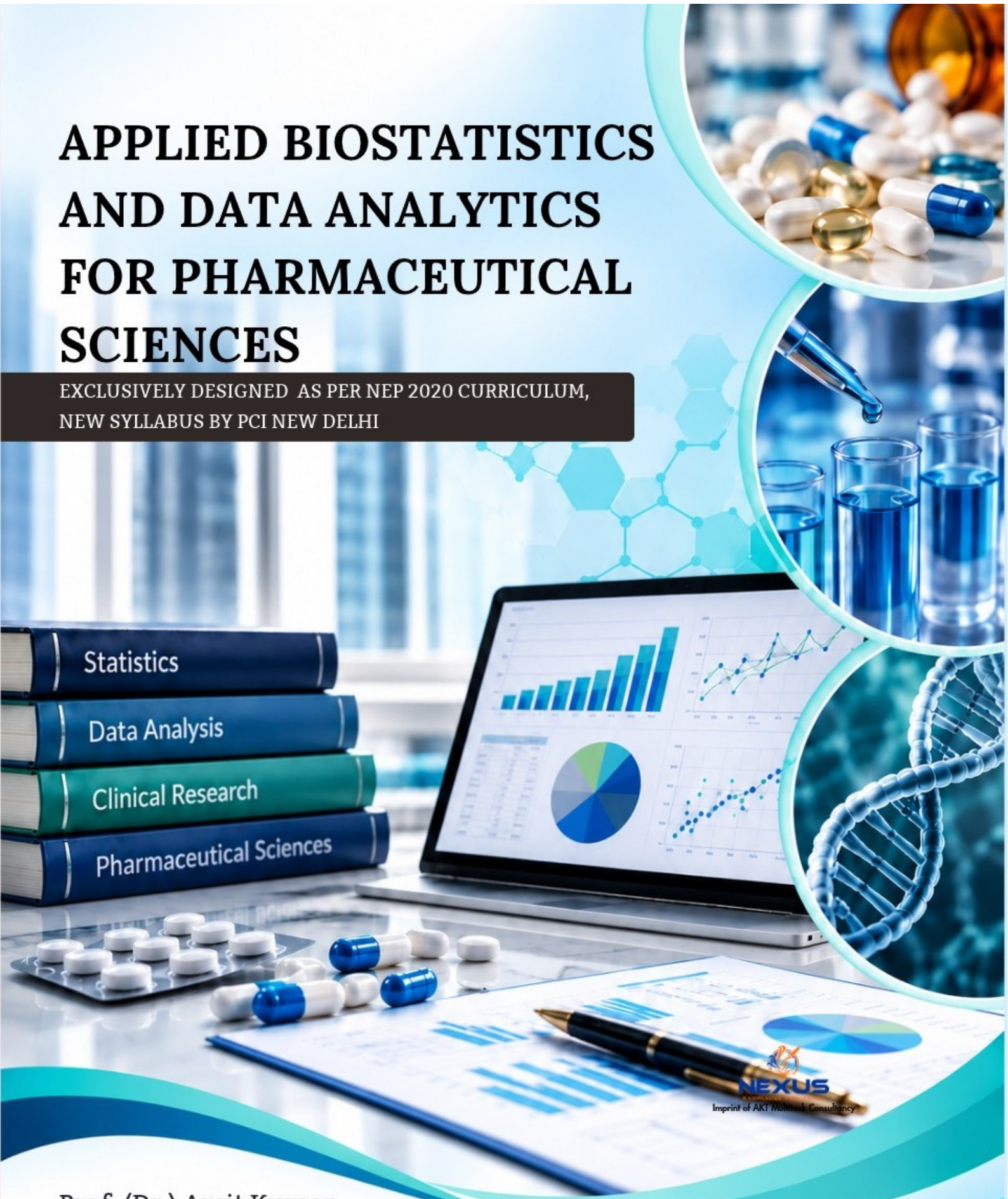
APPLIED BIOSTATISTICS AND DATA ANALYTICS FOR PHARMACEUTICAL SCIENCES

EXCLUSIVELY DESIGNED AS PER NEP 2020 CURRICULUM, NEW SYLLABUS BY PCI NEW DELHI



Prof. (Dr.) Amit Kumar
Mrs. Bhagyashri Vilas Gawade
Dr. Charu Sinha
Dr. Richa Srivastava
Mr. Selvamuthukumar R
Mrs. Manisha Chandrakar

NEXUS
KNOWLEDGE PUBLICATION
Imprint of AKT Multitask Consultancy

APPLIED BIOSTATISTICS AND DATA ANALYTICS FOR PHARMACEUTICAL SCIENCES

EXCLUSIVELY DESIGNED AS PER NEP 2020 CURRICULUM, NEW SYLLABUS BY PCI NEW DELHI

Prof. (Dr.) Amit Kumar

Mrs. Bhagyashri Vilas Gawade

Dr. Charu Sinha

Dr. Richa Srivastava

Mr. Selvamuthukumar R

Mrs. Manisha Chandrakar

NEXUS
KNOWLEDGE PUBLICATION
Imprint of AKT Multitask Consultancy

Imprint of AKT Multitask Consultancy

NEXUS KNOWLEDGE PUBLICATION

All right reserved

© Prof. (Dr.) Amit Kumar, Mrs. Bhagyashri Vilas Gawade, Dr. Charu Sinha
Dr. Richa Srivastava, Mr. Selvamuthukumar R, Mrs. Manisha Chandrakar

No part of this publication may be re-produced, stored in a retrieval system or distributed in any form or by any means, electronic, mechanical, photocopying, recording, scanning, web or otherwise without the written permission of the publisher. NEXUS has obtained all the information in this book from the sources believed to be reliable and true, However, NEXUS or its editors or authors or illustratorsdon't take any responsibility for the absolute accuracy of any information publishedand the damages or loss suffered there upon. All disputes subject to Bilaspur jurisdiction only.

Administrative & Production Office
NEXUS KNOWLEDGE PUBLICATION,
Imprint of AKT Multitask Consultancy,
Near Satbahinia Mandir, Arvind Nagar, Sarkanda Bilaspur, Chhattisgarh, India, Pin- 495006

ISBN: 978-81-180323-0-4
PRICE: Rs 499/-

Manuscript Edited By **Gijeesh Nair**
FIRST PUBLISHED BY NEXUS KNOWLEDGE PUBLICATION IN **2026**
PRINTED & BOUNDED BY JEC PRINTING TECHNOLOGIES (JEC PRESS UNIT)

DOI 10.5281/zenodo.23733585

PREFACE

Applied Biostatistics and Data Analytics for Pharmaceutical Sciences is designed to provide a concise and practical understanding of biostatistical concepts used in pharmaceutical and clinical research. It introduces readers to the basic principles of pharmaceutical data analysis and statistical interpretation.

The book covers descriptive statistics, probability distributions, data presentation, and measures of central tendency and dispersion. These concepts help in summarizing pharmaceutical data and understanding important patterns and variations in research observations.

It also explains sampling and statistical inference, including population and sample, sampling techniques, sampling errors, confidence intervals, hypothesis testing, p-values, and statistical significance. These methods support reliable interpretation of findings from clinical and pharmaceutical studies.

The book further discusses correlation and regression techniques, including Pearson correlation, dose-response relationships, linear regression, odds ratios, and risk analysis. These methods help researchers examine relationships between variables and interpret statistical findings appropriately.

A major feature of the book is the practical use of Python for pharmaceutical data analysis. Tools such as Pandas, SciPy, Stats models, and Scikit-learn are introduced for data preparation, statistical analysis, modelling, inference, and prediction, helping readers connect statistical concepts with practical pharmaceutical research applications.

With sincere appreciation,

Authors

ABOUT THE AUTHORS

Prof. (Dr.) Amit Kumar



Prof. (Dr.) Amit Kumar currently working as a professor in Faculty of Pharmaceutical Sciences, Motherhood University, Roorkee, Uttarakhand, India. Prof. (Dr.) Amit Kumar is a pharmaceutical academician and researcher with over 17 years of experience in teaching, research, and academic administration. He holds a Ph.D. in Pharmaceutical Sciences, M.Pharm in Clinical Pharmacy, and B.Pharm. He has served as Principal and subsequently as Deputy Registrar and Registrar at Smt. Manjira Devi University, Uttarakhand.

His areas of research include Drug Design and Development, Molecular Docking, Ethnopharmacology, and Phytochemical Screening. He has published 60 National and International research papers**,** contributed **to** 06 national and international patents, and authored 06 pharmacy books. He has guided 08 Ph.D. research scholars., 10 Postgraduate, and 15 Undergraduate research projects and has actively contributed to seminars, workshops, conferences, curriculum development, and scientific peer review publication. He is associated with IIP Press as a reviewer/editorial contributor and is a member of the Society of Higher Education (SHER**)**.

Mrs. Bhagyashri Vilas Gawade



Mrs. Bhagyashri V. Gawade is an Assistant Professor in Faculty of Health Sciences at JSPM's University, Wagholi, Pune, India and is pursuing her Ph.D. in Pharmaceutical Sciences at JSPM University. She holds an M. Pharm. in Pharmaceutics from SVPM's College of Pharmacy, Malegaon, Baramati and has more than 15 years of academic and teaching experience in pharmaceutical sciences. Her research interests include novel drug delivery systems, formulation development, Transdermal Drug Delivery System, Self Emulsifying Drug Delivery System and pharmaceutical research. She has authored and co-authored several research publications, published 1 Design Patent and a book chapter in the field of pharmaceutics. She has also received research funding under the ASPIRE scheme of Savitribai Phule Pune University and has guided number of M. Pharmacy and B. Pharmacy students at research level. She has contributed to academic administration, examination, accreditation, and student-development activities.

Dr. Charu Sinha



Dr. Charu Sinha is a healthcare management, quality, auditing, training and academic professional with a BDS and an MBA in Hospital & Healthcare Management, with more than 10 years of experience across Research, Academics, Training and Quality. Complemented by advanced credentials in ISO Quality Management Systems, Lean Six Sigma, Clinical Research and AI-enabled quality improvement. She has professional experience across hospital administration, clinical research, academic coordination, training, healthcare quality and process improvement, with associations including Centre for Sight, Apollo Hospitals, Indian Spinal Injuries Centre, Ikure, and many more

As an ISO Lead Auditor, she has conducted and participated in management system audits across healthcare and diverse service/industrial sectors, covering ISO standards. She is also a Lean Six Sigma Master Black Belt with expertise in QMS, NABH, healthcare KPIs, risk-based thinking, process improvement and regulatory compliance. As Founder of Dental Plus Direct, she leads a multidisciplinary team and develops professional education, quality, accreditation and training initiatives in dentistry, collaborating with professional organizations for webinars and capacity-building programs. She has served as a guest speaker and trainer on healthcare quality, benchmarking and NABH, contributed to multidisciplinary research, publications, and patents. She is also skilled in AI-assisted research and productivity tools, SPSS, Advanced Excel, statistical analysis, data visualization and digital learning platforms, integrating quality management, healthcare expertise and emerging technologies in her professional and academic work.

Dr. Richa Srivastava



Dr. Richa Srivastava is a Ph.D.-qualified Pharmaceutical Scientist and Professor at Ambekeshwar Institute of Pharmaceutical Sciences , with expertise in Pharmaceutics, pharmaceutical research, and novel drug delivery systems. Her academic and research experience includes polyherbal formulations, targeted nasal drug delivery, thermoreversible gels, liposomal drug delivery, and data analytics in polyherbal formulations. She has contributed to peer-reviewed research publications and presented her work at national and international scientific forums.Through Applied Biostatistics and Data Analytics for Pharmaceutical Sciences (Theory),she aims to provide pharmacy students with a clear and practical foundation in biostatistics and data analytics, helping them understand the principles needed to interpret pharmaceutical research and make informed scientific decisions.

Mr. Selvamuthukumar R



Mr, Selvamuthukumar R, M.Pharm., is an academician and researcher in the field of Pharmaceutical Sciences with expertise in Pharmaceutics, novel drug delivery systems, and pharmaceutical formulation development. He is currently serving as an Associate Professor in the Department of Pharmaceutics and has over five years of teaching B. Pharmacy, D. Pharmacy, and Pharm. D and research experience. His research interests include nanosponges, microspheres, transdermal drug delivery systems, emulgels, and the application of artificial intelligence in pharmaceutical formulation development. He has qualified M. Pharmacy in Industrial Pharmacy branch from, Annamalai University, and B. Pharmacy from Annamalai University, Chidambaram, Tamilnadu, India. He has guided many students of B. Pharmacy at research level. He has more than 9 publications. He has actively contributed to scientific research through publications in peer-reviewed journals and participation in national and international conferences. His commitment to academic excellence and innovative research continues to support advancements in pharmaceutical education and drug delivery technology.

Mrs. Manisha Chandrakar



Mrs. Manisha Chandrakar is an Assistant Professor at Chhattisgarh Swami Vivekanand Technical University (CSVTU). She has academic and teaching experience in the field of Pharmaceutical Sciences. Her area of expertise is Pharmacology and related pharmaceutical sciences. She has a keen interest in research, innovation, and the development of scientific knowledge. She has published research and review papers in SCI, Scopus, and UGC-indexed journals. She has contributed to the field of academics through her teaching and research activities. She is the author/co-author of two books in the field of Pharmaceutical Sciences. She is also a patent holder reflecting her interest in research and innovation. Her research work includes pharmacological evaluation and exploration of medicinal plants. She is dedicated to quality education, scientific research, and the overall development of Pharmaceutical Sciences.

ACKNOWLEDGEMENT

We express our sincere gratitude to all those who contributed directly and indirectly to the preparation of this book, **"Applied Biostatistics and Data Analytics for Pharmaceutical Sciences."** The book has been developed with the aim of presenting statistical concepts and data analytical approaches in a clear and application-oriented manner for pharmaceutical and clinical research.

We are thankful to the teachers, researchers, pharmaceutical professionals, and academic colleagues whose knowledge and guidance have helped in shaping the content of this work. Their valuable suggestions encouraged us to include important areas such as pharmaceutical data, descriptive statistics, probability, statistical inference, correlation, regression, and interpretation of research findings.

We also acknowledge the importance of practical and data-driven approaches in pharmaceutical research. The inclusion of applications involving clinical trials, pharmacovigilance, pharmacokinetics, pharmacodynamics, quality control, manufacturing, and hospital records has helped make the subject more relevant to real-world pharmaceutical research.

Special appreciation is extended to everyone who supported the integration of **Python-based statistical analysis and data analytics** into the book. The use of tools and libraries such as Pandas, SciPy, Stats models, and Scikit-learn provides learners with practical approaches for exploring, analysing, modelling, and interpreting pharmaceutical datasets.

Finally, we are grateful to our colleagues, students, academic institutions, and all readers whose interest in pharmaceutical sciences and statistical research continues to inspire this work. We hope this book serves as a useful academic and practical resource for understanding biostatistics, data analytics, and evidence-based research in pharmaceutical sciences.

LIST OF CONTENTS

UNIT 1… **1**

DESCRIPTIVE STATISTICS IN PHARMACEUTICAL SCIENCES **1**
1.1. FUNDAMENTALS OF PHARMACEUTICAL DATA 1
1.2. MEASURES OF CENTRAL TENDENCY 13
1.3. MEASURES OF DISPERSION AND DISTRIBUTION SHAPE 20
1.4. DESCRIPTIVE ANALYSIS USING PYTHON 30

UNIT 2… **47**

PROBABILITY AND STATISTICAL DISTRIBUTIONS IN HEALTHCARE **47**
2.1. FUNDAMENTALS OF PROBABILITY 47
2.2. RANDOM VARIABLES AND PROBABILITY MODELS 58
2.3. IMPORTANT STATISTICAL DISTRIBUTIONS 69
2.4. PROBABILITY DISTRIBUTIONS USING PYTHON 81

UNIT 3… **102**

SAMPLING AND STATISTICAL INFERENCE **102**
3.1. POPULATION, SAMPLE AND SAMPLING TECHNIQUES 102
3.2. CENTRAL LIMIT THEOREM AND CONFIDENCE INTERVALS 112
3.3. HYPOTHESIS TESTING AND STATISTICAL SIGNIFICANCE 118
3.4. STATISTICAL INFERENCE USING PYTHON 123

UNIT 4… **132**

CORRELATION AND REGRESSION IN PHARMACEUTICAL RESEARCH **132**
4.1. FUNDAMENTALS OF CORRELATION 132
4.2. VISUALIZATION OF RELATIONSHIPS IN PHARMACEUTICAL DATA 139
4.3. SIMPLE LINEAR REGRESSION 147
4.4. ODDS RATIO AND CLINICAL RISK ANALYSIS 154

UNIT 5… **166**

STATISTICAL ANALYSIS USING PYTHON THROUGH CASE-BASED LEARNING **166**
5.1. PHARMACEUTICAL DATA ANALYSIS WITH PYTHON 166
5.2. CASE-BASED DESCRIPTIVE AND CORRELATION ANALYSIS 174
5.3. REGRESSION ANALYSIS AND STATISTICAL INFERENCE 179
5.4. STATISTICAL REPORTING AND RESEARCH INTERPRETATION 184

Unit 1...

DESCRIPTIVE STATISTICS IN PHARMACEUTICAL SCIENCES

This unit presents the basics of probability and the distributions in healthcare and biomedical/pharmaceutical research. How to describe uncertainty in terms of probability models using random variables is presented. This unit provides information about discrete and continuous distributions and their applications. Normal, binomial and Poisson distributions are considered and their uses in biological measurements, clinical trials, treatment response, drug adverse reaction and hospital events are provided. Also, the utilization of Python in computing and plotting the probability distributions with various graphical and numerical summaries of data is provided. Such knowledge helps in understanding patterns in the data, predicting the probability of clinical outcomes, choosing appropriate statistical techniques and making decisions based on evidence in healthcare research.

1.1. FUNDAMENTALS OF PHARMACEUTICAL DATA

Pharmaceutical data is the type of information which is gathered, recorded, quantified and analyzed during the process of development, testing, manufacture and administration of pharmaceutical products. Pharmaceutical data forms the basis of scientific research and evidence for making decisions on clinical, quality and regulatory grounds. Pharmaceutical data may include information about drug components, physical and chemical properties, bioactivity, pharmacodynamics, patient data, results of treatment, side effects of drugs, and manufacturing procedures. The nature, sources and methods of data gathering are crucial aspects to understand since the quality of statistics analysis relies greatly upon the reliability and adequacy of data. Classification of data also enables researchers to choose right statistical methods and make valid conclusions.

1.1.1. Types and Measurement Scales of Data

Data can be classified according to its characteristics, nature, and method of measurement. Appropriate data classification is essential in the pharmaceutical sector since distinct methods of data collection, presentation, and analysis are required for each category of data. Qualitative and quantitative data are the two primary categories of data.

A. Qualitative Data

Qualitative data are alternatively referred to as categorical data and pertain to the description of qualities, properties, or classes of people, drugs, or observations. They are usually described in words and categories and rarely in numbers. Numbers, if used, are used only as a coding system.

Examples in pharmaceutical sciences include

- Dosage form type like tablet, capsule, syrup or injection form.
- Gender of participants in a clinical trial.
- Adverse drug effect or no adverse drug effect.
- Disease or disorder type.
- Drug classification into prescription or non-prescription.
- Batch of product status as approved or rejected.

Qualitative data can be used to classify people into categories or groups and to describe a population.

Types of qualitative data include:

1. Nominal Data

The nominal type of data includes categories that are unordered. It is utilized in order to classify the observations and all numerical codes are used just as labels. For example, in pharmacology, there are such types of dosage forms as tablets, capsules, injections, and creams; such types of blood groups as A, B, AB, O; and such ways of administration as oral, intravenous, and intramuscular. Likewise, there could be such types of adverse drug reaction as gastrointestinal, dermatological, and neurological. Those categories cannot be ordered in terms of highest and lowest. If, for instance, we code tablets as 1, capsules as 2, and injections as 3, then the number 3 does not tell us that injection is higher than tablet. Numbers are just the labels of different categories.

There are such tools as frequencies, percentages, and mode that are used to summarize nominal data. Frequencies represent the amount of observation within each category, while percentages represent the part of observation within each category. Mode represents the category with the highest frequency. In case when a researcher observes 100 pharmaceutical dosage forms and makes a record of each of them, he/she can estimate how many of them are tablets, capsules, injections, and creams.

2. Ordinal Data

In ordinal data, the variables are categorized in a certain order or rank. It is unlike the nominal data where there are no ranks in the categories, meaning, it does not imply that an observation has a higher or lower position compared to another observation. However, it should be noted that the differences between categories of ordinal data are not always equal or quantitative. The common examples of

ordinal data in the context of pharmaceuticals include the seriousness of the adverse effects of the drug which is usually categorized into mild, moderate, or severe. Patient satisfaction is also an example of ordinal data where a patient is categorized into dissatisfied, neutral, satisfied, or highly satisfied; while in case of disease severity, it is categorized into low, moderate, or high. “Severe” means that it is higher than “moderate” but it does not mean that the difference between “mild” and “moderate” is equal to the difference between “moderate” and “severe.”

The summary measures that can be used on ordinal data include frequencies, percentages, median, and mode. While the median indicates the middle category of the ordered observation, frequency and percentage provide the distribution of the observations based on the ordered categories. The mode displays the category that appears most frequently in the ordered observation. For instance, in a study of patient satisfaction with a certain medicine, the researcher can collect data regarding the number of patients which are dissatisfied, neutral, satisfied, or highly satisfied.

B. Quantitative Data

Quantitative data refers to numeric data representing quantity. It is acquired by means of counting or measurements, and it can be manipulated using mathematics.

Examples in pharmaceutical sciences include

- Drug concentration in plasma.
- Weight of a tablet.
- Blood pressure of a patient.
- Number of tablets produced in a batch.
- Time required for drug dissolution.
- Amount of active pharmaceutical ingredient in a formulation.
- Patient age and body weight.

Quantitative data are generally divided into discrete data and continuous data.

1. Discrete Data

Numerical numbers that can be counted and are typically integers are known as discrete data. Discrete data includes things like the quantity of stuff or the frequency of an event. The number of patients in a trial, the number of tablets in a batch, the number of drug adverse reactions, the number of defective products in a production process, and the number of dosages given to a patient are all examples of discrete data in pharmaceutical research. Discrete data is obtained via counting and, therefore, normally cannot have fractional values. Thus, for instance, 100 patients can participate in a study,

but not 100.5 patients. Discrete data are used to estimate the number of events, sizes of sets, and quantities of pharmaceutical products.

Discrete data are often summarized by using frequencies, proportions, and count statistics. Frequencies demonstrate how often an event occurs, whereas proportions reveal the fraction of cases included into a certain category. For instance, when 5 defective tablets are found in a batch of 100 tablets, then the frequency of defective tablets is equal to 5 and the proportion is 5 %. Such kind of data can be represented in frequency tables and with the help of bar charts and other appropriate graphs. Discrete data are very useful in pharmaceutical research since they allow estimating different countable outcomes.

2. Continuous Data

Continuous data have values that span across a specific range of possible numerical measurements. They come from measuring the results and might include fractional values. In pharmaceutical research, some continuous data include body weight, which is 62.5 kg; drug concentration, which is 4.75 μg/mL; tablet hardness, which is 6.8 kg/cm²; dissolution time, which is 12.4 minutes; blood glucose, which is 98.6 mg/dL; and body temperature, which is 37.2°C. Contrary to discrete data, continuous data do not have to be whole numbers. For instance, when measuring the weight of a tablet, one can obtain 250.4 mg, 250.45 mg, or other values based on the precision of the measurement tool.

For continuous data, the mean, median, standard deviation, variance, and range are common indicators of central tendency and dispersion. The median is the central value, whereas the mean is the average of all measurements. The variance and standard deviation indicate how far the results deviate from the mean. The range displays the separation between the highest and lowest numbers. For example, while examining a drug's concentration in blood samples, researchers will determine the average concentration and gauge how much those values differ among patients.

C. Measurement Scales of Data

The measurement scales refer to the degree of information possessed by the data and dictate which mathematical and statistical functions are applicable. The four types of measurement scales include nominal, ordinal, interval, and ratio.

1. Nominal Scale

The nominal scale is the simplest level of measurement. This method categorizes the observations into categories which have no natural order. Each category is mutually exclusive; that is, an observation cannot be classified into more than one category at a time. The numerical values used to represent the categories are only for labelling purposes, and do not denote quantity or superiority. In

the context of pharmaceutical science, examples include dosage form type, method of drug delivery, blood type, adverse drug reaction type, and brand/generic categorization of drugs. For example, the numbers 1, 2, and 3 may be assigned to tablets, capsules, and injections respectively, but the use of number 3 does not imply that injections are better than tablets.

