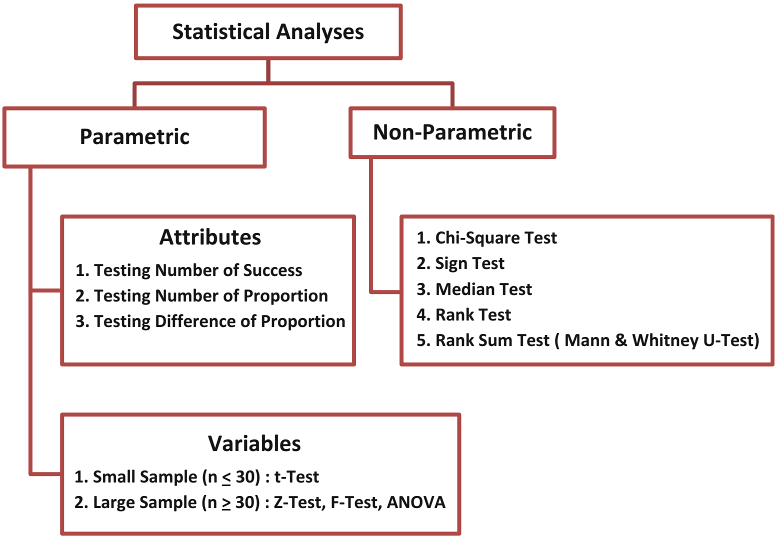
Computed value of F (7.85) is greater than $F_{0.05}$ (3.49) at $V_1 = 3$ and $V_2 = 12$. So, the difference is significant at 5% level of significance. The null hypothesis (H_o) is not accepted. Hence, spectrophotometers differ significantly in mean productivity.

(b) *Decision Regarding Biochemists*

Computed value of F (3.47) is greater than $F_{0.05}$ (3.26) at $V_1 = 4$ and $V_2 = 12$. So, the difference is significant at 5% level of significance. The null hypothesis (H_o) is not accepted. Hence, biochemists also differ significantly in their mean productivity.

Nonparametric Statistical Tests 13

13.1 Statistical Algorithm


Statistical Analyses
Parametric
Non-Parametric
Attributes
1. Testing Number of Success
2. Testing Number of Proportion
3. Testing Difference of Proportion
1. Chi-Square Test
2. Sign Test
3. Median Test
4. Rank Test
5. Rank Sum Test (Mann & Whitney U-Test)
Variables
1. Small Sample (n < 30) : t-Test
2. Large Sample (n > 30) : Z-Test, F-Test, ANOVA

Advantages of Nonparametric Tests

The main advantages of uses of nonparametric tests are:

1. These are distribution-free tests and are based on very few assumptions.
2. These tests are very simple and can be applied with ease.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_13

3. Both the qualitative and quantitative aspects of data are considered for applications of these tests.
4. Parametric tests have a traditional approach, whereas nonparametric tests have a modern approach.
5. Applications of these tests are very useful in medical research.

13.2 Sign Test

The "sign test" makes use of differences in "+" (plus) or "–" (minus) form rather than numeric differences. This test would be applicable for comparing the effect of a treatment on a single group or two treatments on matched pairs in two groups of patients or experimental animals.

Types of Sign Test

1. One-sample sign test (for only one group of subjects)
2. Two-sample sign test (for paired groups of subjects)

One-Sample Sign Test

Steps

1. Plus (+) or minus (−) sign is assigned to the item (observation) with reference to the given or observed value of the median.
2. A number of zero (0) signs are not given any consideration. A number of "minus signs" are termed as "S." The null hypothesis is considered as $H_o = P = \frac{1}{2}$ (0.5 or 50% probability in either way).
3. The critical value (K) is determined by the formula given below:

$$K = \frac{N-1}{2} - 0.98\sqrt{N}$$

4. H_o will be accepted if $S > K$.

Example 1 A cricket coach claims that the players trained by him would definitely make 40 runs in a test match before getting out. We contacted ten batsmen trained by him and recorded their scores in the first test match played by them. The scores of 10 players are 42, 44, 50, 55, 35, 32, 48, 54, 40, and 38, respectively. Test the hypothesis to accept or reject the claim of the coach.

Solution Computations have been provided in Table 13.1.

$$K = \frac{N-1}{2} - 0.98\sqrt{N} = \frac{9-1}{2} - 0.98\sqrt{9} = \frac{8}{2} - 0.98 \times 3 = 4 - 2.94 = 1.06$$

Table 13.1 Sign assigned to scores of players

Players	Runs (x)	Sign (x−40)	Computation
1	42	+	Plus sign = 6
2	44	+	Minus sign = 3
3	50	+	Zero sign = 1
4	55	+	N = sum of (+) + (−)
5	35	−	N = 6 + 3 = 9
6	32	−	
7	48	+	
8	54	+	
9	40	0	
10	38	−	
S		3	

Table 13.2 Sign assigned to weight loss within 3 months

Players	Runs (x)	Sign (x−40)	Computation
1	22	+	Plus sign = 8
2	24	+	Minus sign = 1
3	30	+	Zero sign = 1
4	25	+	N = sum of (+) + (−)
5	25	+	N = 8 + 1 = 9
6	22	+	
7	24	+	
8	24	+	
9	20	0	
10	18	−	
S		1	

Decision

Since $S(3)$ is more than $K(1.06)$, H_o is accepted. Hence, the claim of the cricket coach is not acceptable.

Example 2 A nutritionist claims that any obese person taking diet advised by him would definitely lose 20 kg weight within 3 months. A random sample of 10 persons on his diet reported the weight loss as 22, 24, 30, 25, 25, 22, 24, 24, 20, and 18, respectively. Perform the "one-sample sign test" to accept or reject the claim of the nutritionist.

Solution Computations have been provided in Table 13.2.

$$K = \frac{N-1}{2} - 0.98\sqrt{\mathrm{N}} = \frac{9-1}{2} - 0.98\sqrt{9} = \frac{8}{2} - 0.98 \times 3 = 4 - 2.94 = 1.06$$

Decision

Since $S(1)$ is less than $K(1.06)$, H_o is rejected. Hence, the claim of the nutritionist is acceptable.

Two-Sample Sign Test

Pair sign test is applied when data is continuous and groups contain matched pairs. In "pair sign test," we compare two populations (usual versus special). Plus (+) or minus (−) sign is assigned after subtracting usual from the special ($A-B$). The H_o is always that $P = 0.5$ (50% probability).

Example 3 Two groups of age- and weight-matched inbred rats were classified as "group A" and "group B." Group A rats ($n = 20$) were fed on special diet, and group B rats ($n = 20$) were fed on a usual diet for 1 month. After 1 month rats were weighed and weights recorded in g. Test the hypothesis that the special diet is more effective for weight gain than the usual diet (alternate hypothesis).

Solution Computations have been provided in Table 13.3.

Table 13.3 Sign assigned to weight gain after 1 month

Players	Special diet (A)	Usual diet (B)	Sign ($A-B$)	Computation
1	51	48	+	Plus sign = 14
2	45	40	+	Minus sign = 2
3	44	48	−	Zero sign = 4
4	65	60	+	N = sum of (+) + (−)
5	70	62	+	$N = 14 + 2 = 16$
6	50	50	0	
7	45	45	0	
8	60	55	+	
9	58	53	+	
10	51	48	+	
11	48	45	+	
12	45	42	+	
13	45	41	+	
14	54	48	+	
15	51	46	+	
16	48	43	+	
17	51	48	+	
18	48	48	0	
19	46	46	0	
20	54	56	−	
S			**2**	

$$K = \frac{N-1}{2} - 0.98\sqrt{N} = \frac{16-1}{2} - 0.98\sqrt{16}$$
$$= \frac{15}{2} - 0.98 \times 2 = 7.5 - 3.92 = 3.58.$$

Decision
Since $S(2)$ is less than $K(3.58)$, H_o is rejected. Hence, the special diet is effective for gaining weight.

Large Sample and Sign Test
A sample is considered to be large when items are more than 25. We can apply the sign test by the method of normal approximation. Instead of critical value "*K*," the value of "*Z*" is computed by the following formula:

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{\text{difference}}{\text{SE}\bar{x}}$$

Example 4
In a clinical laboratory, random blood sugar (RBS) level of 10 to 20 patients is done daily. The daily data of the number of RBS investigations in a month of 30 days have been given in tabulated form in Table 13.4. Use the "sign test" to test the hypothesis that 15 RBS investigations are done daily on an average.

Solution

$$\begin{aligned} \text{Plus signs} &= 16(X) \\ \text{Minus signs} &= 11 \\ \text{Zero} &= 3 \\ \text{Probability of plus signs as per } H_o &= \frac{1}{2} \times 30 = 15(np) \end{aligned}$$

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{16 - 15}{\sqrt{15 \times 0.5}} = \frac{1}{\sqrt{7.5}} = \frac{1}{2.73} = 0.366$$

Standard value of Z at 95% confidence limit (5% level of significance) = 1.96

Decision
The computed value of Z (0.36) is less than the standard value (1.96) at 5% level of significance. The null hypothesis is accepted. Hence, the average daily load of RBS investigations could be taken as 15.

Table 13.4 Data of random blood sugar tests daily in 30 days

Day	RBS tests (*Rn*)	Sign (*Rn*−15)	
1	12	−	Plus signs = 16
2	13	−	
3	11	−	Minus signs = 11
4	10	−	
5	15	0	Zero = 3
6	16	+	
7	18	+	
8	17	+	
9	20	+	
10	19	+	
11	15	0	
12	14	−	
13	13	−	
14	12	−	
15	11	−	
16	10	−	
17	13	−	
18	14	−	
19	15	0	
20	18	+	
21	20	+	
22	21	+	
23	20	+	
24	18	+	
25	19	+	
26	22	+	
27	18	+	
28	17	+	
29	16	+	
30	18	+	

Example 5 The following data shows the number of patients examined in the "Out Patient Department" (OPD) of a hospital in a month of 30 days. Use the "sign test" to test the "null hypothesis" that on an average 80 patients are examined in the OPD. The data has been arranged in the tabulated form in Table 13.5.

Solution

As per null hypothesis:

The probability of plus signs $(p) = \frac{1}{2}$

The probability of minus signs $(q) = \frac{1}{2}$

Table 13.5 Data of the number of patients examined in a month

Date	Patients examined	Sign (x−80)	
1	60	−	Plus signs (X) = 17
2	72	−	Minus signs = 12
3	65	−	Zero = 1
4	80	0	Sample size (N) = 30
5	90	+	
6	82	+	
7	71	−	
8	100	+	
9	105	+	
10	110	+	
11	64	−	
12	75	−	
13	73	−	
14	85	+	
15	95	+	
16	68	−	
17	89	+	
18	88	+	
19	100	+	
20	105	+	
21	110	+	
22	64	−	
23	75	−	
24	73	−	
25	85	+	
26	95	+	
27	68	−	
28	89	+	
29	88	+	
30	100	+	

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{17 - 15}{\sqrt{30 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{2}{\sqrt{7.5}} = \frac{2}{2.738} = 0.73$$

Standard value of Z at 5% level of significance = 1.96

Decision

The calculated value of Z (0.73) is less than the standard value 1.96 at 5% level of significance. So, the difference is not significant. The null hypothesis is accepted. So, on an average 80 patients are examined daily in the OPD of the said hospital.