Methods of summarizing nominal data include the calculation of frequency, percentage and the mode. While percentages show the percentage of all observations that fall into each group, frequencies show how many observations fall into each category. The category that appears most frequently is indicated by the mode. For instance, a researcher might tally the number of patients receiving intramuscular, intravenous, and oral medications and present the results as percentages. Pie charts, bar charts, and frequency tables are ways to display nominal data. One statistical test that can be used with such data is the chi-square test.

2. Ordinal Scale

In contrast to nominal, the ordinal scale involves the classification of variables into different groups or levels with an inherent ranking order. The distances between different ranks of ordinal categories are not necessarily constant or measurable. Some common uses of the ordinal scale in pharmaceutical science include assessment of severity of adverse drug reactions, assessment of the level of patient satisfaction, categorization of the severity of a disease, and evaluation of the quality of pharmaceutical products. For instance, the severity of adverse drug reaction can be ranked as mild, moderate, or severe. Although severe is more severe compared to moderate, it cannot be stated that the distance between mild and moderate is equal to that between moderate and severe.

When dealing with data on an ordinal scale, descriptive statistics such as frequency, percentage, median, mode, and percentile can be applied. After the data are sorted by magnitude, the mid-point category is represented by the median value. Conversely, the category that appears most frequently is referred to as the mode. For example, the level of patient satisfaction can be determined by asking patients to classify their responses as being dissatisfied, neutral, satisfied, or highly satisfied. This type of data can be reported as frequency tables or bar charts. The appropriate statistical tests are non-parametric in nature, such as Mann–Whitney U and Kruskal-Walli's test.

3. Interval Scale

Numerical data with a consistent and significant spacing between successive numbers is called an interval scale. It does not, however, have an absolute zero point. It suggests that the absence of the quality being measured is not represented by zero. While ratio cannot be effectively interpreted, addition and subtraction are relevant. For example, a temperature of 20°C is 10°C higher than 10°C, yet stating that 20°C is twice as much as 10°C is meaningless. Temperature in degrees Celsius in

storage studies and some measurement scores in clinical research are two instances of interval scale data in pharmaceutical research.

Depending on the needs of the study, interval scale data can be analysed by computing measures of central tendency like mean and median and measures of dispersion like standard deviation, variance, and range. For instance, one can compute the average temperature reading taken throughout a drug stability study and ascertain the variance of these data. If the assumptions of the chosen test are met, interval scale data may be subjected to further analytical methods such as correlation, regression, and parametric testing.

4. Ratio Scale

The highest level of measurement is represented by the ratio scale. In addition to having an absolute zero, which denotes the complete absence of the measured feature, it has all the characteristics of an interval scale. As a result, all mathematical operations, such as division and multiplication, are applicable, numbers are ordered, and intervals are equal. The concentration of the drug material, body weight, tablet weight, volume of a liquid dosage form, quantity of tablets produced, time required for drug disintegration, and amount of active ingredient in a dosage form are some examples in pharmaceutics. For example, a medication concentration of 20 μg/mL is double that of 10 μg/mL.

The mean, median, standard deviation, variance, and range are typically used to express data on ratio scales. For example, the average tablet weight may be calculated in addition to determining the variation among individual tablets in a manufacturing batch. Ratio-scale data are usually employed in pharmacokinetics experiments, clinical trials and pharmaceutics quality assurance. Appropriate techniques include t-tests, analysis of variance, correlation, regression and other parametric tests, as long as the assumptions of the chosen technique are fulfilled.

1.1.2. Sources of Data in Pharmaceutical Research

Pharmaceutical research involves the systematic gathering of accurate and appropriate data. Data can be gathered through experimentation, clinical trials, manufacturing processes, medical records, literature sources, and many more. The data's source reveals how it was collected, how accurate it is, and even what techniques should be used to analyse it.

Primary and secondary data are the two categories into which data sources are often divided.

A. Primary Data

Primary data are fresh data which are collected firsthand by the researcher for a certain research objective. Primary data are usually collected using methods such as experimentations, observations, measurement, interview, questionnaire, and clinical studies.

Primary data are usually collected based on the aim of the research and therefore regarded as very relevant since they are collected with a view to answering the research question.

Examples in pharmaceutical research include

- Determination of dissolution rates of a new formulation of a tablet.
- Detection of changes in the blood pressure of patients under medication.
- Detection of the levels of a drug in the plasma sample.
- Adverse reactions to drugs are collected from the individuals undergoing clinical trials.
- Laboratory studies on drug stability.
- Hardness, friability and weight variation of tablets.

Advantages of primary data

- High relevance to the research objective.
- More control in the process of collecting data.
- Standardized method for the collection of information.
- The researchers can ensure the measurement of variables that need to be measured.
- Collection of data at the right time and under controlled conditions.

Limitations of primary data

- Data collection may take up much time.
- Data collection may be costly.
- Trained personnel and sophisticated tools may be needed.
- Ethical clearance may be necessary when conducting research on humans and animals.
- Mistakes can be made due to lack of standardization of data collection techniques.

B. Secondary Data

Data that has already been collected for other purposes by another person or organization is referred to as secondary data. Researchers use this data for their investigations.

Examples in pharmaceutical research include

- Published research articles.
- Government health reports.
- Clinical trial databases.
- Hospital records.
- Pharmaceutical company reports.
- Drug safety databases.

- Pharmacopoeias and regulatory documents.
- Previously completed research studies.

Advantages of secondary data

- Easily available in many cases.
- Less expensive than collecting new data.
- Saves time and effort.
- Useful for reviewing existing knowledge.
- Helps researchers identify research gaps and develop new hypotheses.

Limitations of secondary data

- The information may not be completely relevant to the research objective.
- The credibility of the information may be hard to ascertain.
- The information may not be complete and may be out of date.
- The methodology employed to gather the primary information may not be known.
- Some databases may be inaccessible.

Researchers must ensure the reliability, relevancy, accuracy, and timeliness of the secondary data before utilizing it.

C. Major Sources of Pharmaceutical Data

Pharmacological research relies on some significant sources of data. The sources offer information at various levels of drug development and application.

1. Clinical Trials

Clinical trials represent structured research trials designed to examine safety, efficacy, dosage, and therapeutic properties of drugs on humans. This source of pharmaceutical information is among the most valuable since it offers scientific facts about functioning and impact of the medicine. The data derived from clinical trials may involve patients' demographic data, medical background, doses and schedules of administering the medicine, its impact, laboratory results, adverse reactions, pharmacokinetics and pharmacodynamics measurements, compliance, and clinical outcomes. Such data are collected according to a prearranged research protocol.

Phases of clinical trials usually include four stages. In phase I, safety, tolerability, and pharmacokinetics are studied in a limited number of subjects. Phase II explores the preliminary efficacy and sets an appropriate dose of the medicine. Phase III confirms efficacy and safety on a bigger sample size. Finally, phase IV tests safety and effectiveness of the medicine after receiving

marketing authorization. The collected data assist in determining safety and efficacy of a medicine for the intended purpose.

2. Pharmacovigilance

Pharmacovigilance refers to scientific and practical activities that involve monitoring for, assessing, evaluating, and preventing adverse reactions or any other drug-related issues. This is a very significant data source in pharmacy since some of the adverse drug reactions may not have been noted during clinical trials and may come into light later on when a larger and more diverse patient population starts using the drug. The data are gathered from health care practitioners, patients, hospitals, clinics, pharmaceutical organizations, regulatory bodies, spontaneous adverse drug reactions reporting systems, and post-marketing studies.

Some of the important pharmacovigilance data include patient's age and gender, suspected medication, dose and route of administration, nature of the adverse reaction, time of occurrence, medical background, other medications being taken, and results of the reaction. Such data are analyzed by researchers and regulatory bodies to look for safety issues that had not been discovered before, evaluate any relationships between medicines and reactions, and take relevant safety measures.

3. Pharmaceutical Quality Control

The term pharmaceutical quality control can be defined as a process in which tests are done on raw materials, intermediate products, and final pharmaceutical products to check that they are compliant with pre-established criteria regarding identity, potency, purity, safety, and quality of pharmaceuticals. Quality control data are collected by doing laboratory analyses and inspections at various phases of the production process of pharmaceuticals. Data thus generated are used to ascertain compliance of a pharmaceutical product to its specifications and suitability for its intended purpose. Some examples of such data include assays for API, weight variation, hardness, friability, dissolution, disintegration time, pH of the solution, moisture content, microbial limit tests, stability, etc.

The analysis of quality control data is very important for detecting any problem in the manufacturing process, ensuring consistency of batches, and compliance of pharmaceuticals with the quality requirements. Weight variation test helps to know whether there is consistency of the amount of materials present in the tablet form and dissolution test checks whether a medicine is released from its delivery system effectively. Stability data are used to evaluate changes in product quality during its storage period.

4. Pharmacokinetic Studies

Pharmacokinetics is a branch of science concerned with the processes in the body of interest to a given drug. It involves studying the processes of absorption, distribution, metabolism, and excretion. This process is also abbreviated as ADME. Pharmacokinetic tests provide information on how drug concentrations vary in the body over time. The information obtained is used to understand the processes of movement of medicines in the body and to determine dosage regimens. Pharmacokinetic information may include plasma drug concentration, time to peak concentration (Tmax), peak plasma concentration (Cmax), area under concentration–time curve (AUC), half-life, clearance, volume of distribution, and urinary excretion of the drug.

The use of pharmacokinetic information helps in making conclusions about the absorption, distribution, elimination, and the period of staying of the drug in the body. The plasma drug concentration obtained after sampling at various time intervals can be plotted on a graph to obtain a concentration–time curve. The information is also used to make conclusions concerning the dosages, frequency, and route of administration. Comparison of drug formulations and evaluation of the bioavailability of drugs are some other applications of pharmacokinetic information.

5. Pharmacodynamic Studies

Pharmacodynamics refers to the study of the effects of a drug on the body. It involves the analysis of the correlation between the level of drug concentration and the subsequent effect produced by the drug. Pharmacodynamic studies provide data that enable scientists to determine the mode of action of the drug and the effectiveness of the drug. Such data includes reduction in blood pressure, change in blood glucose, reduction in pain intensity, change in heart rate, enzyme inhibition, therapeutic response, and biomarkers.

Pharmacodynamic data are useful in determining the degree of efficacy and variability of the effect that a drug produces. Scientists can perform pharmacodynamic studies to determine the effectiveness of an antihypertensive drug using the changes in blood pressure of a patient. This will involve measuring the changes in blood pressure after the administration of the antihypertensive drug to determine its therapeutic response. The same case applies to measuring changes in blood glucose to determine the efficacy of an antidiabetic drug.

6. Laboratory and Experimental Studies

Physical, chemical, biological and pharmacological properties of pharmaceutical substances are studied through lab and experimental research. Such research might include in vitro tests, animal testing, and other kinds of controlled lab investigations. It is one of the most valuable sources of pharmaceutical information because it contains details on drug properties, formulation

characteristics, biological activity, and possible dangers of use. Some examples of such research include drug solubility testing, stability testing, dissolution testing, enzyme inhibition testing, cell culture testing, antimicrobial testing, toxicity testing, and drug–excipient compatibility testing.

Laboratory information is helpful for many stages of pharmaceutical research such as drug discovery, formulation development, and preclinical evaluation. Solubility testing helps to learn how a drug reacts to various solvents. Dissolution testing allows evaluating how fast and well a drug dissolves in a dosage form. With the help of stability testing, it is possible to study the influence of environmental factors on drug properties. Through experimental analysis, scientists can choose optimal formulations and drugs and receive valuable information for further steps.

7. Hospital and Clinical Records

Hospital and clinical files include data about the patients, diagnosis, treatment, investigations from the laboratories, and the effects of the treatment. These records are useful sources of secondary data for pharmaceutics research because they contain information that is already available from routine healthcare practices. The information may be in the form of patient demographic data, prescription information, drug history, laboratory investigations, hospitalization and discharge, therapy results, and adverse reactions to drugs.

Hospital and clinical files may be applied in retrospective research, treatment results analysis, and in determining the utilization patterns of medicines. In the case of retrospective study, researchers can use prescription information for the purpose of establishing common medicines used. The same information can be used in the comparison of the efficiency of certain treatment approaches. Adverse reactions and health planning can also benefit from such information. Nevertheless, all the researchers have to adhere to certain ethical and privacy guidelines.

8. Pharmaceutical Manufacturing Data

Data from pharmaceutical manufacturing operations are created when pharmaceutical goods are manufactured. These data assist in controlling the manufacturing process, identifying deviations from the process and maintaining consistency in the batches. Some of the data generated from the process may include information about the quality of raw materials used, temperatures and pressures during manufacturing, time of mixing, granulation parameters, compression forces, fill volumes, batch yields, equipment performances, in-process test results, and quality test results of the final product. To ascertain whether the manufacturing process is fulfilling predetermined standards, this data is collected at different phases of the process.

These kinds of data are essential for quality control, manufacturing process validation, and pharmaceutical manufacturing process optimisation. For instance, monitoring compression forces

during manufacturing of tablets can help identify deviations in the process, which may have an effect on the strength or weight of the tablets. Batch yields and in-process test results can help in identifying production problems and maintaining consistency.

9. Regulatory and Government Sources

Regulatory and government sources contain information on pharmaceutical issues such as drugs, health of the general population, drug safety, and pharmaceutical standards. Regulatory and government sources include regulatory agencies, government health departments, and other organizations that gather pharmaceutical and healthcare information. This kind of information includes approved drugs information, drug safety alerts, drug recalls, health statistics at the national level, pharmaceutical standards, regulations concerning clinical trials, medicine utilization reports, and public health surveillance information. This type of information is beneficial when one wants to learn about regulatory framework and the environment of pharmaceutical research.

It is possible to use regulatory and governmental data in the process of studying medicine safety and in order to know the requirements of the regulatory system. For instance, drug recall notices can provide information on quality or safety issues concerning a particular pharmaceutical product, while medicine utilization reports issued by the government can assist in researching the pattern of using medicines.

10. Published Scientific Literature

Scientific literature published in literature consists of research articles, reviews, books, proceedings of conferences, and other scholarly literature. Scientific literature provides information on previous research, research methodologies, and new advances in the field of pharmaceutical sciences. Scientific literature serves as an important source of secondary data since it enables researchers to obtain existing scientific information without collecting all the information themselves. This type of literature can include experimental, clinical, statistical data, and discussions of pharmaceutical research topics.

Scientific literature is used by researchers to review previous knowledge, find gaps in the research area, compare results, formulate research hypotheses, choose research methodology, and reach scientific conclusions. For instance, when formulating a new drug preparation, a researcher needs to review previous research to become familiar with physicochemical characteristics and testing methods. However, any published information needs to be analyzed from the standpoint of its scientific value, relevance, accuracy, and credibility before conducting research.

D. Importance of Data Sources in Pharmaceutical Research

Data from various sources is involved in different steps of drug development. The clinical trial gives data on safety and efficacy; pharmacovigilance gives data on drug safety; quality control gives data on quality of product; and pharmacokinetics and pharmacodynamics give data on how the drug works. Data from manufacturing and regulation provides consistency and compliance.

The reliability of the source of data assists the researcher to:

- Develop safe and effective medicines.
- Improve pharmaceutical formulations.
- Monitor adverse drug reactions.
- Maintain product quality.
- Evaluate treatment outcomes.
- Support regulatory approval.
- Make evidence-based decisions.
- Improve patient care.

The pharmaceutical data is what makes up the basis of scientific research within the field of pharmaceuticals. Knowledge of the many kinds of data and the scales used to measure them helps scientists select the appropriate statistical tests and accurately interpret the results. Understanding primary and secondary data sources guarantees that trustworthy data can be gathered and assessed in domains like clinical trials, pharmacovigilance, quality control, and pharmacokinetics.

1.2. MEASURES OF CENTRAL TENDENCY

A statistical method for describing the average or middle value of a set of data is called a measure of central tendency. One value that represents the trend of the observed values is provided by this method. Large data sets can be condensed into a single representative value using measures of central tendency, making it easier to understand the observations. Measures of central tendency are used in pharmacy practice to characterise patient characteristics, medication concentrations, tablet weight, drug dissolving time, test laboratory measurements, and treatment outcomes, among other things. The mean, median, and mode are the three primary indicators of central tendency.

Measures of central tendency are crucial because they make data interpretation simpler. For example, the average drug concentration level in a group of patients can be ascertained using measurements of central tendency. When some patients exhibit exceptionally high concentrations, the median is a better measure of central tendency than the mean. When identifying the most prevalent category, such

as the typical dosage form or typical adverse medication reactions, mode can also be utilised. However, the measure of central tendency must be used in conjunction with the measures of dispersion because it does not fully characterise a set of data.

In the discipline of pharmacy, measures of central tendency are particularly crucial since the researcher is always working with a large number of observations that must be succinctly reported. Hundreds of measurements of blood pressure, medication concentration, or efficacy are typically used in clinical studies. It is difficult to understand the data trend if all the observations are given. When comparing treatment groups, analysing experiment results, and presenting the results in research reports, the use of an appropriate measure of central tendency provides an easily comprehensible data trend. For example, the typical effectiveness of a certain medication can be described by the mean reduction in blood pressure, but when some people need a lengthy course of treatment, the median treatment time provides a more accurate representation.

The type of data, measurement scale, distribution of observations, and presence of anomalous observations all influence the choice of a suitable measure. The mode aids in determining the most prevalent value or category, the median performs well with skewed data, and the mean is the best metric for numerical data with balanced observations. Several measures of central tendency may be presented in certain research to provide a better understanding of the data trend. For example, reporting both mean and median concentrations of a drug helps to determine whether the mean value is significantly affected by unusual observations.

1.2.1. Mean, Median and Mode

A set of data's central or average value can be described using three different types of measurements of central tendency: mean, median, and mode. To find the mean, tally up all of the observations, then divide the total by the total number of observations. Although outliers may have an impact, it is suitable for numerical data with an equal distribution. The value that results from sorting all of the data is called the median, and it works well for skewed distributions or observations with extremely high or low values. The mode, which can be used with nominal, ordinal, and quantitative data, is the most common observation in the data.

A. Mean

By dividing the total number of observations by their sum, the arithmetic average is computed. The arithmetic average, also known as the mean, is one of the most widely used measures of central tendency since it accounts for every observation in the dataset. The mean is especially helpful when dealing with numerical observations that are more or less evenly distributed and have no outliers.

The arithmetic mean is calculated as:

$$\bar{x} = \frac{\sum x}{n}$$

where:

- $\bar{x}$= arithmetic mean
- $\sum x$= sum of all observations
- n= number of observations

For example, suppose the drug concentrations measured in five blood samples are 10, 12, 14, 16, and 18 µg/mL.

$$\bar{x} = \frac{10 + 12 + 14 + 16 + 18}{5}$$

$$= \frac{70}{5} = 14 \text{ µg/mL}$$

Hence, the mean drug concentration is 14 µg/mL.

The above computation demonstrates that the sum of the concentrations of the five samples is 70 µg/mL and, therefore, dividing this quantity by five results in an average concentration of 14 µg/mL. The mean does not imply that each sample has a concentration of 14 µg/mL. It just offers one summary figure representing the average of all the observations.

Mean is extensively used in pharmaceutical research as it allows summarizing the following quantitative variables: the weight of tablets, drug concentration, blood pressure, dissolution time, assays, etc. For instance, the mean tablet weight in a manufacturing batch could be used as the average mass of material in each tablet. Another example could be the mean blood pressure of patients as the average blood pressure in a certain study group.

Moreover, means are helpful when comparing various groups. Researchers can compare, for example, the mean drug concentration among patients who received two different formulations or the mean dissolution time of tablets in two manufacturing batches. Nonetheless, it is necessary to pay attention to other parameters as well.

One of the major drawbacks of using the mean is its sensitivity to extreme values. Let us look again at the drug concentrations:

10, 12, 14, 16, and 100 µg/mL

$$\bar{x} = \frac{10 + 12 + 14 + 16 + 100}{5}$$

$$= \frac{152}{5} = 30.4 \text{ µg/mL}$$

While many of the observations range from 10 to 16 µg/mL, the outlier of 100 µg/mL results in an average value of 30.4 µg/mL. This observation is not necessarily representative of the concentration for most of the samples. Thus, for cases where outliers are present in the data, the median may be a better measure.

B. Median

The midway value in a set of data, either in ascending or descending order, is called the median. With roughly half of the numbers falling below the median and the other half rising above it, the median divides the organised data into two halves. Because it is unaffected by extremely high or extremely low values, the median is a crucial metric when dealing with skewed data.

The data is first sorted in either ascending or descending order in order to find the median. If the number of numbers in the data set is odd, the median will be the middle number; if the number of numbers is even, the median is calculated by averaging the two middle numbers.

For instance, let us take the dissolution time of five tablets:

8, 10, 12, 14, 16 minutes

The middle observation is **12 minutes**.

Therefore:

$$\text{Median} = 12 \text{ minutes}$$

There are two observations below 12 and two observations above 12, making 12 minutes the central value of the ordered dataset.

For an even number of observations, consider:

8, 10, 12, 14, 16, 18 minutes

The two middle values are 12 and 14.

$$\text{Median} = \frac{12 + 14}{2} = 13 \text{ minutes}$$

Hence, the median dissolution time is 13 minutes.

Median is usually preferred where the data in pharmaceutical and clinical practice do not follow normal distribution or there are outlying values. Suppose that drug concentrations in the vast majority of patients are between 10 and 20 µg/mL, while drug concentration in one patient is 100 µg/mL, then in such case median would be a better estimate of concentration compared to the mean.

Median is especially helpful when dealing with skewed data. Time of treatment response, time of stay at the hospital, some pharmacokinetics measures may demonstrate significant variability of

values, and there is a minority of values much higher than the vast majority. In this case, median can give an idea about the value of central tendency.

Median is also applicable for ordinal data such as severity levels of disease or patient satisfaction with service if there is some ordering between these ordinal levels. But median cannot imply that distances between ordinal categories are the same.

C. Mode

The mode refers to the most recurring number or category in the data. The mode applies to nominal, ordinal, and quantitative data. The data set can have one mode, several modes, or no mode at all.

Consider the following weights of tablets:

250, 250, 251, 252, 250, 253, 251 mg

The value **250 mg** occurs most frequently.

Therefore:

$$\text{Mode} = 250 \text{ mg}$$

The mode is helpful when finding the most frequent observation or classification. In clinical pharmacology studies, one can use the mode to find the most frequently observed dosage form, adverse drug reaction, weight of the tablets, treatment category, etc.

For instance, in a set of data where the dosage forms are:

Tablet, Capsule, Tablet, Injection, Tablet, Capsule

The mode in this case is Tablet, since it occurs more frequently than others.

As opposed to the mean, the mode is not impacted by very large numbers. This means that even if there is an exceptionally high number in a dataset like 250, 250, 250, 251, and 300 mg, the mode will still be 250 mg.

In some cases, the mode fails to give a unique central number. For instance, a set of data can be bimodal; for instance:

10, 10, 12, 12, 14

The mode of data can be either 10 or 12. A dataset where all the observations happen only once lacks a mode. In essence, the mode ought to be understood within the context of the dataset.

1.2.2. Calculation and Interpretation of Central Tendency

It is necessary to select the appropriate measure and interpret its meaning within the context of the data set in order to calculate and interpret measures of central tendency. The mode identifies the observation that appears most frequently in the data set, the median identifies the center position of

the observations when sorted in order, and the mean takes into account all of the observations. Numerous factors influence the choice of central tendency measure.

Consider the following drug concentration values:

10, 12, 14, 16, and 18 µg/mL

The mean is:

$$\bar{x} = \frac{10 + 12 + 14 + 16 + 18}{5} = 14 \text{ µg/mL}$$

Since the observations have previously been sorted in ascending order, the middle value is:

$$\text{Median} = 14 \text{ µg/mL}$$

No value occurs more than once, so:

$$\text{Mode} = \text{No mode}$$

The mean and median values are both 14 µg/mL, while the mode is absent. It is evident that different measures of central tendencies yield distinct information for the same set of data.

In choosing the right measure, it is crucial to consider the distribution of data. In case of symmetry of observations, the two measures will be similar. In cases where there is skewness or outliers, the measures may vary considerably.