13.3 Median Test

Median ($\tilde{x}$) is generally considered as the most appropriate method of average. So, the "median test" is also considered as an important "nonparametric test." It is used to test whether the two or more samples have been taken from the population with a common median ($\tilde{x}$). The pair sign test was used to compare samples from the same population. The median test has been developed to test the different populations (two or samples) which may or may not have the same median.

Small Sample Median Test

For two samples we use 2 × 2 matrix table. If the H_o is true, then sample observations (items) would be more than the median value, and half would be below the median value. Values are plotted in the 2 × 2 matrix table for application of the "median test." Find below the 2 × 2 matrix table for the median test.

2 × 2 matrix table

Number of scores	Sample I	Sample II	Total
Above median	a	b	$a + b = n_3$
Below median	c	d	$c + d = n_4$
Total	$a + c = n_1$	$b + d = n_2$	$a + b + c + d = N$

$$\text{Probability} = P = \frac{n_{1_{C_a}} + n_{2_{C_b}}}{N_{C_{a+b.}}}$$

Note: Factorials are used in calculations

Example 6 Blood urea levels in two groups (samples) of seven adult volunteers each. The data has been tabulated below in Table 13.6. Apply the median test to ascertain that the volunteers come from the same normal population.

Solution

Let us arrange the items in ascending order to find out the median ($\tilde{x}$) of both the samples together.

21, 24, 25, 28, 28, 29, 29, 30, 32, 35, 36, 38, 38, 39

Median ($\tilde{x}$) = size of $\frac{(n+1)^{th}}{2}$ item = 7.5th item = $\frac{29+30}{2} = 29.5$

Table 13.6 Blood urea levels in mg/dl in two groups of volunteers

ID	Group I	Group II
1	28	25
2	29	24
3	36	28
4	32	35
5	21	39
6	38	29
7	38	30

Now let us evaluate and incorporate the values in 2 × 2 matrix table.

2 × 2 matrix table

Number of scores	Sample I	Sample II	Total
Above median	3 (*a*)	4 (*b*)	3 + 4 = 7
Below median	4 (*c*)	3 (*d*)	4 + 3 = 7
Total	3 + 4 = 7	4 + 3 = 7	3 + 4 + 4 + 3 = 14

$$\text{Probability} = P = \frac{\mathrm{n1}_{C_a} + \mathrm{n2}_{C_b}}{\mathrm{N}_{C_{a+b}}}$$

$$= P = \frac{7_{C_3} + 7_{C_4}}{14_{C_7}} = \frac{\frac{7 \times 6 \times 5}{3 \times 2} \times \frac{7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1}}{\frac{14 \times 13 \times 12 \times 11 \times 10 \times 9 \times 8}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}}$$

$$= \frac{1225}{3432} = 0.357$$

Decision

The probability, $P = 0.357$, computed above is very much greater than 0.05 level of significance. Hence, we accept the "null hypothesis" (H_o). So, both the samples belong to the same population.

Large Sample Median Test

When the sample has more than 20 items, it is considered as a large sample. Then we apply "chi square formula" with Yates' correction to test the statistical significance.

Chi square formula with Yates' correction:

$$\chi^2 = \frac{N\left[\left|ad - bc\right| - \frac{N}{2}\right]^2}{n1 \cdot n2 \cdot n3 \cdot n4}$$

Example 7 Fasting blood sugar (FBS) levels determined in a group of 25 normal volunteers exhibited FBS levels in 15 volunteers above the median value, whereas another group of 30 volunteers showed FBS levels in 10 volunteers above the median value.

Solution

Let us incorporate the above data in 2 × 2 matrix table.

Fasting blood sugar levels in two groups of volunteers

Number of scores	Group I	Group II	Total
Above median	15 (*a*)	10 (*b*)	$a + b = n_3 = 25$
Below median	10 (*c*)	20 (*d*)	$c + d = n_4 = 30$
Total	$a + c = n_1 = 25$	$b + d = n_2 = 30$	$N = 55$

$$\chi^2 = \frac{N\left[\left|ad - bc\right| - \frac{N}{2}\right]^2}{n1 \cdot n2 \cdot n3 \cdot n4} = \frac{55\left[\left|15 \times 20 - 10 \times 10\right| - \frac{55}{2}\right]^2}{25 \times 30 \times 25 \times 30}$$

$$= \frac{55\left[\left|300 - 100\right| - 27.5\right]^2}{562500} = \frac{55 \times 29756.25}{562500} = \frac{1636593.75}{562500} = 2.90$$

Degrees of freedom $(V) = (c-1)(r-1) = (2-1)(2-1) = 1 \times 1 = 1$
$\boldsymbol{\chi^2_{0.05}}\ (V_1) = 3.84$

Decision

The computed value $\chi^2_{\mathbf{0.05}} = 2.90$ is less than the table value at 5% level of significance, so the difference is not significant. Hence both the groups (samples) come from the same normal populations and can be pooled up as a single population sample.

13.4 Rank Correlation Test (Spearman's Rank Correlation Test)

Spearman's "rank correlation" is a coefficient used to test an association between the two variables of a sample. The association between the variables may be positive or negative. The coefficient can be tested statistically at 5% or 1% level of significance. The value of rank correlation would always vary between −1 and + 1. The formula for "rank correlation test" is:

$$r_s = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

D = difference of ranks

The "standard error" (SE_r) of "coefficient of correlation" can be computed with the formula given below:

$$\mathrm{SE}_r = \frac{1}{\sqrt{N - 1}}$$

13.5 Rank-Sum Test (Mann-Whitney U-Test)

The U-test is a very strong "nonparametric test" which only considers a larger pair. It provides a way to test if the set of scores has come from the same population. Rank-sum test compares two independent samples that have been treated differently, to ascertain if the two samples are different due to treatment. The data must be distributed continuously for the "U-test" as is the requirement of "sign test" also. *Mann-Whitney U-test* is also known as *Wilcoxon rank-sum test*.

The formula for Mann-Whitney U-test is:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 \text{ and}$$
$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - R_2$$

where:

N_1 is the number of individuals in group A
N_2 is the number of individuals in group B
R_1 is the sum of ranks in group A
R_2 is the sum of ranks in group B

Ranks are assigned to the collective data of both the groups in ascending order.

When ranking is distributed in such a way that one group has all the "lower ranks" and the other group has all the "upper ranks," the value of U would most likely be zero. The value of "U" which is used for the U-test is the smaller one of the U_1 and U_2. The U-table can be seen for m and n.

m = number of items in the large group
n = number of items in the small group

Important Fact If the computed value of "U" is smaller than the table value of U, there must be a significant difference of the two groups. If the table value is not available, we can use transformation of Z for decision.

$$\frac{Z = U - N_1 \times \frac{N_2}{2}}{\sqrt{N_1 \times N_2 \times (N_1 + N_2 - 1)}}$$

Example 8 Two groups of rats of the same age were taken and their weights recorded. Group A ($N_1 = 5$) rats were fed on normal diet, and Group B ($N_2 = 6$) rats were fed on special diet. After 1 month, rats were weighed again, and the weight gained was calculated. Apply Mann-Whitney test (rank-sum test) to test the alternate hypothesis that special diet promotes weight gain.

Solution

The weight gain data of both the groups of rats have been arranged in Table 13.7, and ranks have been assigned in the ascending order of items (weight gain in grams).

Table 13.7 Weight gain in two groups of rats with normal or special diet

	Weight gain in grams and ranks assigned					
Status	I	II	III	IV	V	VI
Group A	5	7	9	8	6	
Rank	1	3	5	4	2	
Group B	13	15	11	12	16	10
Rank	9	10	7	8	11	6

$$\begin{aligned}
N_1 &= 5;\ R_1 = 1+3+5+4+2 = 15\\
N_2 &= 6;\ R_2 = 9+10+7+8+11+6 = 51\\
U_1 &= N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = 5\times 6 + \frac{5(5+1)}{2} - 15\\
&= 30 + 15 - 15 = 30\\
U_2 &= N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = 5\times 6 + \frac{6(6+1)}{2} - 51\\
&= 30 + 21 - 51 = 0\\
U_{0.05} &= 5\ (\text{at}\ \ m = 6\ \text{ and }\ n = 5)\\
U_2 &= 0\ \left(\text{is the smaller value of``}U\text{'' in these two groups}\right)
\end{aligned}$$

Decision

The computed value of $U_2 = 0$ is smaller than the "table value" of $U_{0.05} = 5$. So, the null hypothesis (H_o) is rejected. Hence, there is a significant difference in weight gain with special diet.

Example 9 An IQ testing of random samples of 20 boys and 20 girls was done by the HR department of a company. The scores obtained by them out of 100 are given below in tabulated form in Table 13.8. Apply the "Mann-Whitney *U*-test" to determine if there is a significant difference in the average IQ of boys and girls.

$$\begin{aligned}
N_1 &= 20; R_1 = 326\\
N_2 &= 20; R_2 = 305\\
U_1 &= N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = 20\times 20 + \frac{20(20+1)}{2} - 326\\
&= 400 + 210 - 326 = 284\\
U_2 &= N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = 20\times 20 + \frac{20(20+1)}{2} - 305\\
&= 400 + 210 - 305 = 305\\
U_{0.05} &= 5\ (\text{at}\ \ m = 20\ \text{ and }\ n = 20)\\
U_1 &= 284\ \left(\text{is the smaller value of``}U\text{'' in these two groups}\right)
\end{aligned}$$

Here: m = 20; n = 20,
Table value of $U_{0.05}$ = 138 (at m=20 and n=20)

Table 13.8 Scores of IQ test of boys and girls

ID	Boys' score	Rank (R_1)	Girls' score	Rank
1	56	10	70	17
2	54	9	74	19
3	60	12	62	13
4	58	11	66	15
5	64	14	82	23
6	72	18	88	26
7	56	10	76	20
8	92	28	32	4
9	94	29	84	24
10	90	27	92	28
11	90	27	26	3
12	78	21	42	5
13	52	8	24	2
14	50	7	20	1
15	56	10	80	22
16	66	15	82	23
17	84	24	86	25
18	84	24	84	24
19	46	6	52	8
20	68	16	26	3
Rank sum		**326**		**305**

Decision

The computed value of $U_1 = 284$ is greater than the "table value" of $U_{0.05} = 138$. So, the null hypothesis (H_o) is rejected. Hence, there is a significant difference in the average IQ of boys and girls.

13.6 Coefficient of Variation

Definition The "coefficient of variation" (CV) is defined as the ratio of "standard deviation" (s) to the "mean"($\bar{x}$) expressed as percentage proportion. It is a relative measure of dispersion.

$$\mathrm{CV} = \frac{s}{\bar{x}} \times 100$$

Applications It is applied to the mean ($\bar{x}$) and "standard deviation" (σ or s) for comparing two methods, events, players, or workers to ascertain the better of the two.

Example 1 Mr. Rohit and Mr. Mohit are two cricket players. After five matches Mr. Rohit had an average of 36.0 runs with a “standard deviation” (s) of 17.5 runs, and Mr. Mohit had an average of 48.6 runs with a “standard deviation” of 15.5 runs. In your opinion who is the better player?