Consider the following data on tablet dissolution time:

10, 11, 12, 13, and 30 minutes

The mean is:

$$\bar{x} = \frac{10 + 11 + 12 + 13 + 30}{5}$$

$$= \frac{76}{5} = 15.2 \text{ minutes}$$

The median is:

$$\text{Median} = 12 \text{ minutes}$$

There is no mode as each value appears only once.

Here, the mean is 15.2 minutes, whereas the median is 12 minutes. It happens because of the unusual value of 30 minutes, which raises the value of the mean. As most of the values are 10 – 13 minutes, the median may be more representative in terms of the dissolution time of this drug.

The distribution type may also be revealed by the relationship between the mean, median, and mode. All three statistics could be equal or almost equivalent for a symmetrical distribution. The mean may be higher than the median in a positively skewed distribution because some anomalous values raise the average. The mean is typically lower than the median in a negatively skewed distribution due to

a few exceptionally low values. The researcher can comprehend the pattern of the pharmaceutical data by looking at the relationships between the mean, median, and mode.

For instance, a fairly balanced distribution is shown when the mean and median medication concentrations in clinical trials are equal. However, in the case when the mean drug concentration is much higher than the median, it indicates that some patients have unusually high drug concentrations, which is very important information. Extreme concentrations affect the evaluation of the drug exposure and its safety.

The measures of central tendency should be calculated based on the appropriate units. Thus, the units of measurement for the mean and median should correspond to the original data. For example, when the units of measurement for the drug concentration are expressed in μg/mL, the mean and median concentrations should be expressed in μg/mL as well.

In the course of pharmaceutical research, central tendency calculations are usually carried out via specialized software or spreadsheet applications due to large volumes of observation. Yet, knowledge of the underlying formulae is still essential since a researcher must ensure that he or she uses the appropriate measure and interprets its outcome correctly. Computer programs enable calculating the mean, median, and mode, yet cannot evaluate whether one or another measure is scientific from the point of view of the given research.

As it was mentioned above, measures of central tendency are extensively used in pharmaceutical research. They are helpful for analyzing patient demographics and drug properties. Thus, the mean measure can be used for describing the average drug concentration in a clinical study, the median – for summarizing patients' response times to treatment, and the mode – for identifying the most frequent adverse drug reaction.

Moreover, central tendency measures assist in the work related to quality control in the sphere of pharmaceutical industry and clinical decision-making process. For instance, the mean tablet weight can serve as an indicator of batch consistency whereas the median – when data are distorted by some extremes. Nevertheless, researchers should not focus exclusively on the measure of central tendency; measures of dispersion and data distribution must also be considered.

For instance, there can be two batches of medicines whose mean tablet weight is the same. However, one of the batches can have tablets which vary slightly, and the other – significantly. The difference in the variation indicates the difference in quality and consistency of these two batches of medicine although their means are the same. Thus, measures of central tendency must be analyzed in connection with other measures such as standard deviation, variance, and range.

There can be two treatment groups whose median response time is equal, but the distribution of response time in each of them is quite different. This illustrates the necessity to analyze both the central tendency and the variability of data.

1.3. MEASURES OF DISPERSION AND DISTRIBUTION SHAPE

Statistical methods used to examine the variability, spread, and shape of observations within the data set are called dispersion and distribution shape measures. The data set's center value can be found using measures of central tendency like mean, median, and mode, but they don't reveal anything about how the observations are distributed about the center. Even though the means of two sets of data are comparable, their levels of variation can differ significantly. Range, variance, and standard deviation are examples of dispersion metrics that provide information about how variable the observations are. Measures of distribution shape, such as skewness, provide information about the long tail at either end of the values or the symmetry of the distribution. Measures of dispersion and distribution shape are very useful for examining different aspects of biology and pharmacological research.

1.3.1. Range, Variance and Standard Deviation

Measures of dispersion that provide insight into how the observations within a data collection vary include the range, variance, and standard deviation. The range provides a simple way to gauge the overall spread of observations and is defined as the difference between the data set's maximum and minimum values. The variance accounts for every observation in the data set and is calculated as the average of the square deviations of the observations from the mean. The standard deviation, which provides information on variation using the same units of measurement as the observations, is the square root of the variance. These measures are applied in pharmaceutical studies to determine variation in drug concentrations, weight of tablets, time of dissolution, assays, and blood pressure, among others.

A. Range

The simplest method for figuring out the dispersion is range. The difference between a data set's maximum and minimum values is referred to as the range. Range provides a quick indication of how widely the values are dispersed. A smaller range indicates that the data set's values are near to one another, and vice versa.

The range can be computed using the following formula:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

For example, suppose the drug concentrations measured in five blood samples are:

$$10, 12, 14, 16, 18\ \mu g/mL$$

The maximum value is 18 µg/mL and the minimum value is 10 µg/mL.

$$\text{Range} = 18 - 10 = 8\ \mu g/mL$$

The range of drug concentration is 8 µg/mL. This indicates that the difference between the highest and the lowest concentration measurement is 8 µg/mL.

The range is widely used in pharmaceutical research as a method to get an immediate idea about variation in the measurements. For instance, researchers can report the minimum and maximum weights of tablets in order to find out the range of variations in their weight. In addition, the range of dissolution time is used to reveal the difference between the fastest and the slowest dissolving tablets. Similarly, the range of blood pressure, blood glucose, drug concentration or any other response in clinical trials helps to measure the extent of variation between patients.

The range's ease of use and interpretation is one of its main benefits. The highest and minimum values found in the sample are all that are required. Additionally, because researchers must discover those abnormally significant fluctuations in the data, the range is helpful for analysing initial data and performing quality control. For instance, a wide variation in tablet weight suggests production process variability and necessitates additional research.

The range has one significant flaw: it is solely based on two extreme values and ignores all other values in the sample. Because of this, the range is highly susceptible to abnormally high or low sample values. Consider the situation of medication concentrations:

$$10, 12, 14, 16, 18\ \mu g/mL$$

The range is:

$$18 - 10 = 8\ \mu g/mL$$

If an additional observation of 100 µg/mL is included, the range becomes:

$$100 - 10 = 90\ \mu g/mL$$

Although most measurements stay within the range of 10 to 18 µg/mL, there is an increased range due to the extremely high value. Thus, the range can be used as a very simple measure of variability; however, it should always be considered together with other measures of variability.

B. Variance

The degree to which the values in a data collection depart from the mean is indicated by the variance, a measure of dispersion. Finding the difference between each value and the mean, squaring the results, and calculating the average of the squared results are the steps involved in calculating the

variance. We can prevent the cancelling of the positive and negative deviations by squaring the deviations.

Variance for the entire population is calculated using the following formula:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

where:

- σ^2= Population variance
- x= Individual observation
- μ= Population mean
- N= Number of observations

For a sample, the formula is:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

where:

- s^2= Sample variance
- x= Individual observation
- $\bar{x}$= Sample mean
- n= Number of observations

Use of n-1in the calculation of variance is an unbiased estimator of population variance given the assumption of random sampling.

As an illustration, assume that the following are the weights of five tablets:

$$248, 249, 250, 251, 252 \text{ mg}$$

The mean tablet weight is:

$$\bar{x} = \frac{248 + 249 + 250 + 251 + 252}{5}$$

$$\bar{x} = \frac{1250}{5} = 250 \text{ mg}$$

The deviations from the mean are:

$$-2, \ -1, 0, 1, 2$$

The squared deviations are:

$$4, 1, 0, 1, 4$$

The sum of squared deviations is:

$$4 + 1 + 0 + 1 + 4 = 10$$

Therefore, if these five observations are treated as the complete population:

$$\sigma^2 = \frac{10}{5} = 2 \text{ mg}^2$$

Thus, the population variance is 2 mg².

If the five tablets are instead considered a sample from a larger production process, the sample variance would be:

$$s^2 = \frac{10}{5 - 1}$$

$$s^2 = \frac{10}{4} = 2.5 \text{ mg}^2$$

This difference between population variance and sample variance plays an essential role in pharmaceutical research since researchers operate not with the whole population but with a sample of observations.

Variance is helpful in its ability to include all the observations in calculations instead of taking into account only the maximal and minimal observations. It is widely used in assessing the variability of drug concentration, tablet weight, dissolution time, assays, blood pressure, and many other numerical data. For instance, two different batches of tablets can have the equal average weight, but the one which has the lower variance will be characterized by more constant weight of tablets.

Variance is significant not only for statistical analysis and research methodology because it is the basis for such important methods as ANOVA, regression analysis, probability distributions, etc. In clinical research, variance allows assessing the extent to which patients' reactions differ from the average reaction to a certain treatment. The higher variance corresponds to the higher dispersion and vice versa.

Despite the fact that variance is very useful in mathematical terms, there is some problem with its interpretability because it uses squared units of measurements. For example, if the drug concentration is measured in µg/mL, then the units of variance will be (µg/mL) ^2. This drawback explains why standard deviation is more preferable than variance.

C. Standard Deviation

One of the most important dispersion metrics utilised by academics worldwide is the standard deviation. The deviation of data from the mean is displayed by this metric. The variance's square root is known as the standard deviation. The standard deviation has the advantage of being measured in the same units as the original data.

The population standard deviation can be calculated using the following formula:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

The formula for sample standard deviation is:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

where:

- σ= Population standard deviation
- s= Sample standard deviation
- x= Individual observation
- μ= Population mean
- $\bar{x}$= Sample mean
- N= Population size
- n= Sample size

Using the previous example of tablet weights:

$$248, 249, 250, 251, 252 \text{ mg}$$

The population variance was:

$$\sigma^2 = 2 \text{ mg}^2$$

Therefore:

$$\sigma = \sqrt{2}$$

$$\sigma \approx 1.41 \text{ mg}$$

Therefore, the population standard deviation is approximately 1.41 mg.

It suggests that the weights of the tablets have less variation from the mean weight of 250 mg.

When describing the consistency and variability in pharmaceutical data, standard deviation is a very useful tool. For example, the weights of the pills tend to be near the average when the mean weight is 500 mg and the standard deviation is 2 mg. The tablets are more widely distributed if the standard deviation is 20 mg. Accordingly, a lower standard deviation indicates greater consistency while a higher standard deviation indicates greater unpredictability.

Standard deviation can also be used for evaluating the variation in tablet weight, the amount of drugs, hardness, thickness, dissolution time, and assay of pharmaceutical products. The low standard

deviation in tablet weight means that the process of manufacture results in tablets that have similar weight. In case of high standard deviation, it may indicate the need to investigate the possible reasons for the variability like mixing, compression, filling, and others.

Standard deviation is very popular in clinical and biological researches. Scientists usually provide the mean and the standard deviation for the measurement such as blood pressure, body weight, concentration of the drug, laboratory measurements, and treatment responses. For instance:

$$120 \pm 8 \text{ mmHg}$$

The mean in this instance is 120 mmHg, while the standard deviation is 8 mmHg. The dispersion of data around the mean value is described by the standard deviation.

According to the empirical rule, standard deviation can also be utilised when data have an approximate normal distribution. In this instance, 68% of observations fall within one standard deviation of the mean, 95% fall within two standard deviations, and 99.7% fall within three standard deviations.

For instance, if the standard deviation is 5 μg/mL and the mean drug concentration is 50 μg/mL:

- About 68% of observations may fall into the range 45–55 μg/mL.
- About 95% of observations may fall into the range 40–60 μg/mL.
- About 99.7% of observations may fall into the range 35–65 μg/mL.

This result is only valid under the assumption that the data distribution is roughly normal.

If measurements are fairly comparable and stated in the same units, the standard deviation makes it possible to compare variability between groups. For instance, if the mean drug concentration of two formulations is similar, but one formulation has a lower standard deviation, it indicates that the amounts recorded are more consistent. Nevertheless, sample size, research methodology, measuring techniques, and the pharmacological significance of the outcome should all be considered in this conclusion.

The standard deviation should always be taken into account in relation to the data distribution and mean value. Without considering the measurement's scale, the standard deviation cannot be classified as great or little. For example, standard deviation of 2 mg will be large if the tablet is 10 mg but it will be very small if the tablet is 1,000 mg.

1.3.2. Skewness and Distribution of Biological Measurements

Skewness is a term used in statistics to explain how unbalanced a data set is. Skewness describes how uneven and the direction of unevenness of a data set. When there is an unequal distribution of data on each side of the distribution center, it is referred to as skewness. A distribution is considered

symmetric if it is approximately equal on both sides of the center. A symmetric distribution has comparable means, medians, and modes. A skewed distribution, on the other hand, has an uneven distribution with observations closer to one side and a long tail in the opposite way.

Skewness is necessary as most of the biological and pharmaceutical measurements are not symmetric. Measurements like drug concentration, time taken to react, lengths of hospital stay, enzymes levels, antibodies titter levels, among others, can be positively or negatively skewed.

A. Symmetrical Distribution

Symmetrical distribution refers to the type of distribution where the distribution of observations from both sides of the central value has similarities in nature. For a symmetrical distribution that is perfectly unimodal in nature, the mean, median, and mode have identical values.

There is no tail on either the left or the right side of the observations, which are centred around a single value.

For instance, the observations include:

$$8, 10, 12, 14, 16$$

The mean is:

$$\bar{x} = \frac{8 + 10 + 12 + 14 + 16}{5} = 12$$

The median is also 12, and observations are about equally balanced about this central point. Consequently, the distribution can be said to have a symmetrical nature.

The mean is a suitable indicator of central tendency in a symmetrical distribution, particularly when there are no notable extreme observations. The variation around the mean can also be described using certain metrics, such standard deviation.

The occurrence of symmetrical distribution may be observed in biological and pharmaceutical data where observations are affected by various small independent factors. This is evident where some of the observations done under experimental conditions have a symmetrical appearance. In addition, certain features in relatively homogenous populations may also have an approximately normal distribution.

It is important for pharmaceutical and biological data not to be taken for granted as being normally distributed. Data quality assessment is an important first step for scientists when deciding which statistical methods to use. There will be a combination of numerical and statistical methods used in distribution evaluation.

B. Positively Skewed Distribution

Positive skewness, also known as right skewness, indicates a tail towards larger numbers in the distribution. Most of the data are clustered towards the lower or middle section of the graph while there are few observations with large numbers on the right-hand side.

The general order of magnitude of the measures of central tendency is:

$$\text{Mode} < \text{Median} < \text{Mean}$$

For example, consider the following treatment response times:

$$2, 3, 3, 4, 5, 15 \text{ days}$$

Most patients respond within 2 to 5 days, but one patient requires 15 days. The unusually high observation extends the distribution toward the higher values and increases the mean.

The mean is:

$$\bar{x} = \frac{2+3+3+4+5+15}{6} = \frac{32}{6} \approx 5.33 \text{ days}$$

The median is:

$$\text{Median} = \frac{3+4}{2} = 3.5 \text{ days}$$

The impact of the high outlier is reflected in the difference between the mean and the median. Since the median is unaffected by the extreme point, it may provide a more accurate picture of the typical response time.

Positive skewness may arise in various pharmaceutical and clinical indicators. Concentration of the medication, the time of response to the therapy, hospitalization period, waiting time, and certain biochemical parameters may exhibit positive skewness. For instance, many patients might have low or moderate drug concentration levels while only a few will have high levels of concentration.

It might be more preferable for the researchers to use the median and interquartile range rather than only the mean and standard deviation when dealing with positively skewed data. For some analysis purposes, a logarithmic transformation may occasionally be employed to reduce skewness. The type of the transformation should be determined depending on the nature of the research.

C. Negatively Skewed Distribution

When the distribution's tail extends to the left, it is said to be negatively skewed. This indicates that while the observations with very low values extend the tail to the left, the majority of the data is concentrated on the higher side of the distribution.

The central tendency measures generally have the following relationship:

$$\text{Mean} < \text{Median} < \text{Mode}$$

For instance, assume that the satisfaction level of patients with the medicine is rated using a scale of 1 to 10. If most of the patients provide high scores, such as 8, 9, and 10, but some provide very low scores, then the distribution could have a long tail at the lower scores, and hence the distribution would have negative skewness.

Negative skewness could result if there exists an upper boundary value for a variable and most observations are near the boundary value. For instance, if most students obtain high examination scores, and very few obtain low scores, then examination scores are likely to be negatively skewed. Also, if the majority of the patients improve significantly and only a few slightly, then the clinical improvement score will be negatively skewed.

The mean may be lower than the median because it is drawn to the lower end of the distribution due to a few low observations. The median might be a more accurate indicator of a typical observation in such circumstances.

When dealing with negatively skewed data, one needs to consider this property when determining which statistical method to use. One needs to analyze the nature of the data first before determining which method is suitable.

D. Measurement of Skewness

The degree and kind of a distribution's asymmetry are measured by skewness. It can be quantified both quantitatively and using methods of distribution analysis.

A frequently used measure of skewness using Pearson's approach is:

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

When the mode is difficult to determine, another form is:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

The general interpretation is:

- **Skewness close to 0:** approximately symmetrical distribution
- **Positive skewness:** right-skewed distribution
- **Negative skewness:** left-skewed distribution
- **Larger absolute value:** greater degree of asymmetry

For example, if the calculated skewness is:

$$+1.5$$

the distribution shows positive asymmetry, with a relatively longer tail toward higher values.

If:

$$-1.2$$

the distribution shows negative asymmetry, with a relatively longer tail toward lower values.

If:

$$0.05$$

the distribution is roughly symmetrical based on this measure.

The concept of skewness is important in pharmaceutical studies since the shape of a distribution may impact the choice and interpretation of various statistical procedures. For instance, if the drug concentration or reaction time is highly skewed, it may affect how the data will be summarized and analyzed.

Skewness cannot be used in an unthinking manner. The usefulness of this measure of symmetry largely depends on factors such as the number of observations, degree of asymmetry, purpose of the study, scale of measurement, and statistical procedure under consideration.

Consequently, researchers must combine the assessment of skewness with other features of the data. Knowing whether the pharmacological data are symmetrical, positively skewed, or negatively skewed is crucial for selecting appropriate central tendency and variability metrics.

E. Distribution of Biological Measurements

Often biological measurements are variable due to the impact of genetic makeup, age, sex, lifestyle, environment, illness, medication, error, or individual physiology. Thus, biological data may exhibit a normal, skewed, multimodal, or arbitrary distribution pattern.

There are some biological measurements such as height, weight, blood pressure, or certain laboratory values that can be expected to have a distribution that approximates a normal one in a sufficiently homogeneous population. The distribution pattern may change if the population includes individuals that differ in age, disease condition, treatment history, or other factors.

Frequently, the drug concentration data exhibits skewness since there are inter-individual variations in absorption, distribution, metabolism, and excretion. Some individuals may have slow absorption or elimination, which leads to increased concentrations of drugs. There may be positively-skewed treatment response times since the majority of patients react quickly while several patients need longer time for reaction.

Biological measurements may be impacted by the detection limits and boundaries of the measurements. Laboratory values may have lower detection limits, which means that many of the observations will cluster at zero with only a few observations having very high values. On the

contrary, measurements may have upper physiological boundaries and will be negatively skewed because many observations will be clustered at the maximum value.

When deciding on statistical tests, it is essential to consider the distributional form of biological measures. The data or model residuals are assumed to have a normal distribution according to parametric testing assumptions. Use data transformations, robust statistics, or non-parametric tests if the data exhibit strong skewness. When the data is skew, the median and interquartile range provide a more accurate picture than the mean and standard deviation, which are more suitable when the data is almost symmetrical.

It is unrealistic to expect all biological measurements to follow a normal distribution. Histograms, boxplots, normal probability plots, and skewness measurements are useful tools for analysing distributions. The sample size, research design, existence of outliers, and biological significance of measurements all influence the statistical analysis option.

The knowledge about skewness and distribution shapes gives researchers the opportunity to describe data properly, detect outlying observations, apply appropriate statistical methods, and avoid erroneous conclusions. This is especially relevant for pharmaceutical research where the distribution of drug concentrations, treatment response, laboratory values, and safety measurements has to be considered.

1.4. DESCRIPTIVE ANALYSIS USING PYTHON

Description of Analysis Descriptive analysis is an approach to the systematic presentation of a set of data. This method assists in understanding the major features of the dataset without making any forecasts or conclusions that go beyond the presented information. Descriptive analysis can be applied in pharmaceutical studies to summarize data on drug concentration, tablets weight, dissolution time, blood pressure, age, reaction of patients, and test results. Major descriptive statistics include mean, median, mode, minimum, maximum, range, variance, standard deviation, quartile, and percentage. Powerful and efficient Python software allows researchers to perform calculations of these indicators using NumPy and Pandas packages. Using these instruments, researchers can analyze datasets, calculate statistical summaries, find missing observations, detect outliers, and organize the results of the calculations.

The use of Python to conduct descriptive analysis is especially beneficial since a pharmaceutical dataset can contain several hundred or thousand observations obtained in the course of conducting clinical trials, analyzing the quality control laboratories, pharmacokinetics, and hospital data. Calculation of this quantity of observations manually is time-consuming and can lead to mistakes.

With the help of Python, the researcher can automate the process and repeat the same type of analysis in case new observations appear. Besides, Python allows the researcher to clean, tabulate, visualize and analyze data statistically. However, the use of Python cannot replace the role of the researcher who should choose proper variables, check the dataset, and use the results for interpretation.

1.4.1. Data Summarization with NumPy and Pandas

NumPy and Pandas are two common Python libraries used for numerical and data analysis. The NumPy library is short for Numerical Python and is primarily used for doing numerical computation and array manipulation. Functions for finding the mean, median, standard deviation, variance, minimum, maximum, percentiles, and many more are present in the NumPy library. The Pandas library is used for doing operations on tabular structured data. Pandas' library has two important data structures named Series and Data Frame. The Series data structure contains one column of data, whereas the Data Frame data structure is a tabular data set consisting of rows and columns.

A. Data Summarization Using NumPy

NumPy is one of the important libraries in Python programming language that is very effective especially when the pharmaceutical observations are in the form of a numeric array. An array is defined as a set of values arranged in a manner that can be manipulated through mathematical and statistical operations. For instance, assume that there are five blood samples and their drug concentrations have been measured:

```python
import numpy as np

concentrations = np.array([10, 12, 14, 16, 18])
```

In this case, the array includes five data values on concentrations of drugs in units of µg/mL. After inputting data into the array, NumPy offers a number of functions to compute various descriptive statistics of the data without doing calculations by hand.

For example, to calculate the mean value or arithmetic average:

```python
np.mean(concentrations)
```

The output is:

```
14.0
```

Hence, the mean concentration of the drug is 14 µg/mL. When the data is reasonably symmetric and not significantly impacted by outliers, the mean which provides a measure of central tendency for the observations is commonly employed.

The median is determined by:

```
Python

np.median(concentrations)
```

The output is:

```
14.0
```

Median = 14 µg/mL, as the observations have symmetry about the middle value. Median is used in cases where data are skewed or have any outliers.

The Minimum and Maximum can be calculated as:

```
Python

np.min(concentrations)
np.max(concentrations)
```

The minimum value is 10 µg/mL, while the maximum value is 18 µg/mL. These values can be used to calculate the range:

```
Python

np.max(concentrations) - np.min(concentrations)
```

The output is:

```
8
```

Therefore, the range of drug concentration is 8 µg/mL. The range is a metric that displays the entire range of values, from the lowest to the highest.

Variances and standard deviations are also computed by NumPy, and the measures give an insight into the variation in values from the mean. Variance is computed by:

```
Python

np.var(concentrations)
```

One can compute the population standard deviation using:

```python
np.std(concentrations)
```

For the provided data set, the population standard deviation is roughly 2.83 µg/mL, and the population variance is 8 $\mu g^2/mL^2$. The units of measurement used for the variance are squared units, but for the standard deviation, they are the same as those used for the observations themselves. Thus, the standard deviation is more interpretable in pharmaceutical studies.

Moreover, NumPy includes the percentile () function which helps analyze the positions of observations:

```python
np.percentile(concentrations, 25)
np.percentile(concentrations, 50)
np.percentile(concentrations, 75)
```

25% of the data falls below the 25th percentile. The median is equal to the 50th percentile, while 75% of the data falls below the 75th percentile. In pharmaceutical statistics, percentiles are used to characterise data value variability, particularly when quartiles and interquartile ranges are utilised by the researchers.

When the data values represent a sample of a larger population, the sample variance and standard deviation can be computed using ddof=1:

```python
np.var(concentrations, ddof=1)
np.std(concentrations, ddof=1)
```

The ddof = 1 option makes a degrees-of-freedom correction and gives the frequently applied sample variance and sample standard deviation. The difference is significant since when researchers need to infer something about a more extended population, they tend to rely on sample statistics.

NumPy is highly helpful when it comes to performing numerical computations quickly and efficiently. It can work with large arrays of numbers effectively and lets researchers perform mathematical operations on multiple observations simultaneously. Still, NumPy is oriented towards numerical computations. When it comes to column names, handling of missing values, categorical values, different datatypes, and structured datasets, Pandas is far more convenient.

B. Data Summarization Using Pandas

"Pandas" is a popular library in Python that is used to structure, manipulate, and analyze data. Pandas is useful in pharmacology because the data available often contains multiple forms of data under various columns. For instance, the data from a clinical/pharmacological study may involve patient ID number, age, gender, treatment group, drug concentration, blood pressure, dosage form, and treatment effect.

This kind of data can be arranged in a Pandas Data Frame, which involves rows and columns where each row is an observation/participant and each column is a variable.

For example:

```python
import pandas as pd

data = {
    "Patient_ID": [1, 2, 3, 4, 5],
    "Drug_Concentration": [10, 12, 14, 16, 18],
    "Dissolution_Time": [8, 10, 12, 14, 16]
}

df = pd.DataFrame(data)
```

The DataFrame can be displayed using:

```python
print(df)
```

There are individual data observations represented in rows while variables like Patient ID, Drug Concentration, and Dissolution Time are represented in columns. This table form makes Pandas appropriate for pharmaceutical datasets.

Exploring the Structure of a Dataset

The head () function displays the first few rows:

```python
df.head()
```

This is immediately useful upon import or creation of the data set, as it enables the researcher to verify whether the observations and column labels have been imported properly.

The tail () command shows the last few rows:

```python
df.tail()
```

It can help check the tail end of the data set and ensure that all data have been loaded correctly.

The info () method gives information on the structure of the Data Frame:

```python
df.info()
```

The table usually contains the total number of entries, column labels, data type, and number of non-null entries. This is critical because it allows for the detection of missing entries and determines if the variables have the correct data types.

Age should normally be numeric, while treatment group should be categorical. Detecting improper data type in the early stage can help avoid future problems with statistical analysis.

Descriptive Statistics Using describe ()

The describe () function is one of the most useful functions for obtaining a quick descriptive summary of numerical variables:

```python
df.describe()
```

For numerical columns, the summary generally includes:

- Count of observations
- Mean
- Standard deviation
- Minimum
- 25th percentile
- 50th percentile
- 75th percentile
- Maximum

For instance, the description of Drug_Concentration may show that there are five measurements, the average is 14 µg/mL, the standard deviation associated with it, the lowest level is 10 µg/mL, the median is 14 µg/mL, 75th percentile, and the highest value is 18 µg/mL.

The small table provides a good opportunity for scientists to understand the basic features of the pharmaceutical characteristic.

Descriptive statistics may also be calculated separately. For instance, the average is calculated based on the formula:

```python
df["Drug_Concentration"].mean()
```

The median is calculated using:

```python
df["Drug_Concentration"].median()
```

The minimum and maximum values are obtained using:

```python
df["Drug_Concentration"].min()
df["Drug_Concentration"].max()
```

Variance and standard deviation can be calculated using:

```python
df["Drug_Concentration"].var()
df["Drug_Concentration"].std()
```

By default, Pandas computes sample variance and sample standard deviation with a degree of freedom of 1. It is possible to use ddof=0 for population variance and population standard deviation:

```python
df["Drug_Concentration"].var(ddof=0)
df["Drug_Concentration"].std(ddof=0)
```

The minimum value can be subtracted from the largest value to determine the range:

```python
df["Drug_Concentration"].max() - df["Drug_Concentration"].min()
```

For the concentration values 10, 12, 14, 16, and 18 µg/mL, the range is 8 µg/mL.

Summarizing Categorical Variables

Categorical variables abound in pharmaceutical and clinical data sets. These include dosage form, treatment group, gender, disease class, adverse drug event, and formulation type.

Value Counts of Categories may be derived using:

```python
df["Dosage_Form"].value_counts()
```

The number of observations that fall into each category is provided by this function. For example, we could use it to determine how many patients are receiving injections, tablets, or capsules.

To get the categories' percentage distribution, we might use:

```python
df["Dosage_Form"].value_counts(normalize=True) * 100
```

These values are indicative of the percentage of observations that fall under each category. They are helpful in describing participant demographics, treatment arms, drug formulations, side effects, and diseases among others in pharmaceutic articles.

C. Handling Missing Values

It is common for there to be missing values in data from the pharmaceutical industry, clinical studies, laboratories, and production processes. Missing values occur due to reasons such as lack of compliance from patients on their follow-up visits, no lab tests done, samples being lost, instrument malfunction, or recording mistakes.

Identification of missing values can be done through:

```python
df.isnull()
```

This results in a boolean value which determines whether there are missing values in individual cells. The following function can be used to get the number of missing values in each column:

```python
df.isnull().sum()
```

This enables us to determine which variables require additional attention by providing us with the amount of missing data in each column.

Pandas also has the dropna() method, which drops observations with missing values:

```python
df.dropna()
```

Researchers should not always exclude all the observations with missing values. Excluding observations without knowing the cause of their disappearance may lower the number of samples

and bring bias to the analysis. The researcher must first identify the trend and potential reason behind the missing values.

It may be appropriate to replace missing values by applying proper imputation technique. The selection of the imputation technique depends on various factors like research methodology, amount of missing data, type of variable, and missingness assumptions, etc. Thus, the chosen imputation technique should be scientifically valid and clearly explained in the research report.

D. Grouped Descriptive Summaries

Drug research often involves descriptive statistics for various groups. Researchers may analyze the results for a treatment group and control group, various drug preparations, various batches of tablets, various age groups, or various locations where the clinical trial was conducted.

The panda's library offers the group by () method to compute statistics on separate groups. For instance:

```
Python
df.groupby("Treatment_Group")["Drug_Concentration"].mean()
```

The above command computes the mean value of drugs per treatment group. Rather than giving just one mean, the command gives a mean value for each of the Treatment_Group categories.

A more extensive summary can be gained from:

```
Python
df.groupby("Treatment_Group")["Drug_Concentration"].describe()
```

This produces descriptive statistics such as:

- Number of observations
- Mean
- Standard deviation
- Minimum
- 25th percentile
- Median
- 75th percentile
- Maximum

When comparing measurements of central tendency and dispersion between groups receiving various treatments, the resulting summary is helpful.

For example, a difference in the average concentration of medication between groups is indicated when one treatment group has a higher mean concentration than another. However, the existence of a descriptive difference does not necessarily indicate causality; more statistical analysis may be required to determine the significance and scientific significance of the difference.

Tablet batches, drug formulations, age ranges, patient groups by disease, and clinical trial locations can all benefit from group descriptive analysis. The amount of observations inside each group should be considered before interpreting group results since very tiny groups produce unstable estimates and should be handled carefully.

Finally, it is useful to say that NumPy and Pandas supplement each other in pharmaceutical studies. NumPy proves to be handy for numerical calculations with arrays, while Pandas offers more complete approach to structured data from pharmaceutical area. In the process of research workflow, one can use Pandas for preparation of the dataset, while NumPy functions could be used for numerical calculations.

1.4.2. Interpretation of Descriptive Statistical Results

The interpretation of descriptive statistics means explaining what those statistics say about the given dataset. Statistical results cannot be simply presented as a string of numbers. The researcher needs to describe the central tendency, variability, distribution pattern, and unusual values if any in context to the research problem. It is not enough to say that the mean concentration of the drug is 25 µg/mL without describing the distribution pattern of data, variability, and if the obtained mean represents the majority of observations.

The number of observations, mean or median value, standard deviation or interquartile range, minimum and maximum values, and overall distribution pattern are typically included in the interpretation of descriptive statistics. The distribution pattern and type of data determine which statistics should be utilised. Mean and standard deviation are used for symmetrical numeric data, while median and interquartile range are used for skewed distribution or presence of outliers.

A. Interpretation of the Mean

The term "mean" describes the observations' average value. If the average drug concentration is 14 µg/mL, it indicates that all samples have an average concentration of 14 µg/mL. If there are no significant extreme observations and the data are normally distributed, the mean is a useful indicator of central tendency.

For example, the mean can be used to determine the average blood pressure, average medication concentration, average tablet weight, and average dissolving time in a pharmaceutical trial. The

average tablet weight in the production batch, for example, would be 500 mg if the mean tablet weight was 500 mg. This does not imply that every tablet has a weight of 500 mg, though.

Some pills have weights that are either above or below the average. Because extreme values can affect the mean, it is important to interpret the mean value cautiously if the data are skewed.

B. Interpretation of the Median

The middle value of the organised data set is known as the median. When a data collection contains extreme values or is not normally distributed, this measure can be quite helpful. Half of the subjects had reaction times shorter than six days, while the other half had reaction times longer than six days, if the median for reaction times to treatment is six days.

When biological measures are positively skewed, the median is frequently more descriptive of the sample than the mean. For example, the median is likely to be more informative if most subjects have drug concentrations between 10 and 20 µg/mL, but one subject has a concentration of 100 µg/mL. The data is favourably skewed if the mean is much higher than the median. The data is negatively skewed if the mean is lower than the median.

C. Interpretation of Standard Deviation

The degree to which the observations differ from the mean value is shown by the standard deviation. The data is more concentrated when the standard deviation is smaller, and more dispersed when it is bigger.

For example, let's say that the tablets have an average weight of 500 mg and a standard variation of 2 mg. This indicates that the weight of the tablets is reasonably constant. However, if we look at another group of pills with an average weight of 500 mg and a standard deviation of 20 mg, this indicates that there is a lot of variety even though the average is the same.

A high standard deviation in medicine indicates that patients receiving the same treatment would respond differently. Variations in age, body weight, disease severity, heredity, metabolism, compliance, or other factors could be the cause of this. Individual answers are more uniform when the standard deviation is lower.

It is necessary to take into account the standard deviation in relation to the measurement unit. For example, if a variable's mean is 10 and its standard deviation is 5, this is a far larger standard deviation than if the mean is 1,000.

D. Interpretation of Range and Percentiles

One of the most basic ways to quantify dispersion is by range. It describes the variation between the data set's top and lowest values. Range does not provide information on the distribution of observations inside these two bounds, but it does tell us about the overall variation in the data.

The range is computed using:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

For example, suppose the drug concentration values measured in a group of patients range from 10 to 30 μg/mL. The range is:

$$\text{Range} = 30 - 10 = 20 \text{ μg/mL}$$

Therefore, the total spread for the drug concentration is 20 μg/mL. Nevertheless, it cannot give an idea of where most of the values are concentrated – near 10 μg/mL, near 30 μg/mL, or somewhere in the middle of the range. Furthermore, because it relies on just two numbers, it is susceptible to extreme values.

The percentile, which shows how the values are spread within a data collection, is a more accurate metric. A percentile is a value that falls below a given proportion of the values. For instance, the 50th percentile is the median value, whereas the 25th percentile is the number that roughly 25% of the observations fall below.

The interquartile range (IQR) is the separation between the 75th and 25th percentiles:

$$\text{IQR} = Q_3 - Q_1$$

The interquartile range indicates the dispersion of middle 50 percent of the observations. The IQR is especially valuable for analyzing data in pharmaceutical studies and clinical practice due to its immunity to extremely large or small observations when compared to range and standard deviation.

For instance, let the treatment time for the patients be defined in the following manner:

- 25th percentile = 4 days
- 50th percentile (median) = 7 days
- 75th percentile = 10 days

The interquartile range is:

$$\text{IQR} = 10 - 4 = 6 \text{ days}$$

Therefore, the IQR is 6 days, indicating that the middle 50% of observed treatment response times lie within the interval from 4 to 10 days.

E. Interpretation of Variance

Variance is an indicator of how much the data values vary from the mean of the data. While being very helpful in computations, variance uses squared units, which makes its interpretation somewhat complex. For instance, variance of 25 ((\mu g/mL) ^2) is the measure of variation in the drug concentration; however, it does not use the original units of drug concentration.

Variance is a good measure of variability comparison when applied to different groups whose data is measured using the same units. Thus, if the variance of the drug concentration is smaller for one treatment group compared to the other, it shows that the drug concentration is more stable in the first group. However, it is important for the researcher to look at the standard deviation.

F. Interpretation of Distribution Shape

Shape of the distribution is another factor that gives information about the arrangement of the observations. In the case of symmetrical distribution, there would be about equal dispersion on either side of the center. A distribution that is positively skewed would have a long tail to the right, whereas one that is negatively skewed would have a long tail to the left.

When selecting the measures for descriptive statistics, the distribution's form is crucial. Mean and standard deviation would be appropriate for symmetrical distribution, whereas, median and interquartile range would be more accurate for skewed distribution. Categorical variables should be reported in terms of frequency, percentage and mode.

Researchers can find out shape of distribution through histogram, boxplot and coefficients of skewness. The histogram would tell whether the observations are clustered around the center or are distributed in an uneven manner. Box plot will help in finding out the value of the median and quartiles. This method should not be used as the only tool.

G. Interpretation of Outliers

A value that is significantly higher or lower than other observations are called an outlier. An aberrant biological reaction, a measurement error, a data input error, an instrument malfunction, or true individual variance can all result in an outlier.

For instance, when all drug concentration levels are in the range from 10 to 20 µg/mL but one is at 100 µg/mL, the observation needs to be analyzed. It is essential for the researcher to check if the observation is recorded correctly and if the sample was prepared correctly. A true outlier must not be discarded since it affects the mean.

While the median and interquartile range may be less impacted, outliers have a significant impact on the mean, range, variance, and standard deviation. Thus, it is crucial for researchers to mention the outliers and state how they are treated.

H. Example of a Descriptive Statistical Interpretation

Suppose the drug concentrations obtained from a group of patients have the following summary:

- Number of observations: 50
- Mean concentration: 25 µg/mL
- Median concentration: 23 µg/mL
- Standard deviation: 8 µg/mL
- Minimum concentration: 10 µg/mL
- Maximum concentration: 50 µg/mL

From the above results, it can be seen that the average concentration of drugs in the body is 25 µg/mL while the median concentration of drugs is 23 µg/mL. Since some patients have relatively high medication concentrations, the distribution is probably positively skewed since the mean concentration is greater than the median. There is fluctuation in observations around the mean concentration, as indicated by the 8 µg/mL standard deviation. The readings fall between 10 and 50 µg/mL.

The researcher should examine the distribution to find any outliers because the mean is higher than the median and the highest value is noticeably higher than the central values. With the use of the histogram or box plot, it will be easy to see whether the outliers are biologically plausible or due to measurement errors.

REFERENCES

1. Armstrong, D., & Stratton, R. D. (Eds.). (2016). *Oxidative stress and antioxidant protection: The science of free radical biology and disease*. John Wiley & Sons.
2. Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs.
3. Basile, A. O., Yahi, A., & Tatonetti, N. P. (2019). Artificial intelligence for drug toxicity and safety. *Trends in pharmacological sciences*, *40*(9), 624-635.
4. Baxevanis, A. D., Bader, G. D., & Wishart, D. S. (Eds.). (2020). *Bioinformatics*. John Wiley
5. Berry, D. A. (2016). *Statistical methodology in the pharmaceutical sciences*. CRC Press.
6. Berry, D. A., & Stangl, D. (2018). *Bayesian biostatistics*. CRC Press.
7. Bhattamisra, S. K., Banerjee, P., Gupta, P., Mayuren, J., Patra, S., & Candasamy, M. (2023). Artificial intelligence in pharmaceutical and healthcare research.
8. Binder, H., & Blettner, M. (2015). Big data in medical science—a biostatistical view: Part 21 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *112*(9), 137.
9. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... & Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, *37*(8), 852-857.
10. Chen, D. G., Jin, Z., Li, G., Li, Y., Liu, A., & Zhao, Y. (Eds.). (2017). *New advances in statistics and data science*. Cham: Springer International Publishing.
11. Chow, S. C., & Liu, J. P. (2018). *Statistical design and analysis in pharmaceutical science: validation, process controls, and stability*. CRC Press.
12. Dmitrienko, A. (2017). *Analysis of clinical trials using SAS: a practical guide*. SAS institute.
13. Drăghici, S. (2016). *Statistics and data analysis for microarrays using R and bioconductor*. CRC
14. Elseviers, M., Wettermark, B., Mueller, T., Benko, R., Bennie, M., Gvozdanovic, K., ... & Vogler, S. (Eds.). (2024). *Drug utilization research: methods and applications*. John Wiley & Sons.
15. Foucquier, J., & Guedj, M. (2015). Analysis of drug combinations: current methodological landscape. *Pharmacology research & perspectives*, *3*(3), e00149.
16. Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., & Friede, T. (2023). Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, *32*(2), 425-440.

17. Jarantow, S. W., Pisors, E. D., & Chiu, M. L. (2023). Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. *Current Protocols*, *3*(6), e801.

18. Jeyaraman, N., Jeyaraman, M., Yadav, S., Ramasubramanian, S., & Balaji, S. (2024). Revolutionizing healthcare: the emerging role of quantum computing in enhancing medical technology and treatment. *Cureus*, *16*(8).

19. Kandhare, P., Kurlekar, M., Deshpande, T., & Pawar, A. (2025). A review on revolutionizing healthcare technologies with AI and ML applications in pharmaceutical sciences. *Drugs and Drug Candidates*, *4*(1), 9.

20. Kolluri, S., Lin, J., Liu, R., Zhang, Y., & Zhang, W. (2022). Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: A Review: Machine Learning and Artificial Intelligence in Pharmaceutical R&D. *The AAPS journal*, *24*(1), 19.

21. Li, S. (Ed.). (2020). *Computational methods and data analysis for metabolomics*. Totowa, NJ, USA:: Humana Press.

22. Lo, A. W., Siah, K. W., & Wong, C. H. (2019). Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, *1*(1).

23. Mahato, D. T. K. (2025). Biostatistics And Research Methodology: Pillars of Evidence-Based Pharmaceutical and Health Sciences. *International Journal Pharmaceutical Medicinal Research*.

24. Maitournam, A. (2025). Statistics, Targeted Clinical Trials, Big data, Stratified and Personalized Medicines. In *International Encyclopedia of Statistical Science* (pp. 2647-2654). Berlin, Heidelberg: Springer Berlin Heidelberg.

25. Michel, M. C., Murphy, T. J., & Motulsky, H. J. (2020). New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. *Molecular Pharmacology*, *97*(1), 49-60.

26. Patterson, S. D., & Jones, B. (2017). *Bioequivalence and statistics in clinical pharmacology*.

27. Peace, K. E., Chen, D. G., & Menon, S. (Eds.). (2018). *Biopharmaceutical Applied Statistics Symposium*. Springer.

28. Peace, K. E., Chen, D. G., & Menon, S. (Eds.). (2018). *Biopharmaceutical Applied Statistics Symposium: Volume 2 Biostatistical Analysis of Clinical Trials*. Springer.

29. Roschangar, F., Li, J., Zhou, Y., Aelterman, W., Borovika, A., Colberg, J., ... & Sheldon, R. A. (2022). Improved iGAL 2.0 metric empowers pharmaceutical scientists to make meaningful contributions to United Nations Sustainable Development Goal 12. *ACS Sustainable Chemistry & Engineering*, *10*(16), 5148-5162.

30. Rossi, R. J. (2022). *Applied biostatistics for the health sciences*. John Wiley & Sons.

31. Sakarna, Y., Kalakota, S., Gottipati, A., Kaviandost, P., Chileveru, K., Bhola, R., & Singh, S. (2025). Applications of Biostatistics in Healthcare. *Public Health, Epidemiology, and Beyond. Saudi J Med*, *10*(5), 270-276.

32. Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of librarianship and information science*, *51*(3), 771-780.

33. Shaffer, J. G., Mather, F. J., Wele, M., Li, J., Tangara, C. O., Kassogue, Y., ... & Doumbia, S. O. (2019). Expanding research capacity in sub-Saharan Africa through informatics, bioinformatics, and data science training programs in Mali. *Frontiers in Genetics*, *10*, 331.

34. Shortreed, S. M., Cook, A. J., Coley, R. Y., Bobb, J. F., & Nelson, J. C. (2019). Challenges and opportunities for using big health care data to advance medical science and public health. *American journal of epidemiology*, *188*(5), 851-861.

35. Steinwandter, V., Borchert, D., & Herwig, C. (2019). Data science tools and applications on the way to Pharma 4.0. *Drug discovery today*, *24*(9), 1795-1805.

36. Su, M., Pan, T., Chen, Q. Z., Zhou, W. W., Gong, Y., Xu, G., ... & Li, Y. S. (2022). Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. *Military Medical Research*, *9*(1), 68.

37. Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018). Innovation at the intersection of clinical trials and real-world data science to advance patient care. *Clinical and translational science*, *11*(5), 450-460.

38. Terranova, N., & Venkatakrishnan, K. (2024). Machine learning in modeling disease trajectory and treatment outcomes: An emerging enabler for model-informed precision medicine. *Clinical Pharmacology & Therapeutics*, *115*(4), 720-726.

39. Yan, F., Robert, M., & Li, Y. (2017). Statistical methods and common problems in medical or biomedical science research. *International journal of physiology, pathophysiology and pharmacology*, *9*(5), 157.

40. Zhang, L., Kuhn, M., Peers, I., & Altan, S. (Eds.). (2016). *Nonclinical statistics for pharmaceutical and biotechnology industries*. Switzerland: Springer International Publishing.

Unit 2...

PROBABILITY AND STATISTICAL DISTRIBUTIONS IN HEALTHCARE

The fundamental ideas of probability and statistical distribution in the context of healthcare are covered in this subject. The module uses conditional probability, Bayes' theorem, and probability rules to illustrate the idea of uncertainty, its measurement, and its interpretation. The section also explains probability models and random variables, including their discrete and continuous varieties. The applications of these concepts in the context of healthcare have been covered which include application in the design of clinical trial studies, diagnostic tests, pharmacokinetics, adverse effects of drugs, survival analysis, hospitalization, and public health studies. Important statistical distributions like normal, binomial, and Poisson have been introduced along with their properties and healthcare applications. This module will also provide an insight into the use of the programming language Python to calculate and interpret probability distributions using graphs and statistical functions.

2.1. FUNDAMENTALS OF PROBABILITY

The branch of mathematics that examines uncertainty and the likelihood of events occurring is called probability. The field of pharmaceutical and biological science has numerous instances where uncertainty prevails, like how a patient responds to a medicine, how a drug will react adversely, how a laboratory test detects disease, and how a batch of manufactured products is qualified. Probability gives a mathematical approach in the measurement and description of uncertain events. Probability makes the researcher able to make well-informed decisions, clinical trial designs, risk assessments, test interpretation, and understanding biological observation variability.

Probability can be represented numerically as values ranging from 0 to 1. An event has a probability of 0 if it will never occur and a probability of 1 if it is certain to occur. Probability can also be presented in percentage form. If a probability of 0.25 occurs, this means 25%. Probability does not mean that an event will definitely happen; it just measures its chances of happening.

2.1.1. Basic Probability Concepts and Laws

A technique for calculating the likelihood that an event will occur is called probability. It has been utilized in clinical and pharmaceutical sciences to quantify uncertainty as well as decision making from experimental or observed data. The probability can be estimated for response to a treatment,

adverse drug reactions, success of quality control testing, presence of diseases, and results of clinical or laboratory experiments. The key concepts in probability include experiment, outcome, sample space, event, favourable outcome, and probability.

An experiment is an action or procedure that produces an observable outcome. In the probability concept, experiment entails uncertainty of the outcome before performing the experiment. In pharmaceutical research, the examples of random experiment can include selection of patients for clinical studies, conducting tablet dissolution tests, and checking the presence of adverse drug reaction.

An outcome is one of the possible results from a random experiment. For instance, in the case where a tablet is subjected to a dissolution test, the outcome is either pass or fail depending on the specified specification. When a patient is administered with medication, some of the possible outcomes could include improvement, no improvement, or having an adverse effect.

The collection of all potential results of a random experiment is called sample space, or S. A tablet quality control test with two possible outcomes pass or fail is an example of a sample space:

$$S = \{\text{Pass}, \text{Fail}\}$$

A particular result or group of results that are included in the sample space is referred to as an event. Typically, capital letters like A, B, or C are used to indicate an occurrence. Let's take an example where A is an event in a clinical investigation, such as when patients respond favourably to a treatment.