Solution

Mr. Rohit	Mr. Mohit
Mean ($\bar{x}$) = 36.0 runs	Mean ($\bar{x}$) = 48.6 runs
s = 17.5 runs	s = 15.5 runs
CV $= \frac{s}{\bar{x}} \times 100$	CV $= \frac{s}{\bar{x}} \times 100$
$= \frac{17.5}{36} \times 100$	$= \frac{15.5}{48.6} \times 100$
48.61	**31.89**

Decision

After computations we find that the “coefficient of variation” (CV) of Mr. Mohit (31.89) is less than of Mr. Rohit (48.61). This shows that Mr. Mohit is a better player.

Example 2 Rayat (1985) discovered a method of preserving “cellulose acetate filters” with chemotactic cells after scoring “neutrophil chemotaxis” in μm, as the results of neutrophil chemotaxis assay. A set of filters of ten cases stored by Rayat’s method did not show any deviation after 1 month of storage, whereas readings from filters stored by “conventional method” showed a gross fall in chemotaxis. The chemotactic scores after storage by Rayat’s method were 75, 76, 78, 75, 72, 71, 73, 75, 69, and 70 μm, whereas the set of Millipore filters with chemotactic cell from the same control subjects had the scores as 35, 36, 37, 27, 31, 30, 35, 31, 29, and 30 μm, respectively. Testify that Rayat’s method is better than the conventional method.

Solution

The data has been arranged in Table 13.9 and the “coefficient of variation” for both methods calculated therein.

Decision

After computations we find that the coefficient of variation (CV) of “Rayat’s method” (3.916) is less than the “conventional method” (9.803). This shows that Rayat’s method is a better method for preserving Millipore filters with chemotactic cells.

Table 13.9 Chemotaxis scores in μm after preservation

Volunteer ID	Rayat's method	General method
1	75	35
2	76	30
3	78	37
4	75	27
5	72	31
6	71	30
7	73	35
8	75	31
9	69	29
10	70	31
Mean ($\bar{x}$)	73.4	31.6
Standard deviation	2.875	3.098
CV	**3.916**	**9.803**

13.7 Parameters of Validity of a New Test

Whenever a new "biomedical test" like "agglutination assay" (AR), radioimmunoassay (RIA), or enzyme-linked immunosorbent assay (ELISA) is developed for diagnostic application, it needs to be evaluated for validity. The parameters of validity are:

1. *Sensitivity*: The outcome of assay to detect "true positive" as positive
2. *Specificity*: The outcome of assay to detect the "true negative" as negative
3. *Efficiency*: Percentage proportion of "true results" (positive + negative) out of total samples for quality check
4. *Positive predictive value*: Percentage proportion of "true positive" out of "total positive" detected
5. *Negative predictive value*: Percentage proportion of "true negative" out of "total negative" detected

13.8 General Methodology for Validity of an Assay

Example 1 Suppose we have 27 "true positive" blood samples from patients with dengue fever and 27 blood samples from "normal volunteers" for validating a "new ELISA test." If the ELISA test fails to detect "2" cases out of 27 "true positives" as positive and also fails to detect "1" "true negative" case as negative, what would be the outcome of validation parameters for this ELISA test?

Table 13.10 Results of ELISA test

Status	Dengue patients	Normal volunteers	Total
Test positive	25 (a)	1 (b)	26
Test negative	2 (c)	26 (d)	28
Total	**27**	**27**	**54**

Solution

The data has been arranged in Table 13.10 for understanding the formulae.

General Formulae

1. $Sensitivity = \frac{a}{a+c} \times 100$
2. $Specificity = \frac{d}{b+d} \times 100$
3. $Efficiency = \frac{a+d}{a+b+c+d} \times 100$
4. $Positive\ predictive\ value = \frac{a}{a+b} \times 100$
5. $Negative\ predictive\ value = \frac{d}{c+d} \times 100$

Values for the case cited in the example:

1. $Sensitivity = \frac{a}{a+c} \times 100 = \frac{25}{27} \times 100 = 92.6\%$
2. $Specificity = \frac{d}{b+d} \times 100 = \frac{26}{27} \times 100 = 96\%$
3. $Efficiency = \frac{a+d}{a+b+c+d} \times 100 = \frac{25+26}{54} \times 100 = 94\%$
4. $Positive\ predictive\ value = \frac{a}{a+b} \times 100 = \frac{25}{26} \times 100 = 96\%$
5. $Negative\ predictive\ value = \frac{d}{c+d} \times 100 = \frac{26}{28} \times 100 = 93\%$

Statistical Quality Control in Clinical Laboratories

14

14.1 Quality Perceptions

In a biomedical laboratory, *normal range* is the most important "quality perception" for controlling quality. It has been an essential protocol to give "normal range" for each of the constituents of various body fluids. The normal range is also referred as "reference range" by which we mean that the concentration of a constituent of a biological fluid or sample must be within this range for the individuals considered to be in good health. In other words, we would assume that the values outside these limits warrant an alarm for thorough health checkup. There are set procedure for establishing such ranges. Modern medicine warrants the thorough investigations of biological material derived from the body of a patient to ascertain the cause and effect of disease before the commencement of medical or surgical treatment. Quality management in medical laboratories is a must for accreditation [24].

14.1.1 Normal Distribution Curve

Let us deliberate on an example from industry to understand the concept of "normal distribution" and "normal distribution curve." Suppose a "mineral water bottling plant" is packing water with a standard label of 200 ml per unit and states that there can be 5% natural variation in volume. In this case "normal range" would be 190–210 ml. If we collect a large sample of more than 30 bottles from various batches of mineral water, measure the volume of water in each bottle of the sample, and plot a scatter graph; that graph would be a "normal curve" if the distribution was normal. Such a curve is defined by "mean" ($\bar{x}$) and the "standard deviation" (s) and has been depicted in Fig. 14.1. Mean ($\bar{x}$), the arithmetic average, is determined by dividing the sum of determinations with number of determinations:

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_14

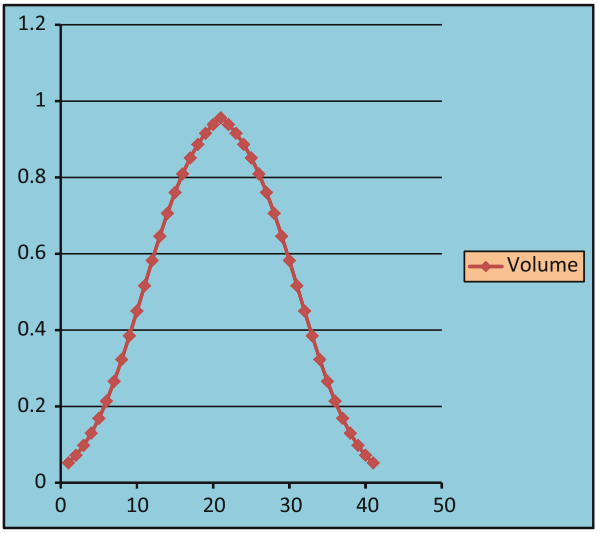


Fig. 14.1 Normal distribution curve for volume of mineral water

$$\bar{X} = \frac{\Sigma x}{n}$$

Standard deviation (sd, s, or σ) is computed by the following formula:

$$s = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n}}$$

where:

n = number of determinations
$\bar{x}$ = Mean
x = each determination or observation

Note Abbreviation used for "standard deviation" could be SD, sd, or s when we use it for "standard deviation" of a sample from population and $\boldsymbol{\sigma}$ for "standard deviation" of population.

We have already learnt in Chapter 10 that for such a curve, 68.3% of values fall within$\bar{x}$ ± 1sd, 95.4% values within$\bar{x}$ ± 2sd, and 99.7% values fall within$\bar{x}$ ± 3sd. It is a universal practice to consider the "normal range" as $\bar{x}$ ± 2sd. With the "normal range" fixed as $\bar{x}$ ± 2sd, it would be obvious that values from 5% of the normal persons would lie outside the range. With a normal range set as $\bar{x}$ ± 3sd, only 0.3% values would lie outside the range; that means only 3 in 1000 persons would fall outside the range.

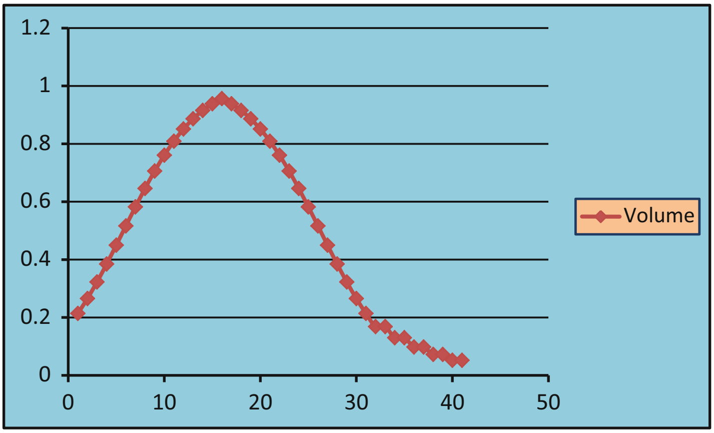


Fig. 14.2 Skewed distribution curve for volume of mineral water

It has been observed that sometimes majority of values at the lower end are close to zero. Under such conditions we get a skewed curve as shown in Fig. 14.2. It is necessary that the working "normal range" should include majority of normal values and very few abnormal values. The $\bar{x} \pm 2\text{sd}$ is the most satisfactory normal range/ reference range as considered by the "International Organization for Standardization" (ISO).

The "International Organization for Standardization" (ISO) is an "International Authority" for setting up standard guidelines for various organizations and laboratories. The International Organization for Standardization is based in Geneva, Switzerland, and 163 countries are its members. The international standard for medical laboratories was first published in the year 2005. It was based on ISO 15189:2003(1st edition) and later revised as ISO 15189:2007 (2nd edition). Further revision was done in the year 2010, and the current version is ISO 15189:2012 (3rd edition). The ISO 15189:2012 provides standard guidelines for "requirements for quality and competence in medical laboratories." The main goal of ISO 15189:2012 is the "Global Harmonization of Medical Laboratories" in terms of quality of their services. The need of the hour is the "total quality management" (TQM) as per ISO 15189:2012. The "statistical quality control" (SQC) is of utmost importance for competence.

14.1.2 Errors

Standard operating procedures (SOPs) should be prepared by all the clinical laboratories and followed to avoid operational and technical errors. Errors may be divided into two groups:

1. Errors due to faulty methods used: These may be inherent due to methodology or due to incorrect composition of reagents used or due to faulty apparatus. Such defects would lead to incorrect determinations even by experienced staff. Faulty apparatus and reagents would affect the determinations of all the batches.
2. Errors due to faulty performance: An inaccurate result may be obtained for a single determination due to wrongly performed step of the procedure. In this case other determinations in the batch could be considered reliable.

14.1.3 Check over Errors

Errors can be kept under control by use of known "standard solutions" or blood samples of known concentration of constituents. A wide range of pooled sera and some urine samples are now available from numerous makers and suppliers for most of the constituents now measured. Various scientific groups/committees have been formed all over the world for accreditation and licensing of medical laboratories for quality services.

Internal and external audit of performance of such checks have been made mandatory for ISO Certifications and Accreditation. Checks are introduced at frequent intervals, and the persons performing the test are kept blind about the predetermined value of the constituent being determined. King and Woottan (1956) were the first to introduce such checks with a batch of each determination. We need to draw charts for each determination, and on these charts daily value of known standard is plotted in order to know the extent of variations from the standard value. When these determinations are outside the minimum permissible range, results are declared invalid, and alarm is raised for repeating the batches.