P(A) is the probability of an event A. The probability could be expressed as follows if every possible outcome has an equal chance of occurring:

$$P(A) = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}}$$

For instance, if quality control is conducted for 100 tablets and only 95 of the tablets meet the required quality standards, then the probability of choosing one passing tablet randomly would be:

$$P(\text{Pass}) = \frac{95}{100} = 0.95$$

Therefore, the probability of picking up a good tablet is 0.95, which is 95%.

The probability's value is always between 0 and 1. If the event cannot occur, the chance is zero; if it will undoubtedly occur, the likelihood is one. By multiplying the chance by 100, it can also be expressed as a percentage.

A. Important Basic Probability Concepts

A simple event comprises of either one outcome or one possible result. In this case, when conducting a tablet quality check with two possible outcomes being pass or fail, then selecting a tablet which will pass in this check will be referred to as a simple event.

A compound event comprises of more than two outcomes or two or more events. In a pharmaceutical quality check, an event whereby a tablet is selected based on whether it passes a weight test or a dissolution test is considered a compound event due to the involvement of more than one condition.

An occurrence is assured if it takes place. Taking into account the entire sample space, the likelihood of a specific event is:

$$P(A) = 1$$

In this case, if all the tablets in the specified data set are either pass or fail, the certainty of a randomly picked tablet being either a pass or a fail is one hundred percent.

An impossible event refers to an event that cannot happen in the circumstances of the experiment. The probability of an impossible event is:

$$P(A) = 0$$

For instance, in a case where testing of tablets is done using a classification system which can produce either pass or fail, picking a tablet that falls under neither of these two classifications would amount to choosing something that is impossible in that classification system.

A complement of an event Ais denoted as A^cor A^. It includes all the possible outcomes where Ahas not happened. For instance, in the case where Adenotes the event of a positive response from a patient on the medication, then A^cmeans non-positive response.

The probability of the complement is:

$$P(A^c) = 1 - P(A)$$

Suppose the probability of a patient responding to treatment is 0.80. The probability that the patient does not respond is:

$$P(A^c) = 1 - 0.80 = 0.20$$

Therefore, the probability of non-response is 0.20, or 20%.

B. Fundamental Laws of Probability

There are many mathematical laws related to probability. It is on the basis of these mathematical laws that probability can be calculated for different events.

1. Law of non-negativity

Any event's probability cannot be negative. Consequently:

$$P(A) \geq 0$$

This implies that all probabilities have to be either zero or higher. The probability cannot take the form of a value such as -0.20.

Practical pharmacological studies may use probabilities ranging from 0 to 1. Thus, for instance, when there is 0.25 probability of something happening, this implies a 25 percent possibility, but 0 probability implies no chance at all.

2. Law of Total Probability

The sample space as a whole has a probability of one:

$$P(S) = 1$$

Sample space comprises all the potential results of the trial. In that regard, one of the potential results must happen whenever the trial is conducted.

For instance, if the potential results of a tablet quality test are just Pass or Fail, then:

$$P(\text{Pass}) + P(\text{Fail}) = 1$$

If the probability of passing is 0.95, then:

$$P(\text{Fail}) = 1 - 0.95 = 0.05$$

Thus, the two possible outcomes together account for the entire probability.

3. Addition Law of Probability

The chance that at least one of two occurrences will occur can be found using the addition law. For events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Here:

- $A \cup B$means event Aor event B, or both.
- $A \cap B$means both $A$and Boccur.

The intersection $P(A \cap B)$is subtracted because it would otherwise be counted twice when $P(A)$and $P(B)$are added.

For example, suppose that in a clinical study:

$$P(\text{Headache}) = 0.30$$

$$P(\text{Nausea}) = 0.20$$

and:

$$P(\text{Headache and Nausea}) = 0.05$$

Then:

$$P(\text{Headache or Nausea}) = 0.30 + 0.20 - 0.05$$

$$P(\text{Headache or Nausea}) = 0.45$$

Thus, the probability that a patient suffers from headache or nausea, or both, is 0.45 or 45%.

Two events are said to be mutually exclusive when they are incompatible. When events are mutually exclusive:

$$P(A \cap B) = 0$$

Therefore, the addition law simplifies to:

$$P(A \cup B) = P(A) + P(B)$$

For instance, in one category of quality testing, the same tablet cannot be classified as having passed and failed the test at once. This implies that the two events cannot occur together.

4. Multiplication Law of Probability

To determine the likelihood that two events will occur simultaneously, apply the multiplication law. For two occasions A and B:

$$P(A \cap B) = P(A) \times P(B \mid A)$$

In this case, P(B|A) is the conditional probability that event B will occur given that event Ahad has already occurred.

Let's say that the first tablet chosen passes the test (A) and the second tablet passes the test (Bis). If the second option is unrelated to the first, then:

$$P(A \cap B) = P(A) \times P(B)$$

Suppose:

$$P(A) = 0.95$$

and:

$$P(B) = 0.95$$

Then:

$$P(A \cap B) = 0.95 \times 0.95$$

$$P(A \cap B) = 0.9025$$

As a result, the probability of both being good is 0.9025, which equals 90.25%.

It is especially useful for researchers to use this rule when determining the joint probability of two events. Independence should be proved in order to use the simple equation P(A)P(B).

5. Complementary Law

According to the complementary law, the likelihood of something not occurring is equal to one minus the likelihood of it occurring:

$$P(A^c) = 1 - P(A)$$

This rule applies when determining the chance that an event won't occur is simpler.

For example, let the likelihood that a patient will experience a negative side effect after taking a certain medication be:

$$P(A) = 0.10$$

The probability that the patient does not experience the adverse drug reaction is:

$$P(A^c) = 1 - 0.10 = 0.90$$

The likelihood that there won't be a negative medication reaction is therefore 0.90, or 90%.

C. Independent and Dependent Events

When two events take place, they are considered independent if the likelihood of the second event is unaffected by the occurrence of the first. In other words, information about one occurrence shouldn't have an impact on the likelihood of another event happening.

When two separate quality control tests are conducted, for example, and the outcome of one test has no bearing on the outcome of the other, the events can be regarded as independent in probability computation. For autonomous events:

$$P(B \mid A) = P(B)$$

Consequently:

$$P(A \cap B) = P(A) \times P(B)$$

If the occurrence of one event affects the likelihood of the other, then the two events are said to be dependant. One example would be selecting two tablets from a limited supply without replacing them. There will be fewer and different tablets remaining in the batch once the first tablet is selected. Therefore, the outcome of the first tablet may have an impact on the likelihood that the second one selected would be good.

The following formula can be used to determine the probabilities of dependent events:

$$P(A \cap B) = P(A) \times P(B \mid A)$$

It is important to learn how to differentiate between independent and dependent events because failure to distinguish between them might result in erroneous probability computations.

D. Mutually Exclusive Events

Events that cannot occur at the same time are referred to as mutually exclusive. One event's occurrence precludes the occurrence of another.

For events that are mutually exclusive:

$$P(A \cap B) = 0$$

Therefore, the addition law becomes:

$$P(A \cup B) = P(A) + P(B)$$

For instance, suppose a certain patient is classified on the basis of being either responsive or non-responsive to treatment by using some predefined criteria. In this case, both these categories will never occur at the same time. In a similar manner, a tablet cannot be found both passing and failing the same quality test at the same time through the same criterion.

It should be mentioned that independent occurrences are not the same as mutually exclusive events. The latter relates to the fact that the occurrence of one event does not affect the likelihood of another, whereas the former refers to events that do not occur simultaneously. This is because two occurrences that are mutually exclusive do not happen simultaneously.

Therefore, it can be said that knowledge about sample space, events, complementary, addition law and multiplication law, independence, dependence and mutual exclusivity constitute the foundation of probability theory in the field of pharmaceutical and clinical research.

2.1.2. Conditional Probability and Bayes' Theorem

The probability of one event occurring given that another event has already occurred is known as the conditional probability. When additional information becomes available and changes the likelihood of an event, the conditional probability can be used. Conditional probability is represented by the notation P(A|B).

The conditional probability formula is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

where:

- $P(A \mid B)$= probability of event Agiven that event Bhas occurred
- $P(A \cap B)$= probability that both $A$and Boccur
- $P(B)$= probability of event B

The formula is applicable when:

$$P(B) > 0$$

As another example, if a medical study consists of 100 patients and out of them 40 are administered Drug A and 30 patients improved after being administered Drug A, then the probability of improving patients when administered Drug A can be calculated by using:

$$P(\text{Improvement} \mid \text{Drug A}) = \frac{30}{40}$$

$$P(\text{Improvement} \mid \text{Drug A}) = 0.75$$

So, the conditional probability of recovery for patients taking Drug A will be 0.75 or 75%.

It is different from the probability of recovery for the whole group of patients in the study since the conditional probability takes into account some extra information, which in our example is the drug taken by the patient, so the population gets smaller.

A. Conditional Probability in Pharmaceutical Research

Conditional probability has a wide variety of applications in pharmaceuticals, clinical studies, diagnostics, and pharmacokinetics. Conditional probability allows pharmaceutical researchers to assess the event in view of some condition or property of the subjects under study.

So, for instance, the probability of occurrence of an adverse reaction to the drug due to the fact that a certain dose of the drug was prescribed to the patient can be estimated. In addition, the probability of response to the therapy due to belonging to a certain age group, certain severity of the disease, a certain genetic feature or prior treatment can be assessed.

The relevance of conditional probability arises in the study of the results of diagnostic tests. For example, it is possible to estimate the likelihood that the disease will be present if the diagnostic test result was positive.

Conditional probability is important also in pharmacokinetic studies. So, for instance, the probability of obtaining a high concentration of the drug can be estimated given that the patient has poor kidney function. Poor kidney function will influence the rate of drug elimination from the body; thus, knowledge about the state of the kidneys can influence the probability of obtaining the specific concentrations of the drug.

Thus, conditional probability allows incorporating extra information when estimating probability of an event.

B. Multiplication Rule Using Conditional Probability

Conditional probability provides the basis for the multiplication rule. For two events $A$and B:

$$P(A \cap B) = P(A) \times P(B \mid A)$$

The same relationship can also be written as:

$$P(A \cap B) = P(B) \times P(A \mid B)$$

These formulas determine the probability of occurrence for both events.

Consider the following example. The probability of randomly selecting a patient who receives Drug A is:

$$P(\text{Drug A}) = 0.40$$

Suppose the probability that a patient improves given that the patient received Drug A is:

$$P(\text{Improvement} \mid \text{Drug A}) = 0.75$$

A randomly chosen patient's chances of receiving Drug A and getting better are:

$$P(\text{Drug A} \cap \text{Improvement}) = 0.40 \times 0.75$$

$$P(\text{Drug A} \cap \text{Improvement}) = 0.30$$

Thus, the probability is 0.30 or 30%.

Here, we have seen how the joint probability of two events may be found using the conditional probability and the probability of one event.

C. Bayes' Theorem

A mathematical method for determining the likelihood of an event depending on further evidence is called the Bayes theorem. When attempting to determine the likelihood of a sickness or the reason behind an observable event, this theorem works best.

The Bayes theorem formula is:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

where:

- $P(A \mid B)$= posterior probability of Agiven B
- $P(B \mid A)$= probability of observing $B$when $A$is true
- $P(A)$= prior probability of A
- $P(B)$= overall probability of observing B

The key parts of Bayes' theorem are explained below:

- **Prior probability:** The likelihood of the occurrence without taking into account any new information.

- **Likelihood:** The likelihood of seeing the evidence in the event that the condition or occurrence materialises.
- **Evidence:** The likelihood of seeing the evidence whether or not the condition materialised.
- **Posterior probability:** The likelihood of an event after considering the evidence.

Bayes' theorem is important because it proves that the interpretation of new evidence is not based on the evidence alone but also on the prior probability of the condition.

D. Example of Bayes' Theorem in Diagnostic Testing

Assume that there is a test with a 90% sensitivity for a particular ailment that affects 1% of the population. This suggests that the test findings will identify 90% of those who have the illness as having it. Five percent of test results are false positives.

Let:

- D= patient has the disease
- D^c= patient does not have the disease
- $+$= test result is positive

The known probabilities are:

$$P(D) = 0.01$$
$$P(D^c) = 0.99$$
$$P(+|\ D) = 0.90$$
$$P(+|\ D^c) = 0.05$$

First, the law of total probability is used to determine the overall likelihood of receiving a positive test result:

$$P(+) = P(+|\ D)P(D) + P(+|\ D^c)P(D^c)$$

Substituting the values:

$$P(+) = (0.90 \times 0.01) + (0.05 \times 0.99)$$
$$P(+) = 0.009 + 0.0495$$
$$P(+) = 0.0585$$

Given a positive test result, Bayes' theorem can now be used to calculate the likelihood that an individual truly has the illness:

$$P(D\ |\ +) = \frac{P(+|\ D)P(D)}{P(+)}$$

Therefore:

$$P(D \mid +) = \frac{0.90 \times 0.01}{0.0585}$$

$$P(D \mid +) \approx 0.1538$$

Therefore, given a positive test result, the likelihood that the person has the illness is around 15.38%.

The case above is illustrative of a key concept that is critical to understanding the process of diagnosis through tests. Positive test results do not always indicate with absolute certainty that the individual has the disease. Even though the test has reasonable sensitivity, the disease is rare in the population, and there are false positives. Consequently, in the event of a positive test, the probability of the disease is influenced by the prior probability of the disease.

E. Bayes' Theorem in Pharmaceutical Research

The various areas where Bayes' theorem can be applied in pharmaceuticals and medicine include interpreting diagnostic results, estimating the likelihood of treatment response, evaluating adverse reactions to medications, and changing conclusions based on new evidence.

Bayesian methods could help in determining if an adverse event that occurred has anything to do with the medicine in question. The previous research evidence can form the prior probability of the relationship, while the collected data about the adverse event forms the new evidence. This posterior probability can then serve as the basis for further investigation.

Bayesian method can be applied in clinical trials through combining the previous information with the new evidence gathered from the clinical study. For instance, the previous study can suggest that there is a moderate probability that the medicine works effectively. The clinical trial would update this probability of success of treatment.

Bayes' methods can also find application in the area of personalized medicine. The probability that a person responds positively to a certain treatment can be updated with the help of evidence related to his or her genetic predisposition, age, the stage of disease, past responses to treatment, and laboratory results.

Lastly, because it distinguishes between the likelihood of having a certain test result if a particular disease exists and the likelihood of having that disease if the specific test result is known, Bayes' theorem is highly useful in the field of diagnostic medicine.

F. Difference Between Conditional Probability and Bayes' Theorem

Bayes' theorem and conditional probability are similar concepts, although they serve slightly different functions.

Conditional probability determines the likelihood of an event given that another event has already occurred. As an example:

$$P(\text{Positive Test} \mid \text{Disease})$$

This represents the likelihood of receiving a positive test result from the population with the illness.

The Bayesian theorem helps to invert the direction of the reasoning as follows:

$$P(\text{Disease} \mid \text{Positive Test})$$

This is the probability that the patient has the illness even after learning that their test results were positive.

They are usually not the same. For instance, a test may have:

$$P(\text{Positive Test} \mid \text{Disease}) = 0.90$$

while:

$$P(\text{Disease} \mid \text{Positive Test}) = 0.1538$$

as demonstrated in the diagnostic example.

As a result, it's critical to distinguish between "probability of a positive outcome given the disease" and "probability of the disease given a positive outcome." Data from clinical research, pharmacological trials, and diagnostics may be misinterpreted if these two probabilities are mixed up.

The Bayes theorem is used to update such probabilities in line with new knowledge, while the theory of conditional probability is founded on the idea of reasoning about specific events given certain conditions.

2.2. RANDOM VARIABLES AND PROBABILITY MODELS

Random variables and probability models, which provide mathematical representations of unpredictable observations, are essential ideas in statistics. Biological variability, measurement errors, patient-related factors, reaction to treatment, and environment all influence the majority of observations in pharmaceutical and health care research. A random variable is a random phenomenon given a numerical number. Some examples of random variables include the number of patients responding to a treatment, a patient's blood glucose level, the amount of time it takes for the drug to dissolve, and the number of side effects encountered in a clinical study. A probability model, on the other hand, specifies the range of possible values for a random variable and gives a probability for each such value. These models are used by researchers to estimate uncertainties, risk assessment, prediction of outcomes, comparisons of treatments and informed decision making.

Probability model involves three key parts. They include the random variable, potential values that the variable can assume, and probability assigned to each value. A probability model's total

probabilities should always equal 1. Probabilities are given to particular values when dealing with a discrete random variable. For a continuous random variable, probabilities are assigned to intervals of values. There exist various ways in which probability models can be obtained. They include theory, literature, observations, and estimations. Probability models in healthcare research are used for a variety of purposes like description of patient population, evaluation of diagnostic tests, investigation of disease prevalence, assessing effectiveness of treatment and estimating probabilities of clinical outcomes.

2.2.1. Discrete and Continuous Random Variables

A random variable is a number whose value is dependent on how a random experiment turns out. It indicates that the value is unclear before the experiment or observation, not that the variable is always uncertain in all situations. Discrete random variables and continuous random variables are the two primary categories of random variables. This classification is based on the kinds of values that the variable can take on and the way that these values are given probabilities.

1. Discrete Random Variables

A random variable that only takes on a countable number of distinct values is called a discrete random variable. The values, which are typically obtained by counting the number of events, people, things, or occurrences, are primarily integers. The number of patients who recover from treatment, the number of tablets that don't pass quality control tests, the number of drug-related side effects, and the quantity of microorganisms in a culture sample are all examples of discrete random variables in pharmacological research.

An uppercase letter X is used to indicate a discrete random variable, whereas a lowercase letter x is used to express the specific value that the random variable assumes. The likelihood that a specific value will be assigned to the discrete random variable X is:

$$P(X = x)$$

A probability mass function (PMF) can be used to define the probability distribution for a discrete random variable. PMF assigns probabilities to every possible value for a random variable. Two essential requirements must be met by the probability:

$$P(X = x) \geq 0$$

and

$$\sum P\,(X = x) = 1$$

It implies, first of all, that probabilities are never negative. Second, the sum of all conceivable outcomes' probability should equal one.

Assume, for example, that three individuals were selected for a clinical trial and that the number of patients who experienced a negative reaction to a particular medication was noted. The number of patients who experienced the reaction is denoted by X. The range of values for Xare:

$$X = 0,1,2,3$$

In this case, X=0 means that none of the three patients have had the reaction, X=1 means that one of the three patients have had the reaction, X=2 means that two of the three patients have had the reaction, and X=3 means that all three patients have had the reaction.

Binomial Distribution

The binomial probability distribution is the most important kind of discrete probability distribution. It is applicable when research consists of a set number of independent trials with just two possible outcomes and a consistent probability of success for each trial.

The two events could be therapeutic success or failure, cure or no cure, death or survival, or the occurrence or non-occurrence of the event under examination in pharmaceutical and medical research.

Let X be the random variable that indicates how many tries were successful. The probability distribution of binomials is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where:

- n= total number of trials
- x= number of successes
- p= probability of success in each trial
- $1 - p$= probability of failure in each trial
- $\binom{n}{x}$= number of possible ways of obtaining $x$successes among $n$trials

Let's consider an example where a treatment has an efficacy probability of 0.80. Let's apply this treatment to five separate people so that each of them has the same effectiveness probability. A binomial random variable is the number of patients responding to the treatment. The binomial distribution is used to determine the likelihood that three, four, or five patients would respond.

Poisson Distribution

The Poisson distribution is another discrete probability model that is worth discussing. This model is widely used to characterise the quantity of events with constant average rates of occurrence that occur across a certain time period or geographical area.

The Poisson distribution can be used in pharmaceutical and medical research to estimate phenomena like the number of hospital emergency admissions on any given day, the number of adverse drug reactions reported within a specific time frame, the number of microbe colonies discovered in laboratory samples, and so on.

The Poisson distribution's probability mass function is:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where:

- X= number of events
- x= particular number of events
- λ= average number of events in the specified interval
- e= mathematical constant approximately equal to 2.71828
- $x!$= factorial of x

For example, if a hospital has five admissions on average each day, the Poisson distribution can be used to calculate the likelihood of having a particular number of admissions in a given day, provided that the underlying assumptions of the distribution are true.

Thus, discrete random variables find great application whenever pharmaceutical information has been gathered by means of counting. The main benefit of discrete random variables is the fact that one may assign probabilities to specific possible values. Discrete random variables cannot be applied to measure body temperature, blood pressure, body weight, or plasma drug concentration since these are measurement variables that assume continuous values.

2. Continuous Random Variables

A random variable that can take on any value over an interval is referred to as a continuous random variable. A continuous random variable does not need to be a whole integer or a group of numbers, in contrast to a discrete random variable.

Some examples of a continuous random variable in pharmaceutical and health care research are:

- Body temperature
- Body weight
- Height
- Blood pressure
- Blood glucose concentration
- Plasma drug concentration

- Dissolution time
- Duration of hospital stay

Consider a scenario in which a patient's body temperature could be 37°C, 37.1°C, 37.15°C, 37.156°C, or any value in between. In this situation, an infinite number can exist between any two numbers. Continuous variables are therefore perceived as having been collected over an interval rather than as discrete values.

The probability density function is used to illustrate the distribution probability of continuous random variables. The picture below displays the likelihood that the variable will take on values between any two numbers, a and b:

$$P(a \leq X \leq b) = \int_a^b f\,(x)\,dx$$

where the probability density function is denoted by f(x).

The likelihood of getting a precise point is zero, which is a highly important characteristic of the continuous random variable:

$$P(X = x) = 0$$

This does not, however, mean that the metric is not observable or quantifiable in and of itself. In contrast, probability is assigned to an interval rather than a single point in the case of a continuous distribution.

For example, researchers calculate the likelihood that a patient's systolic blood pressure will be between 120 and 130 mmHg rather than precisely 125.000 mmHg:

$$P(120 \leq X \leq 130)$$

The probability density function's whole area must be equal to 1:

$$\int_{-\infty}^{\infty} f\,(x)\,dx = 1$$

This means that there is a 100% chance that the variable will take any value within its whole range.

Normal Distribution

One of the most important continuous probability distributions is the normal distribution. It can be characterised by two parameters, namely the mean (μ) and standard deviation (σ), and is characterised by a bell-shaped symmetric distribution.

The following is the normal distribution formula:

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

where:

- x= observed value
- μ= population mean
- σ= population standard deviation
- π= mathematical constant
- e= mathematical constant

The distribution's center is indicated by the mean, while its dispersion is indicated by the standard deviation. The distribution is narrower when the standard deviation is lower. On the other hand, a higher standard deviation indicates greater dispersion.

Many biological and pharmacological measurements may be regularly distributed if all the requirements are satisfied. These metrics could include blood pressure, body height, weight, etc. However, rather than assuming that all biological measurements are normally distributed, it is preferable to test for the distribution of the actual data.

Since most statistical methods rely on normality or use the normal approximation, the normal distribution is crucial to statistics.

Uniform Distribution

The uniform distribution, a continuous probability distribution, is another type of probability model.

Every value within an interval has the same probability in a continuous uniform distribution. For instance, when a measurement is supposed to have a uniform distribution between two defined points, the values in the interval are equally dense. The uniform distribution applies in cases where there is nothing in particular about some values being more common than others within the interval.

Exponential Distribution

One kind of probability distribution that is frequently used to describe waiting or inter-event durations is the exponential distribution. In the context of health and pharmacy, it can be used to indicate how long it will take for a particular event to occur or how long it will take for patients to arrive to a hospital under specific circumstances.

For instance, one may choose an exponential distribution to model the time before a specific event takes place or the interval between two instances of the same event.

Table 1: Difference Between Discrete and Continuous Random Variables

Basis	Discrete Random Variable	Continuous Random Variable
Nature of data	Obtained through counting	Obtained through measurement

Possible values	Countable and separate	Any value within an interval
Examples	Number of patients, number of adverse effects, number of tablets rejected	Blood pressure, body weight, temperature, drug concentration
Probability function	Probability mass function	Probability density function
Individual value probability	May be greater than zero	Usually equal to zero
Common models	Binomial, Poisson, geometric	Normal, uniform, exponential
Representation	Probability assigned to each value	Probability represented by area under a curve

It is essential to distinguish whether a particular random variable is continuous or discrete because it will be necessary to choose an appropriate probability model for further calculations. In case when a continuous model is applied to continuous data or count data is used instead of continuous one, wrong results may be obtained.

2.2.2. Probability Models in Healthcare Research

Probability models in health care research are mathematical modelling of unknown health, biological, and pharmaceutical processes. This is important to the researcher in trying to understand how often events take place and the chance of having certain outcomes as well as predicting possible future outcomes in certain conditions. Given that health care outcomes are always uncertain to some extent, probability models offer the researcher a way of handling uncertainty.