14.1.4 Permitted Variations

Permitted variations for the inorganic and organic constituents of the blood have been decided by quality control bodies/committees. For inorganic constituents such as sodium (Na^+), potassium (K^+), and chloride (Cl^-), permitted variations are ±3%, whereas for organic constituents such as glucose, urea, uric acid, cholesterol, bilirubin, and albumin, permissible variations could range from ±5% to ±10%.

When determinations are made in batches, we may omit control occasionally in case of nonavailability of control sera or standard. In batches we find many results within "normal range." All abnormal results in a batch would lead to suspicion of something wrong with the procedure or the equipment used. When some investigations are done rarely or in very small batches, we should include normal samples in duplicate along with standard control. When spectrophotometric or colorimetric methods are used, it is advisable to avoid delay in taking optical density measurements.

The indication of a possible change in operation of a technique can be obtained from daily "mean" of successive batches by plotting charts. At present we run large batches on “autoanalyzers” for substances such as glucose, urea, uric acid, sodium, potassium, and chloride; the mean remains remarkably constant from day to day. For substances like calcium, alkaline phosphatase, and transaminases, determinations are done in small batches daily; weakly means could be taken.

14.1.5 Accuracy and Precision

We must understand the meaning of these two terms in a clinical laboratory. The term “accuracy” reflects how close is the “mean” ($\bar{x}$) of large number of determinations to the actual amount of the substance present in the test specimens or standards.

The “precision” refers to the extent to which the repeated determinations on an individual specimen or “standard control” by applying a certain technique are consistent. This reflects the range of error of the method used and may vary from technologist to technologist. We understand from the facts that a method with high degree of precision may not be very accurate. The precision refers to the “variable error,” whereas the accuracy refers to the “constant error” inherent in a method. The degree of precision can be obtained by carrying out more than 30 determinations in duplicates and taking difference of each set and calculating “standard deviation” (s) by the following formula:

$$s = \sqrt{\frac{\Sigma d^2}{n}}$$

where $d =$ difference between each pair of duplicates.

$n =$ the number of duplicates.

The limits of precision are generally taken within 95% limits, as worked out by $\bar{x} \pm 2s$, where $\bar{x}$ is the mean of the set of determinations and s is the “standard deviation” as calculated above for at least 30 sets of duplicates. Alternatively, “standard deviation” can be calculated by multiplying the “mean” difference between pairs by 0.88.

$$s = \bar{d}(0.88)$$

14.2 Quality Control Measures

We can take the following measures for control of quality:

1. Use of quality control materials
2. Use of control charts for SQC

3. Participation in interlaboratory comparison programs
4. Analysis of quality control data

14.2.1 Use of Quality Control Materials

The laboratory should use quality control materials that react to the testing procedures in a manner as close as possible to patients' samples. These materials should be periodically depending on the stability of standard operating procedures (SOPs) to avoid risk to the patient from erroneous results.

As stated earlier, quality control standards are available commercially. However, if these are not available, pooled sera from one's own laboratory can be used as standard. Quality control manager should collect surplus sera, urine, and fluids daily and store in a deep freezer and keep on adding these until sufficient volume has accumulated. These should be thawed, Millipore filtered, and analyzed to determine the concentration of constituents, divided in aliquots, and stored in deep freezer for daily use as standards.

14.2.2 Control Charts

The "control charts" are the graphic tools developed for detecting unnatural pattern of variation in laboratory investigations or production process (in industry) and determining the permissible limits of variations. The permissible limits are called "upper control limit" (UCL) and "lower control limit" (LCL).

14.2.3 Advantages of Control Charts

There are three major advantages of control charts:

1. Control charts define the goals to be achieved.
2. These act as tools to attain the determined goals.
3. These enable us to take decision to accept or reject the batch of determinations or items.

14.2.4 The Mean Chart ($\bar{x}$ Chart)

The mean chart is used to express the quality average of given set of determinations or samples drawn from a given process. Make the following calculations:

1. Take mean of various samples (say mean of a batch of fasting blood sugar determinations): $\bar{x} = \frac{\Sigma x}{n}$.

2. Take “grand mean” ($\bar{\bar{x}}$), that is, mean of means all batches (samples) done in a week or a month: $\bar{\bar{x}} = \frac{\Sigma \bar{x}}{n}$.
3. Take “range” of various samples item wise: (range = largest item – smallest item)
4. Take mean of Range: $\bar{R} = \frac{\Sigma R}{\text{No. of Batches of Samples}}$
5. Set up control limits: $\bar{\bar{x}} \pm 2\sigma$
6. We can also set up limits as
   Upper Control Limit (UCL) = $\bar{\bar{x}} + A\bar{R}$
   Lower Control Limit (LCL) = $\bar{\bar{x}} - A\bar{R}$
7. The value of A is calculated as
   $A = \frac{3}{\sqrt{n}}$ (when $n > 25$)

Example 1 Data of the month of January for “blood urea” determinations carried out at emergency laboratory of a hospital has been given as daily mean for the batches as 28, 27, 25, 29, 27, 29, 30, 25, 26, 28, 29, 31, 27, 28, 29, 25, 29, 27, 29, 30, 25, 26, 28, 29, 31, 30, 25, 26, 28, and29 mg/dl. Prepare a “mean chart” for “blood urea” quality control, and discuss its utility.

Solution

Arrange all the daily means as in Table 14.1, and calculate grand mean ($\bar{\bar{x}}$) and “standard deviation” to set the limits of mean chart for “blood urea” quality control.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{99}{30}} = \sqrt{3.3}$$
$$= 1.82 = 2(\text{Rounded off to whole number})$$

$$\textbf{UCL} = \bar{\bar{x}} + 2\sigma = 28 + 2 \times 2 = 28 + 4 = 32$$
$$\textbf{LCL} = \bar{\bar{x}} - 2\sigma = 28 - 2 \times 2 = 28 - 4 = 24$$

From the above calculations, the “control limits” for “blood urea” determinations come out to be 24 to 32 mg/dl. The “mean chart” for the above data has been exhibited as Fig. 14.3.

Comments

All the daily means of ‘blood urea determinations” provided (ranging from 25 to 31 mg/dl) are within “control limits.” Hence, the quality is within “control,” and there is no sign for alert.

14.2.5 The Cumulative Sum Chart (CUSUM Chart)

The “cumulative sum chart” was introduced by Woodward and Goldsmith in 1964. It is just a variant of “mean chart.” A mean value as close as possible to the actual mean of the daily means ($\bar{\bar{x}}$) is chosen, and every day the difference of day’s mean

Table 14.1 Daily means of blood urea determinations in a month

Day ID	Daily mean ($\bar{x}$)	$\bar{x} - \bar{\bar{x}}$	$\left(\bar{x} - \bar{\bar{x}}\right)^2$
1	28	0	0
2	27	−1	1
3	25	−3	9
4	29	1	1
5	27	−1	1
6	29	1	1
7	30	−2	4
8	25	−3	9
9	26	−2	4
10	28	0	0
11	29	1	1
12	31	3	9
13	27	−1	1
14	28	0	0
15	29	1	1
16	25	−3	9
17	29	1	1
18	27	−1	1
19	29	1	1
20	30	2	4
21	25	−3	9
22	26	−2	4
23	28	0	0
24	29	1	1
25	31	3	9
26	30	2	4
27	25	−3	9
28	26	−2	4
29	28	0	0
30	29	1	1
Description	$\bar{\bar{x}} = 28$		Sum = 99

from this is calculated and added to the sum of all previous differences. The result is plotted on a chart. The direction of this graph line depends on the mean chosen. The grid line of mean value chosen as grand mean ($\bar{\bar{x}}$) should pass through the graph as astride. Any upward or downward deflection of the graph of CUSUM would be a warning sign of nonconformance of quality. The "CUSUM chart" would be like "mean chart" if values are within the "control limits."

Suppose the mean value selected was grand mean ($\bar{\bar{x}}$) of previous month, that is, 28 mg/dl. The differences of daily means from this "proposed grand mean" were worked out, and CUSUM with reference to this was computed during the next month

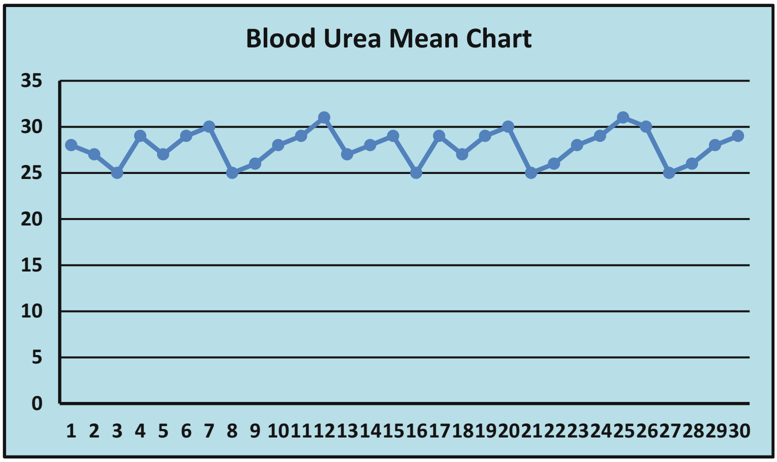


Fig. 14.3 The mean chart for blood urea batches within a month

for the batches of blood urea determinations as depicted in Table 14.2. The "CUSUM chart" for this data has been shown in Fig. 14.4.

$$\begin{aligned} \textbf{UCL} &= \text{Cusum Mean} + 2\sigma = 29 + 2 \times 2 = 29 + 4 = 33 \\ \textbf{LCL} &= \text{Cusum Mean} - 2\sigma = 29 - 2 \times 2 = 29 - 4 = 25 \end{aligned}$$

Comments

Cumulative sum values of "blood urea determinations" for a month are around the grand mean of previous month and conform to the control limits (25–33 mg/dl). Hence, the quality is within "control," and there is no sign for an alert. In the case of nonconformity, corrective actions are taken after identification of the cause of variation.

14.2.6 Participation in Interlaboratory Comparison Programs

Clinical laboratories should participate in the interlaboratory comparison programs such as "external quality assessment" (EQAS) program or proficiency testing programs. Each laboratory should monitor the results of interlaboratory comparison program(s) and implement the corrective actions when the determined criteria are not fulfilled.

Every laboratory should draft documented procedure for participation in interlaboratory comparison. The document should include defined responsibilities and instructions for participation. The interlaboratory comparison program chosen by the laboratory should provide clinically relevant challenges that mimic the

Table 14.2 CUSUM of blood urea determinations in a month

Day ID	Daily mean ($\bar{x}$)	$\bar{x} - \bar{\bar{x}}$	CUSUM
1	28	0	28
2	27	−1	27
3	29	1	28
4	29	1	29
5	27	-1	28
6	27	-1	27
7	26	−2	25
8	31	3	28
9	26	−2	26
10	28	0	26
11	27	−1	25
12	31	3	28
13	29	1	29
14	30	2	31
15	27	−1	30
16	28	0	30
17	30	2	32
18	28	0	32
19	29	1	33
20	26	−2	31
21	29	1	32
22	26	−2	30
23	28	0	30
24	29	1	31
25	27	−1	30
26	30	2	32
27	27	−1	31
28	28	0	31
29	29	1	32
30	28	0	32
Description	$\bar{\bar{x}} = 28$		CUSUM mean = 29
			$\sigma = 2$

patients' samples and have the capacity of examining entire examination process, including preexamination procedures as well as post examination procedures.