For instance, a drug may not work the same way on each patient. The drug could cause the patient to get total recovery, partial recovery, have no effect at all, or side effects. In addition, a test for diagnosing a disease may work on certain people and give false results on others.

1. Probability Models for Disease Occurrence

Models of probability may be applied in estimating the probability that an individual or group of people is going to develop diseases. Probability models can take into consideration several factors such as age, gender, genetics, way of life, environment, and medical background, among others. Estimating a person's likelihood of developing diabetes in five years while taking into account their body mass index, family history, blood sugar levels, and physical activity is an example of this.

When the variable of interest is the presence or absence of an illness, discrete probability models are frequently used to simulate the occurrence of disease. If there are only two possible outcomes, this could be done using a Bernoulli model:

- Disease present
- Disease absent

The binomial distribution can be used to determine the number of individuals in a given sample population who have contracted the disease when there are several observations.

2. Probability Models for Clinical Trials

A clinical trial is full of uncertainties because of differences between subjects in their reaction to the treatment applied. Probabilistic models are useful tools for estimating the number of patients who are likely to be sensitive to the drug, have side effects from the treatment, or quit from the experiment. Besides, the number of participants and the probability of detecting treatment effect in a sample are determined using probability models.

For instance, if it is expected that a new drug will improve patient symptoms in 70% of the population, the binomial probability model will help calculate the probability of the certain number of patients being responsive to the treatment during the clinical trial. These calculations are important for planning the study and evaluating the results of the experiment.

Finally, probability models play an important role in randomization procedures. By means of randomization each subject has a predetermined probability of getting a certain treatment.

3. Probability Models for Diagnostic Testing

Sensitivity, specificity, positive and negative predictive values, and other concepts of probability are used to analyse the test's diagnosis.

- The likelihood that a test would come back positive if a disease is present is known as sensitivity.
- Specificity is the likelihood that a test would come back negative if there were no illness.
- If the test is positive, the positive predictive value indicates the likelihood of having a disease.
- In the event that test findings are negative, the negative predictive value represents the likelihood of not having a disease.

It is crucial to utilise a probabilistic model in the process since the interpretation of test findings depends not only on the test results but also on the likelihood that the disease would exist in the population. If the illness is uncommon in the population, a test may occasionally result in false

positives while having excellent sensitivity and specificity. When diagnosing an illness and analysing its probability, the Bayesian Theorem is frequently utilised.

4. Probability Models for Pharmacokinetics

The branch of pharmacokinetics deals with the movement of drugs through the body and is concerned with such factors as absorption, distribution, metabolism, and excretion. The drug concentration, time of elimination, rate of absorption, and biological half-life can vary from one patient to another. Probability models are used by scientists to explain such variations in order to be able to make predictions about drug concentration depending on certain conditions.

For instance, using a model of probability, scientists can find out the probability that the concentration of the medication in blood stays in the therapeutic range after a certain dose of the drug has been applied. Pharmacokinetic models can contain random effects and residuals in order to show inter-patient variability and the variability due to the measurement process respectively.

5. Probability Models for Adverse Drug Reactions

Adverse drug reactions can happen at random during the course of medication. Models of probabilities could be used to predict the chances of occurrence of a particular adverse drug reaction and also the variables that cause the increase in risk of such a situation. This could be studied to ascertain the probability of the likelihood of development of liver toxicity in patients using a certain drug, several drugs or those that have liver disease.

The binomial distribution may be used for the estimation of the number of patients developing adverse effects, whereas the Poisson distribution could be used for estimating the number of adverse effects reported during a certain period. Logistic regression and survival models may be used for predicting the probability or time when adverse effects will occur. These models help pharmacovigilance in many ways.

6. Probability Models for Survival and Time-to-Event Data

A lot of interesting outcomes of researches in healthcare and pharmacology are dependent on time. It is possible to define the moment when a certain event occurs. Such events are time to death, time to disease recurrence, time to recovery, time to treatment failure, and time to hospital readmission. The data in which the time from a certain moment of start till occurrence of the certain event are usually called survival data or time-to-event data.

The survival analysis is especially important since the event of interest does not necessarily happen in each subject of the research. For example, in a clinical trial where the treatment for cancer is being investigated some subjects can have the disease recurrence during the study and others can remain recurrence-free till the end of the study. Thus, in such cases the exact event time cannot always be

known. With the help of survival analysis such time-related observations can be managed and the probability of absence of the event can be estimated.

There are several probability models which can be applied to describe survival and time-to-event data. They are the exponential distribution, Weibull distribution and the survival probability models. With the help of such models the researchers can investigate how the risk of the event changes over time and can estimate the probability of its occurrence or non-occurrence in a certain period of time.

The survival function is usually described as:

$$S(t) = P(T > t)$$

where:

- T= random variable representing the time until the event occurs
- t= specified time point
- $S(t)$= probability that the event has not occurred by time t

For example, if:

$$S(12) = 0.80$$

This implies that the predicted probability that a patient will be free from a specific event for more than 12 months is 80% in the research scenario.

The cumulative likelihood of an event happening by time t can also be compared to the survival function. If the chance that the event has happened by time t is F(t), then:

$$F(t) = P(T \leq t)$$

and:

$$S(t) = 1 - F(t)$$

Consequently, if the likelihood of an event happening within a year is 0.20, the likelihood of staying event-free after that time is:

$$S(12) = 1 - 0.20 = 0.80$$

A survival probability is applied in studies on cancer, clinical trials, infections, or any other chronic disease. For instance, researchers may compare time to disease relapse for two groups of people receiving different treatments. Also, researchers may analyze time until re-hospitalization among patients receiving different interventions.

In addition, survival analysis may help in identifying any factors that can be associated with the difference in time to an event. In this case, factors like treatment group, age, severity of the disease, clinical characteristics, and other patient covariates may be considered to see if they are associated with the increase or reduction of the time to an event.

Exponential distribution is one of the simplest types of the survival distribution and is based on constant event rate. The Weibull distribution is more flexible in the sense that it may present different shapes of risk depending on the changes in event rate. In this way, it is advisable to choose the probability model based on the data characteristics and assumptions made by the researcher.

One of the most remarkable features in the survival analysis is the occurrence of censoring. Censoring occurs in the situation when the actual event time cannot be observed for an individual within the study duration. It happens either when the person leaves the study without an event occurring or if he or she completes the study without the event occurrence. In such cases, even though we have a censored observation, there is some useful information contained since it shows that the event did not occur at least up to the moment the participant was observed. Therefore, the probability models of survival and time to an event help us to understand when events occur and probability of staying without them.

7. Probability Models for Hospital and Healthcare Services

Probability models have been adopted in the management of health care services. Hospitals can adopt probability models to make estimations for the number of patients coming, waiting time, bed availability, emergency cases, and workload on employees. For instance, the Poisson distribution model is used to make estimations of the number of patients that are expected to come to the hospital in one hour. Queueing models are adopted by hospitals to understand the movement of patients through registration, consultation, tests, and treatment.

8. Probability Models for Public Health and Epidemiology

Probability models are applied in public health studies where scientists assess the risk of an infectious disease spreading in a population and the efficacy of preventive measures. Probability models may determine the likelihood of infection, spread, recovery, or mortality. These models can also forecast the possible effect of vaccination, isolation, social distancing, and treatment initiatives.

For instance, models that are developed for an infectious disease might categorize people into susceptible, infected, and recovered. By using observations and presumptions in mathematical computations, the likelihood of changing states is ascertained.

9. Probability Models for Medical Decision-Making

In many cases, healthcare practitioners have to take decisions based on uncertain situations. Probability models enable the comparison of benefits and costs associated with various medical treatments. A decision-making model can incorporate probabilities of recovery, side effects, disease development, cost, and quality of life, among others.

In a certain situation, a doctor may want to compare two medical treatments. This is accomplished through comparing the probabilities of recovery and complications of both medical procedures. The medical treatment that is likely to yield more benefit at acceptable levels of risk is preferred.

2.3. IMPORTANT STATISTICAL DISTRIBUTIONS

A statistical distribution is a mathematical model that provides a description of the way values of a variable are distributed among a population or sample. All possible outcomes for a particular variable are described by distributions, together with the probabilities or frequencies connected to each outcome. Statistical distributions have several applications in health care and pharmaceutical research including understanding biological variability, laboratory measurement analysis, response to treatment analysis, incidence estimates and clinical events evaluations. Depending on the kind of data and the nature of the issue being resolved, probability distributions are employed. For instance, body weight and blood pressure measurements can be analyzed using normal distribution whereas the number of people responding to a particular treatment can be analyzed using binomial distribution. Similarly, the number of patients experiencing side effects and those being admitted into hospitals during a given period of time can be analyzed using a Poisson distribution.

Statistical distributions can be presented in various ways including through tables, graphs, probability functions or frequency curves. The central tendency, variance, and shape of the data are all shown by statistical distributions. Because it affects the validity of statistical tests, predictions, and conclusions, choosing an appropriate statistical distribution is crucial. Because they are the most widely utilised distributions, the normal, binomial, and Poisson distributions are crucial to pharmacological and clinical research.

2.3.1. Normal Distribution in Biological and Pharmaceutical Measurements

Normal distribution is among the most important and popular probability distributions in statistical analyses. Normal distribution is defined as a continuous probability distribution exhibiting the shape of a symmetrical bell curve. Normal distribution is used in statistics to describe quantities that are affected by a number of factors. Variables like height, weight of the body, blood pressure, body temperature, lab values, drug concentration in the body, weight of tablets, and errors of measurements could approximately have normal distribution under proper conditions.

Normal distribution is a function of two parameters:

- **Mean** (μ)**:** Represents the centre or location of the distribution.
- **Standard deviation** (σ)**:** Represents the spread or variability of observations around the mean.

The normal distribution's probability density function is:

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

where:

- x= observed value
- μ= population mean
- σ= population standard deviation
- e= base of the natural logarithm
- π= mathematical constant approximately equal to 3.14159

The distribution's location is determined by the mean, while its dispersion from the mean is determined by the standard deviation.

1. Characteristics of the Normal Distribution

There are a number of features that distinguish normal distribution, which makes it quite convenient for biological and pharmaceutical statistical purposes.

- **Symmetry:** The normal distribution has symmetry with respect to its mean. In other words, the distribution's left and right sides have the same forms. In other words, observations are dispersed uniformly about the center.
- **Equality of mean, median, and mode:** When a distribution is absolutely normal, the mean, median, and mode all coincide and equal the distribution's center**.**
- **Bell-shaped form:** The distribution's characteristic bell-shaped form. The majority of observations are close to the mean, and less and fewer are farther away.
- **Total area equals one:** The normal distribution curve's total area is equal to one:

$$\int_{-\infty}^{\infty} f\,(x)\,dx = 1$$

It shows the likelihood of every scenario the random variable could have.

- **Two tails tend to extend infinitely:** In a normal distribution, the two tails tend towards infinity both on the negative side and positive side of the number line. The distribution tends to touch the x-axis but doesn't do so.
- **Calculated using mean and standard deviation:** The standard deviation is used to determine the distribution's spread, whereas the mean is used to determine the distribution's center point or location.

- **Unimodal:** A normal distribution has a single hump or mode. Thus, it is a unimodal distribution.

2. Empirical Rule of the Normal Distribution

The empirical rule, commonly known as the 68-95-99.7 rule, applies to the normal distribution. Assuming that the data is normally distributed, this rule describes the proportion of observations that fall within one, two, and three standard deviations of the mean.

Approximately 68% of data are within one standard deviation of the mean:

$$\mu - \sigma \leq X \leq \mu + \sigma$$

Two standard deviations around the mean contain about 95% of the observations:

$$\mu - 2\sigma \leq X \leq \mu + 2\sigma$$

Three standard deviations around the mean contain about 99.7% of the data:

$$\mu - 3\sigma \leq X \leq \mu + 3\sigma$$

For example, if a sample's mean systolic blood pressure is 120 mmHg and its standard deviation is 10 mmHg, then approximately 68% of the data will fall within:

$$120 - 10 = 110$$

and:

$$120 + 10 = 130$$

Thus, approximately 68% of individuals would have systolic blood pressure between 110 and 130 mmHg.

Within two standard deviations:

$$120 - 2(10) = 100$$
$$120 + 2(10) = 140$$

Therefore, approximately 95% of observations would fall between 100 and 140 mmHg.

Within three standard deviations:

$$120 - 3(10) = 90$$
$$120 + 3(10) = 150$$

Therefore, if the data is regularly distributed, around 99.7% of the measurements will fall between 90 and 150 mmHg.

The dispersion of normally distributed biological and pharmacological measurements can be understood using the empirical rule as a guide.

3. Standard Normal Distribution and Z-Score

A z-score can be used to convert any normal distribution into a conventional normal distribution. A z-score displays the number of standard deviations that separate the data from the mean.

Z score formula is represented as follows:

$$z = \frac{x - \mu}{\sigma}$$

where:

- x= observed value
- μ= mean
- σ= standard deviation

A positive z-score indicates that the observation is higher than the mean; a negative z-score indicates that the observation is lower than the mean.

For example, let us take the mean plasma drug concentration as 50 mg/L, standard deviation of 5 mg/L and the observed value from the patient is 60 mg/L. Then the z-score is:

$$z = \frac{60 - 50}{5}$$

$$z = \frac{10}{5} = 2$$

As a result, the patient's medication level deviates from the mean by two standard deviations.

When comparing data and their distributions, identifying exceptionally high or low values, and calculating probabilities using the conventional normal distribution, Z-scores are useful.

4. Normal Distribution in Biological Measurements

Biological measurements inherently differ from individual to individual. This may include differences in height, weight, blood pressure, pulse rate, haemoglobin level, and blood glucose concentration. Where several minor factors contribute to the variability of a measurement, then under some conditions, the measurement's overall distribution may tend to be normally distributed.

The normal distribution may be applied in:

- Capturing the central tendency of a population's characteristics.
- Obtaining estimates of proportions of individuals within some interval.
- Detecting unusual values.
- Establishing reference ranges for laboratory measurements.
- Comparing individual observations with those of populations.
- Helping in conducting statistics-based tests.

Nonetheless, researchers cannot take for granted the fact that all biological variables are normally distributed. Variables such as time spent in hospital, income, viral load, and some hormone concentrations may be significantly skewed either positively or negatively. Others may be strongly affected by outliers.

Where observed data fail to follow a normal distribution pattern appropriately, researchers may apply a suitable data transformation method or another distribution/statistical technique, based on research objectives and nature of data.

5. Normal Distribution in Pharmaceutical Measurements

Normal Distribution is common in pharmaceuticals in areas like quality control, manufacture, laboratory testing, and development. Variables like tablet weight, capsule fill weight, assay, dissolution, content uniformity, and measurement errors in the analytical procedure are approximately normally distributed in favourable manufacture or measurement conditions.

As an illustration, suppose we have tablets labelled to have a weight of 500 mg. some deviations might arise in the manufacture of these tablets, hence causing variations from the target value in the weight of individual tablets. The mean and standard deviation can be used to assess the consistency of the manufacturing process if this variable is roughly regularly distributed.

Normal distribution is also applicable in the following scenarios:

- Drug concentrations in the bloodstream and plasma analysis.
- Results of assays and analytical methods evaluation.
- Measurement error assessment.
- Quality control of pharmaceuticals.
- Mean values in the comparison of treatment groups.
- Confidence intervals calculation.
- Bioequivalence measurements analysis.
- Dissolution and stability measurements evaluation.

Nonetheless, the use of normal distribution has to be assessed depending on the data set and measurement process.

6. Normal Distribution and Reference Intervals

The reference range is meant to define the range that represents values observed in a particular reference population, such as a healthy population. The mean and standard deviation can be used to characterise a laboratory result that is roughly regularly distributed.

In a population with a normal distribution, the most typical theoretical 95% range is:

$$\mu \pm 1.96\sigma$$

Assuming a normal distribution, approximately 95% of the data are included in the confidence interval.

Take an illustration; let the average haemoglobin level in a population be:

$$\mu = 14 \text{ g/dL}$$

and the standard deviation is:

$$\sigma = 1 \text{ g/dL}$$

The approximate 95% interval is:

$$14 \pm 1.96(1)$$

Therefore:

$$14 - 1.96 = 12.04$$

and:

$$14 + 1.96 = 15.96$$

Thus, the theoretical interval is approximately:

$$12.04 \text{ to } 15.96 \text{ g/dL}$$

However, when dealing with real laboratory medicine, the reference intervals must be calculated using proper reference populations, proper procedures, adequate number of samples, and laboratory guidelines, not just the normal distribution formula.

As far as the importance is concerned, the normal distribution can be considered as an important concept that serves as a basis of many other concepts. Calculating and analysing the mean and standard deviation, z-scores, probabilities, confidence intervals, and other methods are made easier by the characteristics of the normal distribution. It is necessary to evaluate the normalcy assumption against actual circumstances, though.

2.3.2. Binomial and Poisson Distributions in Clinical Research

Two important discrete probability distributions that are frequently used to examine discrete outcome possibilities in clinical, pharmaceutical, and medical research are the Binomial and Poisson distributions. The Poisson distribution is used when researchers are attempting to measure the number of events occurring within a particular time period, area, volume, or population, whereas the binomial distribution is typically used when researchers are attempting to measure the number of successes in the predetermined number of trails.

1. Binomial Distribution

The number of successes in a predetermined number of independent trials with two possible outcomes per trial is known as the binomial distribution. Success or failure, positive or negative, recovered or not recovered, treated or not treated, and adverse reaction or no adverse reaction are all possible outcomes.

- The binomial distribution's prerequisites are:
- There are always the same number of trials.
- Each trial has two possible results.
- Trials don't depend on one another.
- The likelihood of success is constant throughout trials.
- The quantity of successes is the random variable.

The probability mass function is as follows if X is the number of successes from n trials:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where:

- n= total number of trials
- x= number of successes
- p= probability of success
- $1 - p$= probability of failure
- $\binom{n}{x}$= number of possible arrangements of $x$successes among $n$trials

The combination term is calculated as:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

For instance, if there is a clinical study involving 10 subjects where every subject has an identifiable chance of responding to medication, If the assumptions of the binomial distribution are sufficiently met, then the number of such responding respondents would form a binomial random variable.

2. Mean and Variance of the Binomial Distribution

The mean, or expected value, of a binomial random variable is:

$$E(X) = np$$

The variance is:

$$Var(X) = np(1-p)$$

The standard deviation is:

$$\sigma = \sqrt{np(1-p)}$$

For instance, if a medication has a probability of treatment success of 0.80 and ten people are treated with the medication, the expected value is:

$$E(X) = np$$
$$E(X) = 10(0.80) = 8$$

Therefore, the expected number of successful outcomes is 8 patients.

The variance is:

$$Var(X) = 10(0.80)(1 - 0.80)$$
$$Var(X) = 10(0.80)(0.20) = 1.6$$

and the standard deviation is:

$$\sigma = \sqrt{1.6} \approx 1.26$$

Therefore, according to the given binomial distribution, the mean of success is 8 while the standard deviation is approximately 1.26 patients.

3. Binomial Distribution in Clinical Trials

The binomial distribution is frequently applied when researchers decide whether a subject will have a certain outcome. Some examples are:

- Whether the patient will respond or not to a drug.
- Whether the patient will recover or not.
- Whether the patient will have or not have an adverse reaction to a drug.
- Whether the patient will test positive or not to the disease.
- Whether the patient will finish or not finish the research.
- Whether the patient will have or not have a specific event.

Suppose the response probability is 0.70 and the treatment is applied to five subjects. Then, the probability of exactly three responses is:

$$P(X = 3) = \binom{5}{3}(0.70)^3(0.30)^2$$

Since:

$$\binom{5}{3} = 10$$

the probability becomes:

$$P(X = 3) = 10(0.70)^3(0.30)^2$$
$$P(X = 3) \approx 0.3087$$

According to the binomial distribution, the likelihood that exactly three people will respond to the treatment is approximately 0.309, or 30.87%.

4. Applications of the Binomial Distribution

The following are a few situations where the binomial distribution can be applied in the pharmaceutical industry:

- Responses to clinical trials.
- Success rate in treatments.
- Adverse reactions to drugs.
- Result of diagnostic tests.
- Vaccination efficacy studies.
- Patient compliance studies.
- Quality control tests in pharmaceuticals.
- Success or failure results from a certain number of participants.
- The proportion of those who improve their condition.

The binomial distribution is especially suitable in cases where each participant provides only one binary result. However, the use of the binomial distribution in some cases can be unsuitable, especially when there is variation in the probability of success in various trials.

5. Poisson Distribution

Poisson distribution is used to estimate the number of occurrences of some events that take place over an interval of time and are independent of each other and have constant mean rates of occurrence.

Some of the examples where Poisson distribution can be applied are:

- Number of patients attending the emergency room per hour.
- Number of adverse drug reaction cases reported per month.
- Number of hospital admissions per day.
- Number of infections that may develop in a particular population.
- Number of bacterial colonies in a laboratory test specimen.
- Number of medication errors per specified time.
- Number of diseases that have been reported from a certain geographical location.

The Poisson distribution's probability mass function is provided by:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where:

- x= number of events
- λ= average number of events in the specified interval
- e= base of the natural logarithm
- $x!$= factorial of x

The predicted number of events during the given timeframe is represented by the parameter λ.

6. Mean and Variance of the Poisson Distribution

The Poisson distribution's equal mean and variance is a key feature:

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

The standard deviation is:

$$\sigma = \sqrt{\lambda}$$

For example, that a hospital sees four emergency patients every hour on average. Then:

$$\lambda = 4$$

The expected number of emergency patients per hour is therefore:

$$E(X) = 4$$

The variance is also:

$$Var(X) = 4$$

The standard deviation is:

$$\sigma = \sqrt{4} = 2$$

For the given interval, the Poisson model has a mean of four occurrences, a variance of four, and a standard deviation of two events.

7. Poisson Distribution in Clinical Research

Clinical events that occur within a specified time frame can be analysed using the Poisson distribution. This includes counting the number of adverse effects that occur during a clinical trial, the number of hospitalisation cases that occur in a given day, or the number of illness cases that occur within a specified time frame.

If there are three medication errors in a hospital per week, then:

$$\lambda = 3$$

Two medication errors occurring in a given week are likely to be:

$$P(X = 2) = \frac{e^{-3}3^2}{2!}$$

This gives:

$$P(X = 2) \approx 0.2240$$

This means that the probability of having exactly two mistakes with medicines in one week, using the above Poisson distribution model, is about 22.40%.

These types of calculations may prove useful to healthcare researchers and managers who need to know how often certain events occur.

8. Applications of the Poisson Distribution

The Poisson distribution finds application in several areas in pharmaceutical sciences and health care research, such as:

- Hospital admissions and patient arrivals.
- Disease monitoring.
- Epidemiological studies.
- Pharmaceutical safety and adverse events reporting.
- Event analysis in clinical trials.
- Infection control studies.
- Error analysis in medicine and drugs.
- Microbiology in laboratories.
- Health service demands.
- Resource management in hospitals.

Research questions concerning the interval's event frequency rather than an individual's suffering from that event are well-suited to the Poisson distribution.

9. Relationship Between the Binomial and Poisson Distributions

Although they are both discrete probability distributions, binomial and Poisson probability distributions are employed in distinct contexts. When the number of trials is known, each trial produces two results, and the statistician counts the number of successes, the binomial probability distribution is appropriate. The Poisson probability distribution, on the other hand, works well when researchers count the number of occurrences within a given time frame and the number of possible occurrences is unknown in terms of the number of trials.

There is a relationship between the two distributions. The Poisson distribution can be used to approximate the binomial probability distribution if specific requirements are met:

- $n$is large.
- $p$is small.
- The product $np$is moderate.

The Poisson parameter is then approximated by:

$$\lambda = np$$

Consider, for example, a highly rare adverse drug reaction that has a very low chance of happening among a large number of individuals. Sometimes the Poisson probability distribution formula can be used to ascertain the number of people that experienced such a reaction instead of utilising the binomial probability distribution formula to determine the necessary probability.

It should be mentioned that the Poisson approximation can only be used under certain circumstances. To put it another way, if there is a definite number of independent trials with an equal chance of success, the binomial distribution formula should be applied.

In general, the Poisson distribution examines the number of occurrences within a certain time frame, whereas the binomial distribution deals with the number of successes from a given number of trials.