14.2.7 Analysis of Quality Control Data

Quality control data should be analyzed and reviewed periodically by "quality control agencies" to detect trends in examination performance that may indicate

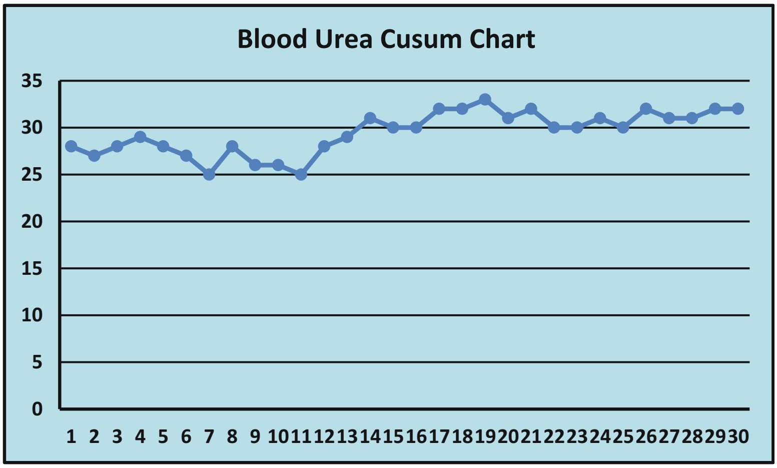


Fig. 14.4 CUSUM chart for blood urea

problems in the examination systems. Corrective actions should be taken in case of nonconformities.

Suppose 50 clinical laboratories participate in "external quality assessment program" for blood urea determination following same SOP. These laboratories would return the results of blood urea value of the external quality control sample to the "external quality control body or agency." The "external quality control body or agency" would find out the "group mean"(GM) and "standard deviation" (SD) of 50 results received and then compute Z distribution for the result of each laboratory (LR) by the formula:

$$Z = \frac{|\,\mathrm{LR} - \mathrm{GM}\,|}{\mathrm{SD}}$$

14.2.8 Decision

$$\begin{array}{ll} Z \leq 2.0 & \text{(OK)} \\ Z > 2.0 & \text{Alert is issued.} \end{array}$$

Suppose GM = 30 mg/dl with SD = 1.9 and the result of your clinical laboratory for the provided standard sample is 28 mg/dl.

$$Z = \frac{|\ 28 - 30\ |}{1.9} = \frac{2}{1.9} = 1.05 \quad \text{(Perfect)}$$

Hence, the “quality control” of your clinical laboratory would be adjudged within permissible limits for blood urea determination.

15 Applications of Microsoft Excel in Statistical Methods

15.1 Excel Worksheet and Finding Mean

The Excel worksheet is the platform where we arrange our data for statistical analysis. During the last two decades, Microsoft Excel has undergone tremendous evolution in terms of its capacity and features. Microsoft Excel worksheets are composed of "rows and columns." Each new workbook has three sheets by default in all the versions of Microsoft Excel except "Microsoft Excel 2016." The creation of additional worksheets in "Microsoft Excel-2016" depends on the CPU memory of a computer. The number of rows and columns in Microsoft Excel worksheets has been exhibited in Table 15.1 for various versions.

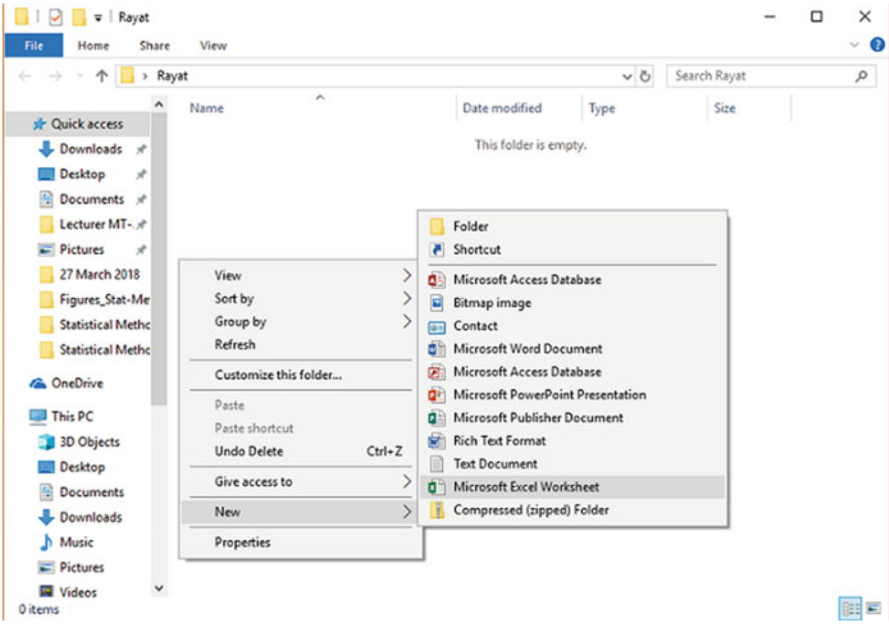
Let us learn to open a new "Microsoft Excel worksheet" document. To open a new "Microsoft Excel worksheet," click the right button of the mouse of your PC in the free space of your FOLDER or desktop, and in the POP-UP menu, go to NEW and in the associate MENU select the "Microsoft Excel worksheet" by clicking the left button of the mouse to start the new document. Rename the "New Microsoft Excel Worksheet.xlsx" folder, and double-click to open the document. The route has been depicted in Fig. 15.1, and a "Blank Excel Worksheet" has been depicted in Fig. 15.2.

The data can be entered in rows or columns in the Excel worksheet, but the data entered in a column for a sample is preferred for calculations and statistical analysis. The "statistical functions" in an Excel worksheet are operational with reference to ranges. Figure 15.3 shows the data typed in cell B3 to B11 and C3 to C11 from a group of nine patients with hypertension. As we desired to find sum of the numeric values of the data in a column ranging from cell B3 to B11 in this Excel worksheet; In the cell B12, the syntax =SUM(B3:B11) was typed, and on pressing Enter/Return key, the sum was entered automatically in the cell B12. For finding the mean of the same data, the syntax =AVERAGE (B3:B11) was typed in the cell B13, and on pressing the Enter/Return key, the average/mean was entered automatically in the cell (B13), as has been exhibited in Fig. 15.3.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7_15

Table 15.1 Capacity of Microsoft Excel worksheets

Version	Rows	Columns	Total cells
Excel 5	16,384	256	4,259,328
Excel 95	16,384	256	4,259,328
Excel 97	65,536	256	16,777,216
Excel 2002	65,536	256	16,777,216
Excel 2002 (XP)	65,536	256	16,777,216
Excel 2003	65,536	256	16,777,216
Excel 2007	1,048,576	16,348	17,142,120,448
Excel 2010	1,048,576	16,348	17,142,120,448
Excel 2010+	1,048,576	16,348	17,142,120,448



This folder is empty.

Fig. 15.1 Opening a new Microsoft Excel worksheet

Note A general point about worksheet functions like SUM( ) or AVERAGE( ) is that we must not leave a space between the name of the function and the opening parenthesis; otherwise Excel will consider it an error.

Relative and Absolute Address

In the Excel worksheet used for the above data, the formula =AVERAGE(B3:B11) was copied and pasted in the cell C13 as = AVERAGE(C3:C11), and the average was returned in the cell C13 on pressing Enter/Return key. The B3:B11 is a relative

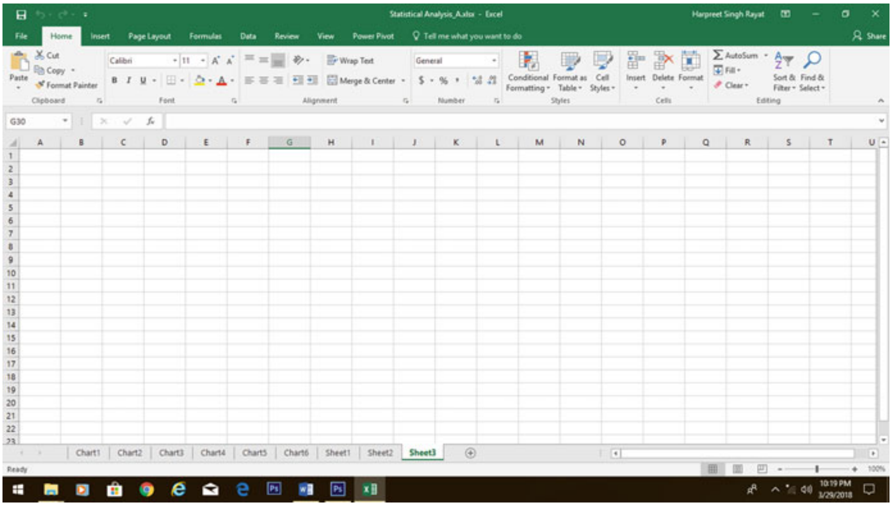
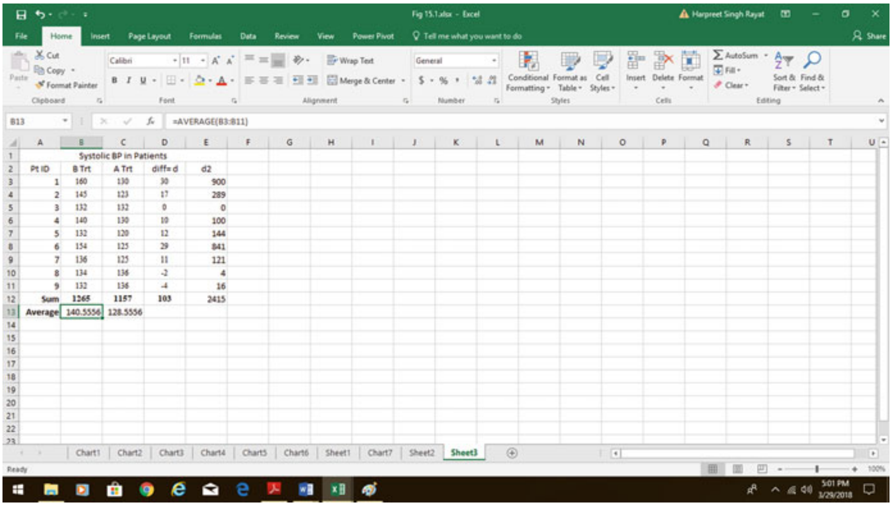


Fig. 15.2 Microsoft Excel worksheet



B13 =AVERAGE(B3:B11)

Pt ID	B Trt	A Trt	diff= d	d2
1	160	130	30	900
2	145	123	17	289
3	132	132	0	0
4	140	130	10	100
5	132	120	12	144
6	154	125	29	841
7	136	125	11	121
8	134	136	-2	4
9	132	136	-4	16
Sum	1265	1157	103	2415
Average	140.5556	128.5556		

Systolic BP in Patients

Fig. 15.3 Screenshot of Excel worksheet displaying the mean of data

address and gives exact result when we copy and paste this in an appropriate cell, whereas $B$3:$B$11 would be termed as an “absolute address,” which when copied to any other cell would refer to the same range of cells and would not give desired output like “relative address.”

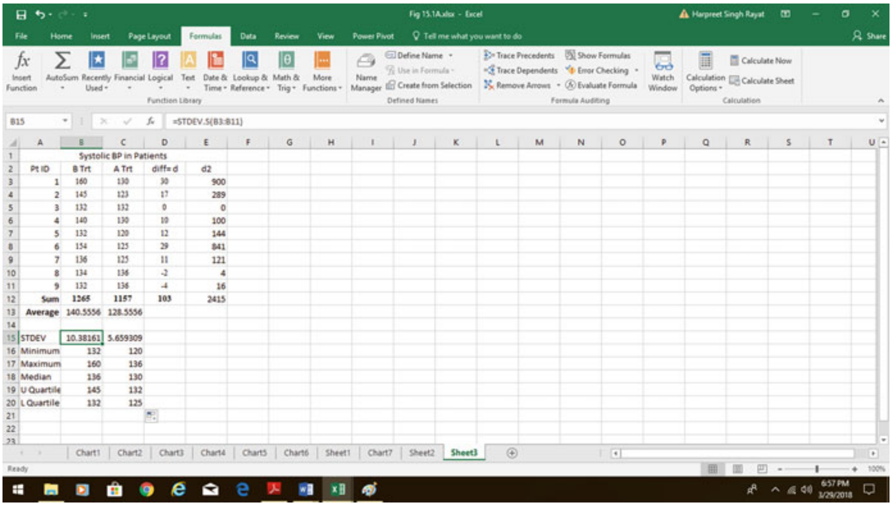
15.2 Excel Worksheet Functions for Statistics

As deliberated above, the AVERAGE( ) is only one function out of the large number of functions provided by Excel which provides assistance in the manipulation of numbers. It is often difficult to remember the precise name of a function and/or the order of its parameters. Excel provides us the function wizard for our convenience. If we click the button on the toolbar labelled ***fx***, this would invoke the function wizard, which starts off by trying to determine which function we need. As we use the program, we would find the functions that we call upon are kept in the function category called "Most Recently Used." But to start with, we will have to select the function from INSERT FUNCTION MENU.

Let Us Practice with the Function Wizard

Do as follows in column A: In cell 15A, type Standard Deviation; in cell 16A, type Minimum; in cell 17A, type Maximum; in cell 18A, type Median; in cell 19A, type Upper Quartile; and in cell 20A, type Lower Quartile.

Do as follows in columns B and C: In cell 15B to 20B as well as in cells 15C to 20C, type the formulae to calculate STDEV, MIN, MAX, MEDIAN, and QUARTILES, respectively. First four we can type the formula directly into the cell without forgetting to begin typing the formula with an EQUAL (=) sign. Or we can use the Function Wizard for guidance. The QUARTILE function would also need an argument for lower and upper limit (one for lower and three for upper). MIN, MAX, and MEDIAN like 4s would need more than one argument. Here we would enter only Group Range and leave all the other arguments as blank. The results of the above exercise have been exhibited in Fig. 15.4.



B15 =STDEV.S(B3:B11)

	A	B	C	D	E
1		Systolic BP in Patients			
2	Pt ID	B Trt	A Trt	diff= d	d2
3	1	160	130	30	900
4	2	145	123	17	289
5	3	132	132	0	0
6	4	140	130	10	100
7	5	132	120	12	144
8	6	154	125	29	841
9	7	136	125	11	121
10	8	134	136	-2	4
11	9	132	136	-4	16
12	Sum	1265	1157	103	2415
13	Average	140.5556	128.5556		
14					
15	STDEV	10.38161	5.659309		
16	Minimum	132	120		
17	Maximum	160	136		
18	Median	136	130		
19	U Quartile	145	132		
20	L Quartile	132	125		

Fig. 15.4 Exhibiting the use of Excel worksheet for multiple functions

Table 15.2 Systolic blood pressure (BP) in nine patients before and after the treatment

Subjects	BP-before treatment (A)	BP-after treatment (B)	Difference $d = A{-}B$	d^2
1.	160	130	+30	900
2.	145	123	+17	289
3.	132	132	0	0
4.	140	130	+10	100
5.	132	120	+12	144
6.	154	125	+29	841
7.	136	125	+11	121
8.	134	136	−2	4
9.	132	136	−4	16

Other Simple Functions of Excel Worksheet

`COUNT(range)`: the number of cells in the range which contain numerical data
`VAR(range)`: the sample variance of the range
`SUM(range`: the sum of the values in the range
`SUMSQ(range)`: the sum of the squares of the values in the range
`DEVSQ(range)`: $(x_i - \bar{x})^2$

Student's *T*-Test Using Microsoft Excel Function

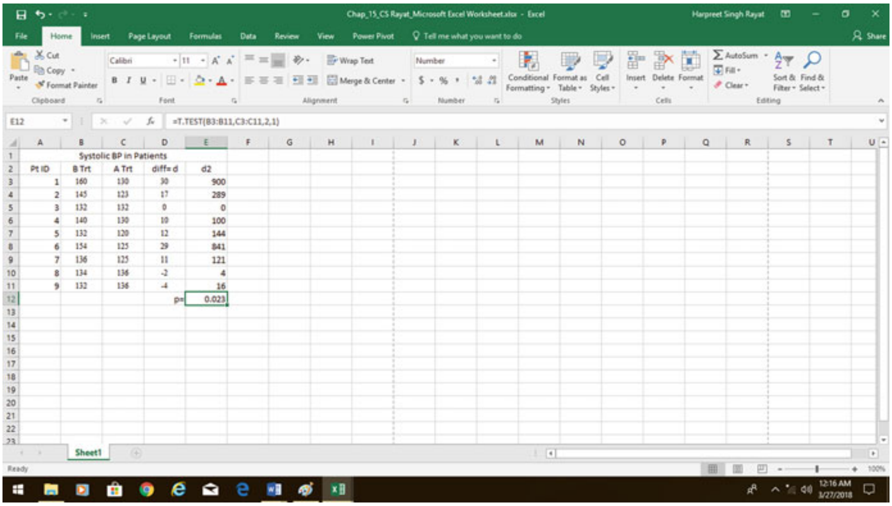
Let us revisit Chap. 11 to solve an example through application of Microsoft Excel function for paired *t*-test.

Example In a group of nine hypertensive patients, systolic blood pressure was recorded in mm of Hg, before and after treatment as exhibited in Table 15.2. Test the significance of treatment on patients or accept the null hypothesis (H_o).

The application of paired t-test through Microsoft Excel function gives out t-probability as $p = 0.023$ which in other words could be called $p < 0.05$. So, H_o stands rejected. Hence, the treatment is very much effective to control hypertension. Screenshot as Fig. 15.5 is produced herewith to exhibit the Excel worksheet function: =TTEST(B3:B11,C3:C11,2,1).

15.3 Designing Graphics with Microsoft Excel

We can design a variety of charts or graphics for the data arranged in the Excel worksheets using the insert function and thereby inserting the chart or graphic of choice. All the graphics depicted in this book have been crafted using Microsoft Excel worksheet.



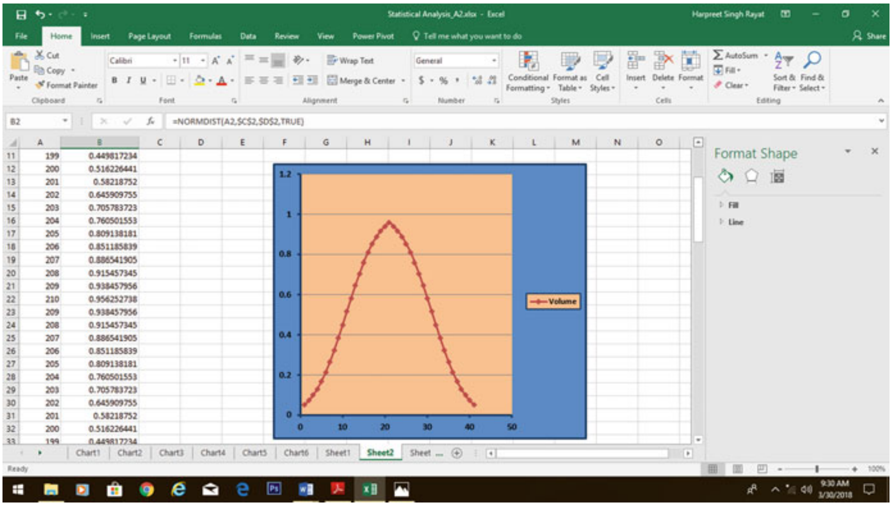
E12 =T.TEST(B3:B11,C3:C11,2,1)

	Systolic BP in Patients			
Pt ID	B Trt	A Trt	diff= d	d2
1	160	130	30	900
2	145	123	17	289
3	132	132	0	0
4	140	130	10	100
5	132	120	12	144
6	154	125	29	841
7	136	125	11	121
8	134	136	-2	4
9	132	136	-4	16
			p=	0.023

Fig. 15.5 Screenshot depicting paired t-test on Excel worksheet ($p = 0.023$)

Suppose we want to draw a "normal distribution curve" (bell-shaped curve). For that we will have to arrange our data from normal standard or from the population, in a column on Excel worksheet. The 50% data would be in the ascending order followed by 50% data in descending order. If we revisit the previous chapter and look into the Excel worksheet to understand the essential steps to be followed, it would come out as follows:

1. Data was arranged in column A, A3:A42.
2. The mean of the data has been calculated by statistical function in the cell C3and rounded off to whole number.
3. Standard deviation of data was calculated by statistical function in cell D3 and rounded off to three decimal places.
4. Normal distribution was computed in cell B3 by typing the following "syntax" in the formula bar after selecting the cell B3, =NORMDIST(A2,C2,D2, TRUE), which returned the value 0.05182661.
5. The formula was copied down in cells B4 to B42, which returned the distribution values as are visible in the Excel worksheet in Fig. 15.6.
6. Distribution values (B3:B42) were selected, and scatter chart was applied for generating normal distribution curve as depicted in Fig. 15.6.



B2 =NORMDIST(A2,C2,D2,TRUE)

	A	B
11	199	0.449817234
12	200	0.516226441
13	201	0.58218752
14	202	0.645909755
15	203	0.705783723
16	204	0.760501553
17	205	0.809138181
18	206	0.851185839
19	207	0.886541905
20	208	0.915457345
21	209	0.938457956
22	210	0.956252738
23	209	0.938457956
24	208	0.915457345
25	207	0.886541905
26	206	0.851185839
27	205	0.809138181
28	204	0.760501553
29	203	0.705783723
30	202	0.645909755
31	201	0.58218752
32	200	0.516226441
33	199	0.449817234

Fig. 15.6 Screenshot depicting normal distribution curve

15.4 Excel Worksheet Functions for Calculating Probabilities

For calculating probabilities for standard distributions, we use the following functions:

1. `BINOMDIST(x,k,`θ`,0:` This is $P(X = x)$, i.e., the probability function, when $X \sim \mathrm{Bin}(k, \theta)$.
2. `BINOMDIST(x,k,`θ`,1:` This is $P(X \leq x)$, i.e., the distribution function, when $X \sim \mathrm{Bin}(k, \theta)$.
3. `POISSON(x,`μ`,0) :` When $X \sim \mathrm{Poisson}(\mu)$.
4. `POISSON(x,`μ`,1) : POISSON(x,`μ`,1).`
5. `NORMDIST(x,`μ`,`σ`,1) : That is,` $P(X \leq x)$ when $X \sim N(\mu, \sigma 2)$.
6. `NORMDIST(x,`μ`,`σ`,0) :` The density function of X when $X \sim N(\mu, \sigma 2)$.