Table 2: Difference Between Binomial and Poisson Distributions

Basis	**Binomial Distribution**	**Poisson Distribution**
Type of data	Number of successes	Number of events
Number of trials	Fixed	Not necessarily fixed
Outcomes	Success or failure	Event occurrence
Probability	Constant probability of success	Constant average event rate
Main parameter	(n) and (p)	(\lambda)
Mean	(np)	(\lambda)
Variance	(np(1-p))	(\lambda)
Example	Number of patients responding to treatment	Number of adverse events reported per month
Common use	Clinical response and diagnostic outcomes	Disease cases, hospital arrivals, and event counts

2.4. PROBABILITY DISTRIBUTIONS USING PYTHON

The Python language is one of the most popular software tools used in healthcare, pharmaceutics, biology, and other clinical fields in order to conduct data processing and statistical analyses. There are numerous libraries available in Python which can help to create, analyze, and interpret probability distributions, namely NumPy, Pandas, SciPy, Matplotlib, and Seaborn. A probability distribution displays a random variable's potential values together with the likelihood that these values will occur.

By using Python, it will be possible to make probability calculations, generate random observations, make calculations of descriptive statistics, analyze theoretical probability distributions, and compare statistical models with real data from the pharmaceutic industry. Also, by using Python, one may automatize repetitive calculations and apply the same process of analyzing data to several sets.

Probability distribution in the field of pharmaceutics can be used for studying of various variables and events in the sphere, such as drug concentration, weight of tablets, patient response to treatment, adverse drug reactions, hospital admission rates, disease incidence rates, lab results, and time to event.

Some popular probability-distribution analysis Python libraries are:

- **NumPy:** Employed for performing numerical computations, working with arrays, and generating random variables based on probability distributions.
- **Pandas:** Helps organize, clean, filter, and analyze structured data of pharmaceuticals and clinical studies.
- **SciPy:** Supplies the functions of probability distribution, probability calculations, parameter estimates, and many other statistics methods.
- **Matplotlib:** Used to generate graphs such as histograms and theoretical distribution plots.
- **Seaborn:** Offers extra visualization tools and allows generating distribution plots.

A typical procedure for carrying out probability-distribution analysis using Python will entail:

1. Loading of the necessary Python libraries.
2. Preparation and examination of the data set.
3. Determining if the variable is discrete or continuous.
4. Specifying a suitable probability distribution.
5. Either estimating or specifying the necessary distribution parameters.
6. Computing of probabilities or statistics.
7. Model fit to the data if applicable.
8. Interpretation of the findings in connection with the pharmaceutical/clinical study.

For instance, NumPy could be used to sample from a normal distribution:

```python
import numpy as np

data = np.random.normal(loc=50, scale=5, size=1000)
```

In this code:

- loc=50 specifies the mean of the distribution.
- scale=5 specifies the standard deviation.
- size=1000 specifies that 1,000 observations should be generated.

The generated observations can be examined using descriptive statistics. For example:

```python
np.mean(data)
np.std(data)
```

These techniques calculate the sample's mean and standard deviation. The calculated mean and standard deviation should be close to, but not precisely equal to, 50 and 5, respectively, because the data were produced at random.

In case the need arises for graphically examining the observations, they may be plotted on a histogram using Matplotlib library:

```python
import matplotlib.pyplot as plt

plt.hist(data, bins=30, density=True)

plt.xlabel("Measurement")
plt.ylabel("Density")
plt.title("Normal Distribution")

plt.show()
```

In this case, the histogram serves as a graphical tool that helps to understand how the generated observations are distributed. The use of density=True option ensures that the histogram reflects the density rather than just the number of observations.

When it comes to pharmaceutical research, the same procedure can be applied to any research data. It means that researchers can generate or analyze normally distributed data which include such

measurements as tablet weight or lab results and other approximately continuous variables. At the same time, SciPy can be helpful to researchers who need to work with such theoretical probability distributions as normal, binomial, Poisson, exponential, and Weibull distributions.

It should be noted that there is a difference between generating random data according to a certain probability distribution and analyzing real pharmaceutical data. Randomly generated data will help to learn statistics and visualize some probability models; however, real conclusions should be made on the basis of properly collected research data.

Thus, Python allows researchers to easily deal with the probability-distribution analysis. With the help of numerical computation, data management, statistical calculations, and, when needed, visualization, researchers can easily analyze probability models.

2.4.1. Graphical Visualization of Probability Distributions

Graphical visualization is defined as the representation of probability distributions and observed data through use of graphs, diagrams and curves. Graphical visualization makes it easier for researchers to grasp the nature, center, dispersion, variability and exceptional features of data rather than numeric representations alone. Graphical visualization can help researchers to ascertain whether the observations are centered around some value, the degree of symmetry or skewness of a distribution and whether outliers or group differences may be present.

In health care and pharmaceutical research, graphical visualization may be applied in analyzing variables such as concentrations of a drug, blood pressure, blood glucose levels, time of dissolution, response to treatment and number of adverse events. Graphical visualization may also be applied by researchers in determining whether an assumed probability distribution like normal, binomial or Poisson distribution is reasonable to the observations made.

1. Histogram

One of the most popular ways to plot the distribution of numerical variables graphically is through the use of a histogram. A histogram splits the observations into intervals referred to as bins and plots their frequency or density in those bins. Each bar's height indicates the quantity or percentage of observations within that bin.

The following is an example of how to plot a histogram using Python and Matplotlib:

```python
import matplotlib.pyplot as plt

plt.hist(data, bins=20, edgecolor="black")

plt.xlabel("Observed value")
plt.ylabel("Frequency")
plt.title("Histogram of Observed Measurements")

plt.show()
```

When density=True is used, the histogram is scaled to represent probability density rather than frequency:

```python
plt.hist(data, bins=30, density=True, edgecolor="black")

plt.xlabel("Observed value")
plt.ylabel("Probability density")
plt.title("Probability Distribution")

plt.show()
```

It is also possible to compare the probability density function with the density scaled histogram. The number and width of bins chosen to plot a histogram can have an impact on the histogram's appearance, therefore the researcher shouldn't make any inferences about the histogram's form based on a single bin choice.

A preliminary impression of the data distribution's shape, such as whether the data are uniformly distributed, skewed, regularly distributed, or randomly distributed, may be provided by the histogram. Nevertheless, graphical depiction should always be complemented by numerical summary and statistical analysis, when necessary.

2. Probability Density Curve

A smooth mathematical function that simulates the probability distribution of a continuous random variable is known as the probability density function. The area between any two locations indicates the likelihood that the random variable will fall between those points, and the sum of the areas under the probability density function equals 1.

For example, SciPy can be used to build the normal probability density function:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

x = np.linspace(30, 70, 500)

y = norm.pdf(x, loc=50, scale=5)

plt.plot(x, y)

plt.xlabel("Measurement")
plt.ylabel("Density")
plt.title("Normal Probability Density Curve")

plt.show()
```

In this example:

- loc=50 specifies the mean.
- scale=5 specifies the standard deviation.
- norm.pdf () calculates the probability density for the specified values of x.

The density curve thus generated is symmetric with a bell-shaped form, with the centre of the curve being dictated by the mean and the spread by the standard deviation.

Alternatively, a histogram of the data can also be overlaid on the normal density curve:

```python
plt.hist(data, bins=30, density=True, alpha=0.6)

x = np.linspace(min(data), max(data), 500)

y = norm.pdf(x, np.mean(data), np.std(data))

plt.plot(x, y)

plt.xlabel("Measurement")
plt.ylabel("Density")
plt.title("Observed Data and Normal Distribution")

plt.show()
```

This comparison gives a first approximation on how well the data observed are normally distributed. But even if there is very good correspondence between the two, we cannot simply use this observation as evidence of our assumption being valid.

3. Bar Chart for Discrete Probability Distributions

A bar chart is one of the ways that discrete probability distributions are often presented since discrete random variables assume distinct values. For instance, the number of individuals that respond to a drug could have distinct values like 0, 1, 2, 3, and so forth.

Representation of binomial distribution using SciPy:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom

n = 10
p = 0.7

x = np.arange(0, n + 1)

probabilities = binom.pmf(x, n, p)

plt.bar(x, probabilities)

plt.xlabel("Number of successes")
plt.ylabel("Probability")
plt.title("Binomial Probability Distribution")

plt.show()
```

Here:

- n=10 represents 10 trials.
- p=0.7 represents a probability of success of 0.70.
- binom.pmf() calculates the probability associated with each possible number of successes.

The chart formed from this is one that indicates the probabilities of getting various numbers of successful results from a sample of 10 patients.

4. Visualization of the Poisson Distribution

A model that characterises the quantity of events in a given time frame is the Poisson distribution. Data such as the number of hospital admissions per day, the number of adverse drug reactions per month, the number of emergency cases per hour, and the number of prescription errors within a given time frame may be counted using the distribution.

The Poisson distribution can be expressed in terms of:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import poisson

lam = 4

x = np.arange(0, 15)

probabilities = poisson.pmf(x, mu=lam)

plt.bar(x, probabilities)

plt.xlabel("Number of events")
plt.ylabel("Probability")
plt.title("Poisson Probability Distribution")

plt.show()
```

It is evident from this example that lam=4 indicates that the average number of occurrences throughout the specified time frame is four. The probability of different occurrence numbers is displayed in these bars.

It is important to note that the average λ value affects the Poisson distribution's form. It tends to be more symmetrical at high λ values and righter skewed at low λ values.

5. Box Plot

A box plot, sometimes referred to as a box-and-whiskers plot, is a visual depiction that provides an overview of the data values. The median, lower quartile, upper quartile, and potential outlier data values are displayed.

A box plot may be constructed from:

```python
plt.boxplot(data)

plt.ylabel("Measurement")
plt.title("Box Plot of Observed Data")

plt.show()
```

The box contains 50% of the data and displays the IQR. The median is represented by the line inside the box. The data points that fall under the software's plotting rule are shown by the whiskers, while those outside the whiskers may show up as possible outliers.

Box plots are especially helpful when comparing numerical distributions among groups. This is common when comparing data like blood pressure of patients in different groups or drug concentration in patients receiving different dosages.

Box plots can provide the following information about:

- The median value.
- The interquartile range.
- The variability of observations.
- Potential outliers.
- Differences in the distribution between groups.

Outliers detected using a box plot need not necessarily be deleted. The researcher must assess the observation itself, the method used to obtain it, and the science behind it before determining if the observation is an error or a valid outlier in its own right.

6. Cumulative Distribution Function

The probability that a random variable will be less than or equal to a given value is measured by the Cumulative Distribution Function. It might be stated as:

$$F(x) = P(X \leq x)$$

The total area below the probability density function up to x is the CDF for a continuous distribution.

SciPy may be used to calculate the CDF of the normal distribution in Python:

```python
from scipy.stats import norm

probability = norm.cdf(55, loc=50, scale=5)

print(probability)
```

This calculates:

$$P(X \leq 55)$$

for a variable with a mean of 50 and a standard deviation of 5 that is regularly distributed.

Two CDF values are subtracted to get the probability that a random variable falls between two points:

```python
probability = norm.cdf(60, loc=50, scale=5) - \
              norm.cdf(40, loc=50, scale=5)