These probabilities are also helpful when we have to calculate p-values for non-normal distributions. For example, the *p*-value associated with the observed value $\mathrm{T}_{\mathrm{obs}}$ of the t-statistic for a two-sided *t*-test with 24 degrees of freedom is given by `TDIST`($\mathrm{T}_{\mathrm{obs}}$, 24, 2). Functions `FDIST` and `CHIDIST` are also available, for the F and χ^2 distributions, but should be treated with caution because the command syntax is not the same as for the `TDIST` command.

Now, let us practice the computation of *p*-value with “Excel worksheet function” after calculating *t*-value, *F*-value, or χ^2-value through manual applications of formulae. We have already solved many examples in previous chapters but referred to probability table values for taking decisions. Applications of "Excel worksheet function" are equally potent methods to work out probability of observed t-value, F-value, or χ^2-value.

Example 1 In a study it was observed that out of 50 non-vaccinated children, 20 got measles and out of 50 vaccinated children, 10 got measles. Evaluate significance of effectiveness of vaccination for immunity against measles.

Solution

In this example under Chap. 9, we came out with χ^2-value = 4.8, and degrees of freedom were (2–1)(2–1) = 1. Decision was rejection of H_o.

Here when we applied Excel worksheet function, CHISQ.DIST.RT: =CHISQ.DIST.RT(4.8,1), the p-value came out to be $p = 0.028$. That is very much less than p: 0.05. So, decision is the same; $p < 0.05$. Hence, H_o is rejected. The vaccination is significantly effective for immunity against measles.

Example 2 An experiment was conducted for assessing the effect of vitamin A-deficient diet. Out of the 20 inbred rats, 10 rats were fed on normal diet, and the other 10 were fed on vitamin A-deficient diet. The amount of vitamin A in the serum of rats of both groups was determined, and the mean and standard deviation were worked out as shown in Table 11.2 of Chap. 11 (please refer back to Chap. 11).

(a) Find out whether the mean value ($\bar{X}_2$) of rats fed on vitamin A-deficient diet is the same as the mean value ($\bar{X}_1$) of those fed on normal diet.
(b) If the difference is there, prove that this difference is due to sampling variation or due to the deficiency of vitamin A.

Solution

In this case student's t-test gave $t = 3.096$, and degrees of freedom were 10 + 10–2 = 18. The decision was rejection of H_o as $p < 0.01$ as worked out with the help of t-distribution table.

Here when we applied Excel worksheet function, T.DIST.2 T: =T.DIST.2 T (3.096,18), the p-value came out to be $p = 0.006$. That is very much less than p 0.01 ($p < 0.01$) So, the decision is the same. Hence, H_o is rejected. The difference in serum levels of vitamin A is due to vitamin A-deficient diet.

We conclude with the remarks that applications of "Excel worksheet functions" are a major substitute for "statistical tables" and are equally potent for delivering accurate p-values. However, we are free to consult "statistical tables" for taking statistical decisions.

Appendices

Appendix I: Table of Powers, Roots, and Reciprocals

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\frac{1}{n}$
1	1	1	1.000	1.000	1.000
2	4	8	1.414	1.260	0.500
3	9	27	1.732	1.442	0.333
4	16	64	2.000	1.387	0.250
5	25	125	2.236	1.587	0.200
6	36	216	2.449	1.817	0.167
7	49	343	2.646	1.913	0.143
8	64	512	2.828	2.000	0.125
9	81	729	3.000	2.080	0.111
10	100	1000	3.162	2.154	0.100
11	121	1331	3.317	2.224	0.091
12	144	1728	3.464	2.289	0.083
13	169	2197	3.606	2.351	0.077
14	196	2744	3.742	2.410	0.071
15	225	3375	3.873	2.466	0.067
16	256	4096	4.000	2.520	0.063
17	289	4913	4.123	2.571	0.059
18	324	5832	4.243	2.621	0.056
19	361	6859	4.359	2.668	0.053
20	400	8000	4.472	2.714	0.050
21	441	9261	4.583	2.759	0.048
22	484	10,648	4.690	2.802	0.045
23	529	12,167	4.796	2.844	0.043
24	576	13,824	4.899	2.884	0.042
25	625	15,625	5.000	2.924	0.040
26	676	17,576	5.099	2.962	0.038
27	729	19,683	5.196	3.000	0.037
28	784	21,952	5.292	3.037	0.036

(continued)

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\frac{1}{n}$
29	841	24,389	5.385	3.072	0.034
30	900	27,000	5.477	3.107	0.033
31	961	29,791	5.568	3.141	0.032
32	1024	32,768	5.657	3.175	0.031
33	1089	35,937	5.745	3.208	0.030
34	1156	39,304	5.831	3.240	0.029
35	1225	42,875	5.916	3.271	0.029
36	1296	46,656	6.000	3.302	0.028
37	1369	50,653	6.083	3.332	0.027
38	1444	54,872	6.164	3.362	0.026
39	1521	59,319	6.245	3.391	0.026
40	1600	64,000	6.325	3.420	0.025
41	1681	68,921	6.403	3.448	0.024
42	1764	74,088	6.481	3.476	0.024
43	1849	79,507	6.557	3.503	0.023
44	1936	85,184	6.633	3.530	0.023
45	2025	91,125	6.708	3.557	0.022
46	2116	97,336	6.782	3.583	0.022
47	2209	103,823	6.856	3.609	0.021
48	2304	110,592	6.928	3.634	0.021
49	2401	117,649	7.000	3.659	0.020
50	2500	125,000	1.000	3.946	0.020
51	2601	132,651	7.141	3.708	0.020
52	2704	140,608	7.211	3.733	0.019
53	2809	148,877	7.280	3.756	0.019
54	2916	157,464	7.348	3.780	0.019
55	3025	166,375	7.416	3.803	0.018
56	3136	175,616	7.483	3.826	0.018
57	3249	185,193	7.550	3.849	0.018
58	3364	195,112	7.616	3.871	0.017
59	3481	205,379	7.681	3.893	0.017
60	3600	216,000	7.746	3.915	0.017
61	3721	226,981	7.810	3.936	0.016
62	3844	238,328	7.874	3.958	0.016
63	3969	250,047	7.937	3.984	0.016
64	4096	262,144	8.000	4.000	0.016
65	4225	274,625	8.062	4.021	0.015
66	4356	287,496	8.124	4.041	0.015
67	4489	300,763	8.185	4.062	0.015
68	4624	314,432	8.246	4.082	0.015
69	4761	328,509	8.307	4.102	0.014
70	4900	343,000	8.367	1.121	0.014
71	5041	357,911	8.426	4.141	0.014
72	5184	373,248	8.485	4.160	0.014

(continued)

n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\frac{1}{n}$
73	5329	389,017	8.544	4.179	0.014
74	5476	405,224	8.602	4.198	0.014
75	5625	421,875	8.660	4.217	0.013
76	5776	438,976	8.718	4.236	0.013
77	5929	456,533	8.775	4.254	0.013
78	6084	474,552	8.832	4.291	0.013
79	6241	493,039	8.888	4.291	0.013
80	6400	512,000	8.944	4.303	0.013
81	6561	531,441	9.000	4.327	0.012
82	6724	551,368	9.055	4.344	0.012
83	6889	571,787	9.110	4.362	0.012
84	7056	592,704	9.165	4.380	0.012
85	7225	614,125	9.220	4.397	0.012
86	7396	636,056	9.274	4.414	0.012
87	7569	658,503	9.327	4.431	0.011
88	7744	681,472	9.381	4.448	0.011
89	7921	704,969	9.434	4.465	0.011
90	8100	729,000	9.487	4.481	0.011
91	8281	753,571	3.000	4.498	0.011
92	8464	778,688	9.592	4.514	0.011
93	8649	804,357	9.644	4.531	0.011
94	8836	830,584	9.695	4.547	0.011
95	9025	857,375	9.747	4.563	0.011
96	9216	884,736	9.798	4.579	0.010
97	9409	912,673	9.849	4.596	0.010
98	9604	941,192	9.899	4.610	0.010
99	9801	970,299	9.950	4.626	0.010
100	10,000	1,000,000	10.000	4.642	0.010

Appendix II: Table of Normal Distribution (Single Tail)

This table helps us to ascertain the percentage of “area” or “observations” in the “normal distribution curve” falling beyond various “standard deviation units.” $Z = \frac{x - \bar{x}}{\sigma}$; $Z = 0$ at the mean ($\bar{x}$).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4482	0.4442	0.4403	0.4363	0.4324	0.4285	0.4246
0.2	0.4207	0.4167	0.4128	0.4090	0.4052	0.4013	0.3974	0.3936	0.3898	0.3858
0.3	0.3820	0.3782	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.3181	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.14.01	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0437	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0219	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010

Values of *Z* up to one decimal point (0.1, 0.2, 0.3, etc.) have been given in vertical column, whereas horizontal values to the right of Z in the table give values up to two decimal point (0.1, 0.02, 0.03, etc.).

To determine the percentage of area or observations lying beyond the *Z* value if 1.96, first find the 1.90 in the first column at the left side and then go across the column and the value under 0.06. This value shows area $= 0.0250$. Therefore, percentage of observations will be $0.0250 \times 100 = 2.5\%$.

Appendix III: Table of Standard Normal Distribution Areas – Under the Standard Normal Curve

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.0844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3138	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3990	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4964	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4972	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4499	0.4499	0.4499	0.4499	0.4499	0.4499	0.4499	0.4499	0.4499	0.4499
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Appendix IV: Table of Critical Values of "*t*"

Level of significance for one-tailed test						
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
Level of significance for two-tailed test						
df	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.078	6.314	12.706	3.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.706	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Appendix V: Table of Critical Values of Correlation Coefficient – "*r*"

df	0.1	0.05	0.02	0.01	0.001
1	0.988	0.997	1.000	1.000	1.000
2	0.900	0.950	0.980	0.990	0.999
3	0.805	0.878	0.934	0.959	0.991
4	0.729	0.811	0.882	0.917	0.974
5	0.669	0.755	0.833	0.875	0.951
6	0.622	0.707	0.789	0.834	0.925
7	0.580	0.666	0.750	0.798	0.898
8	0.549	0.632	0.716	0.765	0.872
9	0.521	0.602	0.685	0.735	0.847
10	0.497	0.576	0.658	0.708	0.823
11	0.476	0.553	0.634	0.684	0.801
12	0.458	0.532	0.612	0.661	0.780
13	0.441	0.514	0.592	0.641	0.760
14	0.426	0.497	0.574	0.623	0.742
15	0.412	0.482	0.558	0.606	0.725
16	0.400	0.468	0.543	0.590	0.708
17	0.389	0.456	0.529	0.575	0.693
18	0.378	0.444	0.516	0.561	0.679
19	0.369	0.433	0.503	0.549	0.665
20	0.360	0.423	0.492	0.537	0.652
25	0.323	0.381	0.445	0.487	0.597
30	0.296	0.349	0.409	0.449	0.554
35	0.275	0.325	0.361	0.418	0.519
40	0.257	0.304	0.358	0.393	0.490
45	0.243	0.288	0.338	0.372	0.465
50	0.231	0.373	0.312	0.354	0.443
60	0.211	0.250	0.295	0.325	0.408
70	0.195	0.232	0.274	0.302	0.380
80	0.183	0.217	0.257	0.283	0.357
90	0.173	0.205	0.242	0.267	0.338
100	0.164	0.195	0.230	0.254	0.321

Appendix VI: Table of χ^2 Distribution

df	0.1	0.05	0.02	0.01	0.001
1	2.71	3.84	5.41	6.64	10.83
2	4.61	5.99	7.82	9.21	13.82
3	6.25	7.82	9.84	11.34	16.27
4	7.78	9.49	11.67	13.28	18.47
5	9.24	11.07	13.39	15.09	20.52
6	10.65	12.59	15.03	16.81	22.46
7	12.02	14.07	16.62	18.48	24.32
8	13.36	15.51	18.17	20.09	26.13
9	14.68	16.92	19.68	21.67	27.88
10	15.89	18.31	21.16	23.21	29.59
11	17.28	19.68	22.62	24.73	31.26
12	18.55	21.03	24.05	26.22	32.91
13	19.81	22.36	25.47	27.69	34.53
14	21.06	23.69	26.87	29.14	36.12
15	22.31	24.99	28.26	30.58	37.70
16	23.54	26.30	29.63	32.00	39.25
17	24.77	27.59	30.99	33.41	40.79
18	25.99	28.87	32.35	34.81	42.31
19	27.20	30.14	33.69	36.19	42.82
20	28.41	31.41	35.02	37.57	45.32
21	29.62	32.67	36.34	38.93	46.80
22	30.81	33.92	37.66	40.29	48.27
23	32.01	35.17	38.97	41.64	49.73
24	33.20	36.42	40.27	42.98	51.18
25	34.38	37.65	41.57	44.31	52.62
26	35.56	38.89	42.86	45.64	54.05
27	36.74	40.11	44.14	46.96	55.48
28	37.92	41.34	45.42	28.28	56.89
29	39.09	42.56	46.69	49.59	58.30
30	40.26	43.77	47.96	50.89	59.70

Appendix VII: Mann Whitney U-Test (One Tailed at 0.05 Level; Two Tailed at 0.10 Level)

	n																			
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–																			
2	–	–																		
3	–	–	0																	
4	–	–	0	1																
5	–	0	1	2	4															
6	–	0	2	3	5	7														
7	–	0	2	4	6	8	11													
8	–	1	3	5	7	10	13	15												
9	–	1	4	6	8	12	15	18	21											
10	–	1	4	7	9	14	17	20	24	27										
11	–	1	5	8	10	16	19	23	27	31	34									
12	–	2	5	9	11	19	21	26	30	34	38	42								
13	–	2	6	10	12	20	24	28	33	37	42	47	51							
14	–	3	7	11	13	23	26	31	36	41	46	51	56	61						
15	–	3	7	12	14	23	28	33	39	44	50	55	61	66	72					
16	–	3	8	13	15	25	30	36	42	48	54	60	65	71	77	83				
17	–	3	9	14	16	26	33	39	45	51	57	64	70	77	83	89	96			
18	–	4	9	15	18	28	35	41	48	55	61	68	75	82	88	95	102	109		
19	0	4	10	16	19	30	37	44	51	58	65	72	80	87	94	101	106	116	123	
20	0	4	11	17	20	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
21	0	5	11	18	22	34	41	49	57	65	73	81	89	97	105	113	121	130	138	146
22	0	5	12	19	23	36	44	52	60	68	77	85	94	102	111	119	128	136	145	154
23	0	5	13	20	25	37	46	54	63	72	81	90	98	107	116	125	134	143	152	161
24	0	6	13	21	26	39	48	57	66	75	85	94	103	113	122	131	141	150	160	162
25	0	6	14	22	28	41	50	60	69	79	89	98	108	118	128	137	147	157	167	177
26	0	6	15	23	29	43	53	62	72	82	92	103	113	123	133	143	154	164	174	185
27	0	7	15	24	30	45	55	65	75	86	96	107	117	128	139	149	160	171	182	192
28	0	7	16	25	32	46	57	68	78	89	100	111	122	133	144	156	167	178	186	200
29	0	7	17	26	33	48	59	70	82	93	104	116	127	138	150	162	173	185	196	208
30	0	7	17	27	35	50	61	73	85	96	108	120	132	144	156	168	180	192	204	216
31	0	8	18	28	36	52	64	76	88	100	112	124	136	149	161	174	186	199	211	224
32	0	8	19	29	38	54	66	78	91	103	116	128	141	154	167	180	193	206	218	231
33	0	8	19	30	39	56	68	81	94	107	120	133	146	159	172	186	199	212	226	239
34	0	9	20	31	40	57	70	84	97	110	124	137	151	164	178	192	206	219	233	247
35	0	9	21	32	42	59	73	86	100	114	128	141	156	170	184	198	212	226	241	255
36	0	9	21	33	43	61	75	89	103	117	131	146	160	175	189	204	219	233	248	263
37	0	10	22	34	45	63	77	91	106	121	135	150	165	180	195	210	225	240	255	271
38	0	10	23	35	46	65	79	94	109	124	139	154	170	185	201	216	232	247	263	278
39	1	10	23	37	49	67	82	97	112	128	143	159	175	190	206	222	238	254	270	286
40	1	11	24	38	53	68	84	99	113	131	147	163	179	196	212	228	245	261	278	294

Appendix VIII: Mann Whitney U-Test (One Tailed at 0.025 Level; Two Tailed at 0.05 Level)

	n																			
m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–																			
2	–	–																		
3	–	–	–																	
4	–	–	–	0																
5	–	–	0	1	2															
6	–	–	1	2	3	4														
7	–	–	1	3	5	6	8													
8	–	0	2	4	6	8	10	13												
9	–	0	2	4	7	10	12	15	17											
10	–	0	3	5	8	11	14	17	20	23										
11	–	0	3	6	9	13	16	19	23	26	30									
12	–	1	4	7	11	14	18	22	26	29	33	37								
13	–	1	4	8	12	16	20	24	28	33	34	41	45							
14	–	1	5	9	13	17	22	26	31	36	40	45	50	55						
15	–	1	5	10	14	19	24	29	34	45	44	49	54	59	64					
16	–	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75				
17	–	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87			
18	–	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99		
19	–	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	
20	–	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	–	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	–	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	–	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	–	3	10	17	25	33	42	50	59	67	6	85	94	102	111	120	129	138	147	156
25	–	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145	154	163
26	–	4	11	19	28	37	46	55	64	74	83	93	102	112	122	132	141	151	161	171
27	–	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158	168	178
28	–	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164	175	186
29	–	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171	182	193
30	–	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177	189	200
31	–	5	14	24	34	45	56	67	78	90	101	113	125	136	148	160	172	184	196	208
32	–	5	14	24	35	46	58	69	81	93	105	117	129	141	153	166	178	190	203	215
33	–	5	15	25	37	48	60	72	84	96	108	121	133	146	159	171	184	197	210	222
34	–	5	15	26	38	50	62	74	87	99	112	125	138	151	164	177	190	203	217	230
35	–	6	16	27	39	51	64	77	89	103	116	129	142	156	169	183	196	210	224	237
36	–	6	16	28	40	53	66	79	92	106	119	133	147	161	174	188	202	216	231	245
37	–	6	17	29	41	55	68	81	95	109	123	137	151	165	180	194	209	223	238	252
38	–	6	17	30	43	56	70	84	98	112	127	141	156	170	185	200	215	230	245	259
39	0	7	18	31	44	58	72	86	101	115	130	145	160	175	190	206	221	236	252	267
40	0	7	18	31	45	59	74	89	103	119	134	149	165	180	196	211	227	243	258	274

Bibliography

1. Aitken AC. Determinants and matrices. University Mathematical Texts-1. 3rd ed. Edinburgh/London: Oliver and Boyd; 1944.
2. Aitken AC. Determinants and matrices. University Mathematical Texts-2. 3rd ed. Edinburgh/London: Oliver and Boyd; 1944.
3. Bancroft H. Introduction to biostatistics. New York: Harper and Row; 1963.
4. Bartlett MS. On the theory of statistical regression. Proc R Soc. 1933;53:54.
5. Belk WP, Sunderman FW. Am J Clin Path. 1947;17:853.
6. Bocher M. Introduction to higher algebra. New York: Macmillan; 1908.
7. Cramer H. Mathematical methods of statistics. First Indian edition. Bombay: Asia Publishing House; 1962.
8. David FN. Tables of correlation coefficient. London: Biometrika Office, University College; 1938.
9. Fisher RA. Statistical methods for research workers. 8th ed. Edinburgh/London: Oliver and Boyd; 1941.
10. Fisher RA. The design of experiments. 2nd ed. Edinburgh/London: Oliver and Boyd; 1937.
11. Fisher RA. Applications of student's distribution. Metron. 1925;5(3):90.
12. Fisher RA, Corbet AS, Williams CB. The relation between the number of species and number of individuals in a random sample of an animal population. J Anim Ecol. 1943;12:42.
13. Fisher RA, Yates F. Statistical tables. 2nd ed. Edinburgh/London: Oliver and Boyd; 1944.
14. ISO 15189. Medical Laboratories – requirements for quality and competence. 2012.
15. Jefereys H. Theory of probability. Oxford: Clarendon Press; 1939.
16. Kaushal TL. Statistical analysis. 3rd ed. New Delhi: Kalyani Publishers; 1999.
17. Kendell MG. The advanced theory of statistics, I. London: Charles Griffin; 1943.
18. Keynes JM. A treatise on probability. London: MacMillan; 1921.
19. King EJ, Wootton IDP. Micromethods in medical biochemistry. 3rd ed. London: Churchill. p. 1956.
20. Mahajan BK. Methods in biostatistics. 7th ed. New Delhi: JAYPEE Brothers Medical Publishers (P) Ltd; 2010.
21. Pearson K. On the criterion that a given system of deviations from the probable in case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag. 1900;50(V):157.
22. Pearson K. Tables for statisticians and biometricians, I. 2nd ed. Cambridge: Cambridge University Press; 1924.
23. Pearson K. Tables for statisticians and biometricians, II. 2nd ed. Cambridge: Cambridge University Press; 1931.
24. Rayat CS. Quality management in medical laboratories for accreditation. Austin J Pathol Lab Med. 2017;4(1):1019.
25. Rayat CS, Dutta U. Preserving Millipore filters with chemotactic cells. Bull PGI. 1985;19(1):13–5.

C. S. Rayat, *Statistical Methods in Medical Research*,
https://doi.org/10.1007/978-981-13-0827-7

26. Sen PK. Biostatistics: statistics in biomedical, public health and environmental sciences. Amsterdam: Publications of BG Greenberg; 1985.
27. Sheppard WF. Tables of probability integral in British Association, mathematical tables vol-7. London: Cambridge University Press; 1939.
28. Student. The probable error of a mean. Biometrika. 1908;6:1.
29. Student. The probable error of a correlation coefficient. Biometrika. 1908;6:302.
30. Uspensky JV. Introduction to mathematical probability. New York: McGraw Hill; 1937.
31. Wilks SS. The theory of statistical inference. Ann Arbor; 1937.
32. Woodward HR, Goldsmith PL. Cumulative sum technique (published for I.C.I.). Edinburgh: Oliver and Boyd; 1964.
33. Wootton IDP, King EJ. Lancet. 1953;1:470.
34. Yates F. Contingency tables involving small numbers and the χ^2 test. JRS Suppl. 1934;1:217.
35. Yule GU, Kendall MG. An introduction to theory of statistics. 12th ed. London: Charles Griffin & Co; 1940.