print(probability)
```

This calculates:

$$P(40 \leq X \leq 60)$$

Consequently, CDF is useful when scientists need to determine the likelihood that a pharmacological or biological test result will fall below a predetermined threshold or fall within a specific range.

For example, CDF can be used to estimate the likelihood that a drug's concentration will be below the specified value, that laboratory results will fall within the reference range, or that biological measurements will fall within a specific clinical range.

Visualization through graphs serves as an important tool alongside the analysis of probabilities. While histograms serve as a good tool to analyze numerical distributions, probability density functions serve as continuous theoretical distributions, bar charts help to depict discrete probability distributions, box-and-whisker diagrams summarize distribution properties as well as possible outliers, and CDF depicts cumulative probability distribution. All these techniques come handy in analyzing data and probability models before undertaking formal statistical analysis in pharmaceutical studies.

2.4.2. Interpretation of Distribution Patterns and Results

The interpretation of distribution pattern requires understanding the shape, center, spread, and features of the probability distribution to see what information it holds. While in the fields of pharmaceutical and biological sciences, mere creation of a chart or finding probabilities will not suffice; researchers need to interpret findings based on the research problem and features of the research population. The use of Python in this case would allow one to find descriptive statistics, histograms, box plots, probability functions, and probability based on the distribution; however, the final interpretation would require statistical and scientific interpretation.

1. Interpretation of the Centre

The central value of the data points is referred to as the distribution's center. The mean, median, and mode are a few typical measures of the distribution's center. The mode is the highest frequency of occurrence, the mean is the average of the values, and the median is the intermediate value when the data points are organised in either ascending or descending order.

The mode will be located at the center of a symmetrical distribution, particularly if it is normally distributed, and the mean and median will be comparatively close. Significant differences between the mean and the median indicate that the distribution may not be symmetrical and that extreme values may exist.

Assume, for example, that some patients had a mean plasma medication concentration of 50 mg/L and a median concentration of roughly 50 mg/L. The distribution is comparatively symmetrical, as evidenced by the similarity between these two indices. However, if the median concentration is 50 mg/L and the mean concentration is 58 mg/L, this discrepancy could indicate the presence of some high outliers. Researchers can learn more about medication concentration, tablet weight, blood pressure, treatment response, and other measurable variables by interpreting the center.

2. Interpretation of Spread

The degree of variability within observations is indicated by the distribution's dispersion. The mean value of two sets of data may be identical, but their dispersion may differ noticeably. Range, variance, standard deviation, and interquartile range (IQR) are a few often used metrics for assessing dispersion.

Because it calculates the average deviation of observations from the mean in the same units as the measurement, the standard deviation is a crucial indication that shows how far off the observations are from the mean on average. The observations are closer to the mean when the standard deviation is smaller, and vice versa.

Consider two treatment groups, for example:

Table 3: Comparison of Mean Response and Standard Deviation Between Treatment Groups

Group	Mean Response	Standard Deviation
Treatment A	50	4
Treatment B	50	12

Response averages for both groups are equal at 50. Nevertheless, Treatment A is characterized by significantly smaller variance, suggesting high consistency in patients' responses. Treatment B demonstrates higher variability in patients' responses.

Low variability of tablet weight, content, dissolution time, or drug concentration in pharmaceutical manufacture might suggest higher consistency in manufacturing/measurement processes.

3. Interpretation of Symmetry

Distribution is regarded as being symmetrical when the observations on either side of its center demonstrate roughly the same patterns. Normal distribution is the most common type of symmetrical distribution.

A relatively equal mean and median combined with a bell-shaped histogram can suggest that the data are regularly distributed. The application of normal-based statistical methods might be justified in such a situation.

For example, the distribution will be nearly normally distributed if the tablet weight is scattered uniformly around the target weight and the histogram is substantially bell-shaped.

However, as the number of bins and sample size have an impact on histograms, researchers should never rely just on visual evaluation to identify the distribution of the data. Formal tests for determining normalcy must also be considered, along with additional numerical descriptions and graphical assessments like Q-Q plots.

The normality of the data is also dependent on the objective of the study as some statistical tools may be robust or not very sensitive to deviation from normality while others are highly sensitive to it.

4. Interpretation of Skewness

Skewness is an indicator that describes the degree and direction of asymmetry in the probability distribution.

- **Positively skewed distribution –** a distribution with a long tail on the high-value right side. Mean is often larger than the median.
- **Negatively skewed distribution –** a distribution with a long tail on the low-value left side. Mean is often smaller than the median.
- **Zero skewness –** a nearly symmetrical distribution; however, zero skewness is not enough evidence for the normal distribution.

One of the most popular examples in pharmaceutical and clinical sciences is the hospital stay duration. A great majority of patients may have a short hospital stay, while only a few may stay much longer. As a result, we can discuss the long tail on the right side and positive skewness.

The median and interquartile range may indicate central tendency and variation more accurately than the mean and standard deviation in severely skewed data. However, depending on the type of study and its objectives, researchers should consider potential changes.

Therefore, rather than following a general rule, the choice between using a mean-based or median-based measure of central tendency depends on the distribution itself and the purpose of the study.

5. Interpretation of Peaks and Multiple Modes

Mode refers to a certain point/area where data tends to be clustered. The term for a distribution with only one peak is "unimodal" distribution, the term for a distribution with two peaks is "bimodal" distribution, while the term used for a distribution with multiple peaks is "multimodal."

Bimodal distributions may mean that the data points belong to two separate populations. For instance, in case the data points regarding drug concentration levels of subjects have two peaks, then there may be a possibility that the individuals have significantly different metabolic profiles.

Multiple peaks may result from:

- Different treatment groups.
- Differences in age or sex.
- Different disease categories.
- Different dosage levels.
- Data collected at different locations.
- Different laboratory or measurement conditions.
- Combining datasets from different populations.

In case there are several peaks, it is important to examine whether there are any groups that can be distinguished. Splitting of the data based on some clinical or scientific features may help clarify the nature of the distribution.

Nevertheless, researchers should not split a data set in the case when a histogram is bimodal. In addition, the peaks could be generated by a sample variability or the bin size.

6. Interpretation of Outliers

Outliers refer to data that is extremely far away from the other data. They can be detected either by the use of graphs like box plots or numerically through the use of interquartile range.

An outlier may arise because of:

- Measurement error.
- Data-entry error.
- Incorrect units.
- Sample contamination.
- Incorrect dosing information.
- An unusual patient characteristic.
- A genuine extreme biological response.

For instance, an unexpected high value for the plasma drug concentration can be due to a laboratory analysis error, inaccurate dosing, interactions which may interfere with drug metabolism, decreased drug clearance, or an actual physiological situation.

Accordingly, an outlier cannot be deleted blindly. Researchers need to investigate the record, the laboratory documentation, the dosage, and the data entry process. In case the outlier turns out to be true, it might have some important scientific knowledge and hence should be kept unless there are reasons to delete it based on an established scientific approach. Python can help to detect the possible outliers by generating box plots and IQR values among other scientific analyses.

7. Interpretation of Normal Distribution Results

In case where there is a bell shaped and symmetrical histogram together with density curve with mean and median being equal, then the distribution can be considered to be normally distributed.

Assume that tablets weigh approximately the target weight and follow normal distribution. In case where the mean weight of the tablets is 500 mg with the standard deviation of 10 mg, majority of the observations will be found close to 500 mg.

The following approximate ranges are expected for approximately normally distributed data according to the empirical rule:

$$\mu \pm \sigma$$

contains about 68% of observations,

$$\mu \pm 2\sigma$$

contains about 95% of observations, and

$$\mu \pm 3\sigma$$

contains 99.7% of the data points.

Hence, with μ = 500 mg and σ = 10 mg, about 95% of the observations will

$$500 \pm (2 \times 10)$$
$$= 480 \text{ mg to } 520 \text{ mg}$$

This information could be used by scientists to determine expected ranges and analyze drug measurement results.

Normal distribution will also justify the application of statistical methods like t-test, analysis of variance, and confidence interval, based on the assumptions that the particular test and experimental design are founded on. Normality must be determined with regard to sample size, design, analysis, and the particular variable.

8. Interpretation of Binomial Distribution Results

When there are only two possible outcomes for an experiment, the binomial distribution describes the likelihood of success for a certain number of experiments within a fixed number of trails.

The function of probability is:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $n$is the number of trials and $p$is the probability of success.

The expected number of successes is:

$$E(X) = np$$

Suppose a medicine has a response probability of 0.80 and is administered to 20 patients. The expected number of responders is:

$$E(X) = np$$
$$= 20(0.80) = 16$$

Therefore, there will be an average of 16 responders using the given binomial distribution. It is not implied that 16 patients have to respond every time in the samples; it can be 14, 15, 17, 18, etc., due to randomness.

If pis large, then the binomial distribution is more concentrated around large values of success. If pis small, then it is more concentrated around small values of success.

9. Interpretation of Poisson Distribution Results

The Poisson distribution is used for modelling the number of occurrences of an event in a certain interval of time, space or another observational unit.

The probability function of the Poisson distribution is:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where λrepresents the expected number of events in the specified interval.

An important property of the Poisson distribution is:

$$E(X) = \lambda$$

and

$$Var(X) = \lambda$$

For example, suppose a hospital records an average of five emergency admissions per hour. Then:

$$\lambda = 5$$

The model can be employed to determine the probability of getting zero, one, two, five, or other admission numbers in an hour.

The researcher must also make sure that the observed features match the features of the Poisson distribution. In case the variance is significantly higher compared to the mean, there may be overdispersion of data. If this is the case, then a Poisson model may fail to account for the variability, and other models, such as negative binomial distribution, must be taken into account.

10. Interpretation of Probability Values

The term "probability" refers to the measure of an event's chances of occurring under certain conditions and is between 0 and 1.

- The closer to 0 the probability, the more improbable the event is.
- The closer to 1 the probability, the more probable the event is.
- The probability 0.5 means that there are equal chances of an event happening or not happening.

For example:

$$P(A) = 0.90$$

means that the estimated probability of event $A$is **90%** under the specified conditions.

Similarly,

$$P(A) = 0.10$$

Probability of 10%

Probability must always be explained in light of the population, sample, time period, experiment, and probability model assumptions. Just because probability is derived from a statistical model does not mean it guarantees whether or not the outcome will happen to one person.

11. Interpretation of Model Fit

The concept of model fit can be described as the degree of match between the probability distribution and the observed data. It is important that a good model would properly reflect such characteristics as central tendency, variability, shape, and frequency of occurrences (including outliers if necessary).

Possible methods of checking model fit include:

- Comparison of observed and expected frequencies.
- Histograms and density curves analysis.
- Probability plots.
- Descriptive statistics comparison.

- Goodness-of-fit tests application.
- Residuals analysis.

It is crucial to remember that a model shouldn't be selected merely because it is well-liked. The model's presumptions must align with the data and research methodology.

The Poisson distribution, for instance, might be applied when a certain number of independent events happen at a steady pace. However, the Poisson model is likely to fit the data poorly if events cluster together, the variance is much higher than the mean, or there is a considerable shift in the rate of occurrence of events.

Moreover, choosing a normal distribution for a highly skewed variable just because of its popularity can result in misinterpretation of probabilities.

12. Example of Interpretation

Suppose Python analysis of drug concentration data produces the following results:

Table 4: Summary of Statistical Characteristics of Drug Concentration Data

Statistical characteristic	Result
Mean concentration	50 mg/L
Median concentration	49.8 mg/L
Standard deviation	4 mg/L
General distribution pattern	Approximately bell-shaped
Main concentration range	42–58 mg/L
Extreme observations	A few values above 60 mg/L

From the findings, it can be deduced that there is an approximate normal distribution for the drug concentration at 50 mg/L. The mean concentration of 50 mg/L and the median concentration of 49.8 mg/L are almost identical, giving the indication of approximate symmetry. Standard deviation of 4 mg/L is used to identify how far off from the mean value each point lies, whereas the concentration lying between 42 and 58 mg/L shows how close to the mean value they lie.

The range of 42 to 58 mg/L is almost equal to:

$$50 \pm 8$$

Because 8/4=2, this means that the observed value is about two standard deviations away from the mean. Hence, it follows that the observed values are generally consistent with those of a normally distributed set of data.

Observations that occur above 60 mg/L do not automatically make one assume that these values are erroneous. Researchers will have to review patient records, information on dosages, information on sample collection and handling, and other related information. In the event that they are valid, they may be useful for scientific purposes.

Generally, what we can observe from the above is that there is some symmetry in the observed values, some amount of variability, and the existence of some high values. The observed values seem to be in line with the assumption of normality, but further consideration must still be done.

REFERENCES

1. Addie, R. D., Balluff, B., Bovée, J. V. M. G., Morreau, H., & McDonnell, L. A. (2015). Current state and future challenges of mass spectrometry imaging for clinical research. *Analytical Chemistry, 87*(13), 6426–6433.
2. Alamo, T., Reina, D. G., & Millán, P. (2020). Data-driven methods to monitor, model, forecast and control COVID-19 pandemic: Leveraging data science, epidemiology and control theory. *arXiv Preprint arXiv:2006.01731*.
3. Alizadeh, M., Yousefi, L., Pakdel, F., Ghotaslou, R., Rezaee, M. A., Khodadadi, E., ... & Kafil, H. S. (2021). MALDI-TOF mass spectroscopy applications in clinical microbiology. *Advances in Pharmacological and Pharmaceutical Sciences*, *2021*(1), 9928238.
4. Baxter, S. L., Saseendrakumar, B. R., Paul, P., Kim, J., Bonomi, L., Kuo, T. T., ... & All of Us Research Program Investigators. (2021). Predictive analytics for glaucoma using data from the all of us research program. *American journal of ophthalmology*, *227*, 74-86.
5. Behl, T., Kaur, I., Sehgal, A., Singh, S., Bhatia, S., Al-Harrasi, A., Chigurupati, S., Alrashdi, I., Bungau, S., & Uddin, M. S. (2021). Bioinformatics accelerates the major tetrad: A real boost for the pharmaceutical industry. *International Journal of Molecular Sciences, 22*(12), 6184.
6. Bhatt, P., Singh, S., Kumar, V., Nagarajan, K., Mishra, S. K., Dixit, P. K., ... & Kumar, S. (2024). Artificial intelligence in pharmaceutical industry: Revolutionizing drug development and delivery. *Current Artificial Intelligence*, *2*(1), E051223224198.
7. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., ... & Caporaso, J. G. (2018). *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science* (No. e27295v1). PeerJ Preprints.
8. Cheng, F., Wang, F., Tang, J., Zhou, Y., Fu, Z., Zhang, P., ... & Cummings, J. (2024). Artificial intelligence and open science in discovery of disease-modifying medicines for Alzheimer's disease. *Cell Reports Medicine*, *5*(2).
9. Chiolero, A., & Buckeridge, D. (2020). Glossary for public health surveillance in the age of data science. *Journal of epidemiology and community health*, *74*(7), 612-616.
10. Crow, E. L., & Shimizu, K. (Eds.). (1988). *Lognormal distributions: Theory and applications*. Marcel Dekker. Note: Google Scholar displayed a 2018/republished record in your supplied results, but the original book is from 1988.

11. Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., ... & Toga, A. W. (2016). Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PloS one*, *11*(8), e0157077.

12. Donoho, D. (2024). Data science at the singularity. *Harvard Data Science Review*, *6*(1).

13. Dwivedi, A. K. (2022). How to write statistical analysis section in medical research. *Journal of Investigative medicine*, *70*(8), 1759-1770.

14. Eckerle, M., Ambroggio, L., Puskarich, M. A., Winston, B., Jones, A. E., Standiford, T. J., & Stringer, K. A. (2017). Metabolomics as a driver in advancing precision medicine in sepsis. *Pharmacotherapy, 37*(9), 1023–1032.

15. Galkina Cleary, E., Jackson, M. J., Zhou, E. W., & Ledley, F. D. (2023). Comparison of research spending on new drug approvals by the National Institutes of Health vs the pharmaceutical industry, 2010–2019. *JAMA Health Forum, 4*(4), e230511.

16. Geskus, R. B. (2015). *Data analysis with competing risks and intermediate states*. CRC press.

17. Ghosh, T. K., & Jasti, B. R. (Eds.). (2021). *Theory and practice of contemporary pharmaceutics*. CRC Press.

18. Gibson, E. W. (2019). Leadership in statistics: Increasing our value and visibility. *The American Statistician, 73*(2), 109–116.

19. Gibson, E., Koblbauer, I., Begum, N., Dranitsaris, G., Liew, D., McEwan, P., Ortiz, A., & Stull, D. (2017). Modelling the survival outcomes of immuno-oncology drugs in economic evaluations: A systematic approach to data analysis and extrapolation. *PharmacoEconomics, 35*(12), 1257–1270.

20. Hajnajafi, K., & Iqbal, M. A. (2025). Mass-spectrometry based metabolomics: An overview of workflows, strategies, data analysis and applications. *Proteome Science, 23*, Article 5.

21. Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial intelligence for clinical trial design. *Trends in pharmacological sciences*, *40*(8), 577-591.

22. Hazra, A., & Gogtay, N. (2017). Biostatistics series module 10: brief overview of multivariate methods. *Indian Journal of Dermatology*, *62*(4), 358.

23. Hazra, A., & Gogtay, N. (2017). Biostatistics series module 9: Survival analysis. *Indian Journal of Dermatology, 62*(3), 251–257.

24. He, K. Y., Ge, D., & He, M. M. (2017). Big data analytics for genomic medicine. *International journal of molecular sciences*, *18*(2), 412.

25. Ho, M., van der Laan, M., Lee, H., Chen, J., Lee, K., Fang, Y., ... & White, R. (2023). The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. *Statistics in Biopharmaceutical Research*, *15*(1), 43-56.

26. Hsieh, C.-Y., Su, C.-C., Shao, S.-C., Sung, S.-F., Lin, S.-J., Kao Yang, Y.-H., & Lai, E. C.-C. (2019). Taiwan's National Health Insurance Research Database: Past and future. *Clinical Epidemiology, 11*, 349–358.

27. Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (Eds.). (2016). *Handbook of survival analysis*. CRC Press.

28. Lawrence, K. D. (2019). *Robust regression: analysis and applications*. Routledge.

29. Manrai, A. K., Cui, Y., Bushel, P. R., Hall, M., Karakitsios, S., Mattingly, C. J., Ritchie, M., Schmitt, C., Sarigiannis, D. A., Thomas, D. C., Wishart, D., Balshaw, D. M., & Patel, C. J. (2017). Informatics and data analytics to support exposome-based discovery for public health. *Annual Review of Public Health, 38*, 279–294.

30. Mark, P. B., Stafford, L. K., Grams, M. E., Aalruz, H., Abd ElHafeez, S., Abdelgalil, A. A., ... & Do, T. H. P. (2025). Global, regional, and national burden of chronic kidney disease in adults, 1990–2023, and its attributable risk factors: a systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*, *406*(10518), 2461-2482.

31. Marques, L., Costa, B., Pereira, M., Silva, A., Santos, J., Saldanha, L., ... & Vale, N. (2024). Advancing precision medicine: a review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare. *Pharmaceutics*, *16*(3), 332.

32. Noguchi, Y., Tachi, T., & Teramachi, H. (2020). Comparison of signal detection algorithms based on frequency statistical model for drug-drug interaction using spontaneous reporting systems. *Pharmaceutical Research, 37*(5), Article 86.

33. Peixoto, L., Risso, D., Poplawski, S. G., Wimmer, M. E., Speed, T. P., Wood, M. A., & Abel, T. (2015). How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic acids research*, *43*(16), 7664-7674.

34. Playdon, M. C., Joshi, A. D., Tabung, F. K., Cheng, S., Henglin, M., Kim, A., Lin, T., van Roekel, E. H., Huang, J., Krumsiek, J., Wang, Y., Mathé, E., Temprosa, M., Moore, S., Chawes, B., Eliassen, A. H., Gsur, A., Gunter, M. J., Harada, S., … Zeleznik, O. A. (2019). Metabolomics analytics workflow for epidemiological research: Perspectives from the Consortium of Metabolomics Studies (COMETS). *Metabolites, 9*(7), Article 145.

35. Russ, T. C., Woelbert, E., Davis, K. A. S., Hafferty, J. D., Ibrahim, Z., Inkster, B., John, A., Lee, W., Maxwell, M., McIntosh, A. M., Stewart, R., & MQ Data Science Group. (2019). How data science can advance mental health research. *Nature Human Behaviour, 3*(1), 24–32.

36. Sadatsafavi, M., Adibi, A., Puhan, M., Gershon, A., Aaron, S. D., & Sin, D. D. (2021). Moving beyond AUC: decision curve analysis for quantifying net benefit of risk prediction models. *European Respiratory Journal*.

37. Sverdlov, O. (Ed.). (2015). *Modern adaptive randomized clinical trials: statistical and practical aspects*. CRC Press.

38. van Ommen, G.-J. B., Törnwall, O., Bréchot, C., Dagher, G., Galli, J., Hveem, K., Landegren, U., Luchinat, C., Metspalu, A., Nilsson, C., Solesvik, O. V., Perola, M., Litton, J.-E., & Zatloukal, K. (2015). BBMRI-ERIC as a resource for pharmaceutical and life science industries: The development of biobank-based Expert Centres. *European Journal of Human Genetics, 23*(7), 893–900.

39. Vu, T., Ma, P., Chen, J. S., de Hoon, J., Van Hecken, A., Yan, L., ... & Vargas, G. (2017). Pharmacokinetic-Pharmacodynamic Relationship of Erenumab (AMG 334) and Capsaicin-Induced Dermal Blood Flow in Healthy and Migraine Subjects: Vu et al. *pharmaceutical research*, *34*(9), 1784-1795.

40. Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports, 10*, Article 15534.

Unit 3...

SAMPLING AND STATISTICAL INFERENCE

This unit covers the basics of sampling and statistics in clinical, pharmaceutical, and health care studies. Topics include the definition of population and sample, various sampling methods, sampling error, and sampling bias that may distort the results of research. This unit also covers the concept of Central Limit Theorem and confidence intervals as ways to estimate the characteristics of a population. In addition, hypothesis testing and its various components such as null and alternative hypotheses, Type I and Type II errors, p-values, and statistical significance are covered. Lastly, this unit covers the use of Python for performing calculations of confidence intervals, hypothesis testing, and interpreting results of inferential statistics using pharmaceutical data.

3.1. POPULATION, SAMPLE AND SAMPLING TECHNIQUES

Researchers who conduct studies in clinical settings and in the field of pharmaceuticals normally wish to conduct their research on a broad scale and on the entire population, but due to the unavailability of information for every individual, they tend to select a smaller number of people out of that larger population for their study. The entire population which is the whole collection of individuals about which a researcher wishes to generalize is called a population, whereas the selected small group of that population is referred to as a sample. This process of selection of samples from the population is termed as sampling.

3.1.1. Population and Sample in Clinical Research

Some basic concepts of population and sample in clinical research. Population is defined as the whole group of people involved in a research study, for instance all patients who suffer from certain disease while sample is a subgroup of population, which will be really studied. Target population represents the group to which the results of research should be generalized while accessible population represents a subgroup of population available for research. Sample should have some qualities including representativeness, adequate size, proper selection and relevance to the research question. Researchers utilize sample statistics like sample mean or proportion to estimate population parameters including population mean or proportion. It means that an appropriate sample provides an opportunity to acquire reliable information concerning a large population.

❖ **Population**

Population refers to the entire group of people, patients, participants, or subjects having common characteristics and involved in a research study. The population of a clinical study could be the patients who are affected by a specific disease, subjects on a given medication, normal participants, or subjects with specific demographic characteristics.

Example: Let us assume that a researcher is interested in finding out the efficacy of a newly developed drug on diabetics.

The population in this case may consist of:

All eligible adult patients suffering from Type 2 diabetes. The population may be large, making it impossible to conduct an analysis of all the subjects.

Types of Population

1. Target Population: Target population refers to the entire population to which the results of the study can be generalized by the researcher.

Example: In case the research is conducted to test a new drug on patients suffering from high blood pressure, the target population could be:

All patients of high blood pressure who fulfill the required criteria.

2. Accessible Population: The accessible population refers to that part of the target population which is really available to the researcher.

Example: A researcher might wish to research upon all the hypertensive patients in India, but the research might be carried out in certain hospitals located in Bhopal alone. The eligible patients available at these hospitals would be the accessible population.

So,

Target population → All eligible patients

Accessible population → Eligible patients available to the researcher

❖ **Sample**

A sample can be defined as a small number of people who have been chosen from the population for conducting research.

Researchers conduct research on the sample and then draw conclusions about the entire population from the findings of the research.

For instance, consider that there are 10,000 eligible patients of hypertension, but out of them, only 500 are selected by the researcher for conducting a study.

- Population = 10,000 eligible patients

- Sample = 500 selected patients

Characteristics of a Good Sample

The ideal sample will possess the following features:

1. Representative – It will contain features typical for the population.
2. Sufficiently large – There should be enough subjects in the sample to make the findings credible.
3. Properly chosen – An appropriate sampling technique should be applied.
4. Bias-free – The selection process should not be biased towards any specific subjects.
5. Related to the research problem – The subjects should meet all the eligibility requirements of the research.

Population Parameter

A population parameter is a numerical value that describes a characteristic of the entire population.

Examples include:

- Population mean
- Population proportion
- Population variance
- Population standard deviation

For instance, the true mean of blood pressure in all eligible participants within a population is a population parameter.

A population parameter is mostly unknown because researchers hardly examine all members.

Sample Statistic

A sample statistic is a numerical value calculated using data collected from a sample.

Examples include:

- Sample mean
- Sample proportion
- Sample variance
- Sample standard deviation

Example: If the mean systolic blood pressure of 500 patients is 132 mmHg, then 132 mmHg becomes a sample statistic.

Sample statistics are used to infer population parameters.

Table 1: Difference Between Population and Sample

Population	Sample
Entire group of interest	Smaller part of the population
Usually, large	Usually, smaller
May be difficult to study completely	Easier to study
Contains the true population characteristics	Used to estimate population characteristics
Example: all eligible diabetic patients	Example: 300 selected diabetic patients

3.1.2. Sampling Methods, Sampling Error and Bias

❖ **Sampling**

Sampling refers to the act of choosing people who will participate in the research study from the entire population. Therefore, if there are 5,000 patients available to participate in a study but only 300 are required, then the researcher uses some sampling techniques to choose 300 of those people. The different types of sampling techniques are:

1. **Probability sampling**
2. **Non-probability sampling**

A. Probability Sampling

In probability sampling, the process followed in selecting individuals is random or probability based, where each individual has an equal probability of being selected. Probability sampling can usually be applied in cases where the researcher needs a representative sample and generalizes findings about a larger population. A sampling technique is a method through which a researcher selects subjects from the population for use in a research study. For clinical research studies, the use of correct sampling allows the researcher to select the subject systematically and gather information that gives him or her more knowledge about the population. The main types of probability sampling are simple random sampling, systematic sampling, stratified random sampling, cluster sampling, and multistage sampling.

1. Simple Random Sampling

Simple random sampling refers to a form of sampling where each individual of the population under consideration has an equal probability of being selected as part of the sample. This process is achieved through the use of a random mechanism, such as random number generation through computer

programs, lottery, and random number tables. This is because the researcher does not choose the individuals to be selected in the sampling process.

Example:

In a hospital, there is a list of 1,000 patients that qualifies for an experiment, and the researcher requires 100 patients for the research. In such a case, numbers can be assigned to each of the patients, and the researcher will randomly pick 100 patients using random number selection using the computer.

2. Systematic Sampling

Systematic sampling refers to a technique whereby sampling units from a sequential population are chosen using a regular or fixed interval. In systematic sampling, the researcher begins by establishing the interval size for sampling and then goes ahead to sample using the same interval.

The interval is arrived at using the following formula:

$$k = \frac{N}{n}$$

Where:

- N = Population size
- n = Required sample size
- k = Sampling interval

Example: Suppose a researcher has a list of 1,000 patients and needs to select 100 participants.

$$k = \frac{1000}{100} = 10$$

Hence, the sampling interval will be 10. Once the starting point has been chosen, the researcher can then pick out every 10th individual from the population list.

For instance, when the starting point is patient number 4, the selected individuals will be:

4, 14, 24, 34, 44, 54, …

This process continues until the required sample size is obtained.

3. Stratified Random Sampling

Stratified random sampling is a technique wherein the target population is segmented into smaller and homogeneous groups known as strata. Stratification is done in accordance with criteria significant for the research process. Once the population is segmented into strata, samples are drawn randomly from each stratum.

Strata can be formed using criteria like:

- Age
- Sex
- Disease severity
- Geographic location
- Treatment group
- Educational level

The participants can be chosen from the strata either proportionally based on their size within the population or as per a predetermined sampling strategy.

Example: An investigator wishes to examine the efficacy of a drug among individuals from various age groups.

These individuals can be categorized into:

- Young adults
- Middle-aged adults
- Older adults

Random sampling is conducted from the various age categories. This makes sure that all age categories are represented in the sample.

4. Cluster Sampling

In cluster sampling, the population is first divided into groups or clusters. The researcher then takes samples not of individuals from the population but of the groups themselves. Clusters may include naturally occurring groups, such as:

Clusters may include naturally occurring groups such as:

- Hospitals
- Clinics
- Villages
- Schools
- Districts
- Communities

Cluster sampling is particularly useful when the population is large and geographically distributed.

Example: The scientist wants to analyze the problem of drug compliance in patients residing in one of the large states. Instead of creating the list of all the potential respondents in this state, the scientist

selects hospitals as clusters. Some hospitals are randomly chosen, and the respondents from these hospitals join the research sample.

Hence, the selection procedure may be schematically presented as:

State → Hospitals → Eligible Patients

5. Multistage Sampling

Multistage sampling refers to a sampling technique whereby the sample participants are chosen in two or more stages. Instead of picking the participants directly from the whole population, the researcher selects smaller groups successively until he arrives at the needed participants.

For instance, the sampling procedure may include:

State → District → Hospital → Patient

For the first step, the selection of some states can take place. In the second step, some districts can be selected from the chosen states. In the third step, hospitals can be selected from the selected districts. Lastly, the eligible patients can be selected from the chosen hospitals.

Example: Suppose researchers want to conduct a large epidemiological study among patients with diabetes across several regions. They may:

1. Select specific states.
2. Select specific districts within those states.
3. Select hospitals or healthcare centers within those districts.
4. Select eligible patients from those hospitals.

Thus, multistage sampling can be represented as:

State → District → Hospital → Patient

This method allows researchers to conduct studies involving large and geographically dispersed populations in a structured manner.

B. Non-Probability Sampling

Non-probability sampling refers to the sampling process in which individuals are chosen by a sampling technique that does not involve random selection based on probability. In non-probability sampling, there is no guarantee that each individual in the population will have an equal opportunity or chance of being included in the sample. Individuals may be chosen because they are easily accessible, possess certain traits, or can aid the researcher in obtaining access to other qualified individuals.

1. Convenience Sampling

Convenience sampling is a non-probability sampling technique wherein individuals are chosen because of their easy accessibility and availability to the researcher. Individuals who can be contacted conveniently and satisfy the inclusion criteria of the researcher are recruited by the researcher.

Example: A researcher intends to conduct research on patient satisfaction regarding the pharmaceutical services in the hospital. Eligible patients who are conveniently available in the OPD are recruited by the researcher. These individuals constitute the sample.

2. Purposive Sampling

Purposive sampling is where the researcher selects participants who have particular qualities or traits. This is done through the researcher knowing what kind of participants he or she requires and then going out of his or her way to select people who match these requirements.

Example: A researcher needs to conduct research on the experience of clinical pharmacists operating in an intensive care unit setting. He/she selects pharmacists who have worked in the ICU environment before. In this case, the choice of the participants depends on the purpose of the research.

3. Quota Sampling

In quota sampling, the population is divided into several groups and the number of participants that need to be selected from each group is pre-decided. The participants are recruited till the required number or quota of participants from each group is achieved.

Example: A researcher aims to study the awareness of diabetes among 200 adults and he plans to include:

- 100 male participants
- 100 female participants

Recruitment of suitable respondents continues until the attainment of both quotas is achieved.

It can be noted that quota sampling ensures that the desired number of respondents is recruited into the study.

4. Snowball Sampling

Snowball sampling is a type of non-probability sampling approach that is applied in situations where the needed respondents cannot be easily found or accessed. Here, the researcher starts by identifying few potential respondents, who then assist in finding others who qualify to participate in the study. This process continues in that way.

Example: The researcher wishes to conduct a study on the experience of patients in a healthcare facility, but it is not easy to find respondents. By recruiting one qualified respondent, the respondent introduces the researcher to other potential respondents who qualify for the study.

This process can be shown as:

Initial participant → Other eligible participants → Additional participants → Larger sample

The snowball sampling technique is very effective in situations where the population under consideration cannot be accessed from an existing list but rather through networks of individuals.

❖ **Sampling Error**

Sampling error is the discrepancy between the value obtained from sampling and the true value. This arises since researchers conduct studies using a selected part of the population, and not all of the population.

Example: Suppose the true average systolic blood pressure in a population is:

130 mmHg

A randomly selected sample has an average of:

133 mmHg

The difference:

$$133 - 130 = 3 \text{ mmHg}$$

is an example of sampling error.

A sampling error does not imply that there has been any error made by the researcher in conducting the research.

Factors Affecting Sampling Error

The following factors influence the sampling error:

1. Sample Size: The larger the sample size, the greater the stability in estimating and the less the sampling error.
2. Variability in Population: When there is a lot of variation between individuals in a population, sampling variability could be higher.
3. Sampling Technique: There should be a proper sampling technique used in order to get the sample.
4. Representative Sample: The sample which does not represent the population will yield biased results.

❖ **Sampling Bias**

Sampling bias refers to the situation in which the sampling method employed tends to favor or exclude certain features of the population. Sampling bias differs from general sampling variability in that the former involves a systematic issue with the sampling process.

Example: Assume that a scientist wishes to investigate physical activity amongst adult people but samples people from a gym. The fact of the matter is that those individuals who visit a gym will most likely be more physically active than other people.

Types of Sampling Bias

Sampling bias occurs whenever the process of selecting participants leads to some people or groups being over-represented or underrepresented in the sample relative to the population. Sampling bias is characterized by different types of biases that occur at different stages of recruiting participants for the study.

1. **Selection Bias:** Selection bias occurs whenever the process of selecting participants in the study results in the over-representation or under-representation of certain individuals/groups from the target population. In effect, the sample that is selected differs from the population from which the researcher intends to make generalizations. The selection process could influence the validity of the results and make the results ungeneralizable to the target population.
2. **Volunteer Bias:** Volunteer bias is where volunteers in a research study are systematically different from non-volunteers in various ways such as having certain characteristics, interests, attitudes, health-related behaviours and experience. Volunteering for the study might mean that people who decide to be part of the study have something to gain from participating in the study.
3. **Non-Response Bias:** Non-Response bias occurs when people selected to be in the study are not willing to participate in the study or to provide the necessary information required to answer the research questions. In this case, characteristics of those selected but who fail to participate in the study may differ from those of participants.
4. **Under coverage Bias:** Under coverage bias occurs when certain members or groups within the target population have little or no chance to be part of the sampling procedure. In this case, the under representation could be as a result of the poor sampling frame used by the researcher.
5. **Convenience Bias:** Convenience bias occurs when researchers rely more on convenience in terms of access and availability in selecting participants for their study. People selected through convenience could have characteristics that differ from other less available people in the population.

Table 2: Difference Between Sampling Error and Sampling Bias

Sampling Error	Sampling Bias
Occurs because only a sample is studied	Occurs because of systematic problems in participant selection
Generally, represents random sampling variation	Represents systematic distortion
Can occur even with proper random sampling	Often results from inappropriate sampling or recruitment
Usually decreases with increasing sample size	Increasing sample size does not necessarily remove it
Example: sample mean differs from population mean	Example: selecting participants only from one convenient hospital

3.2. CENTRAL LIMIT THEOREM AND CONFIDENCE INTERVALS

The CLT (Central Limit Theorem) and confidence intervals have been used in statistics and clinical research to draw inferences about populations of interest. First, the CLT is a theorem in statistics that demonstrates the tendency of the distribution of sample means to approximate a normal distribution as long as large enough samples are repeatedly drawn from the same population. It helps to conclude anything statistically about a population based on samples drawn from the population. Secondly, confidence intervals are a set of possible values of an unknown population parameter like a population mean or proportion.

3.2.1. Conceptual Understanding of the Central Limit Theorem

The Central Limit Theorem (CLT) is one of the basic concepts of statistics. This concept outlines the way sample means act in case many samples are drawn from a population. The Central Limit Theorem states that if many random samples are drawn from a population, then the sampling distribution of sample means approaches the shape of a normal distribution, regardless of the population distribution shape. In other words, according to the Central Limit Theorem, the averages of many random samples approach the shape of a bell curve.

❖ **Sampling Distribution of the Mean**

The sampling distribution of a statistic is defined as the probability distribution of that statistic, which is calculated from the data of a sample randomly drawn from a population. In this regard, let us

consider an example where researchers draw different samples consisting of 50 patients each from a large population, and they compute the mean blood pressure of each sample. This will constitute the sampling distribution of the sample mean, and with a large enough sample size, it becomes normally distributed.

Important Features of the Central Limit Theorem

Some of the key properties of the Central Limit Theorem include the following:

1. **Sample Statistics:** The CLT is basically focused on the statistics drawn from the various samples, especially the sample mean.
2. **Bigger Sample Sizes give Better Approximation:** With an increase in the sample size, the sampling distribution of the sample mean tends to approximate towards the normal distribution.
3. **Not Necessarily Normal Population Distribution:** One of the key properties of the CLT is that the population need not be normally distributed. With the help of the large samples of random nature, we can get a normally distributed sample mean distribution.
4. **Mean of the Sampling Distribution of Sample Mean:** The mean of the sampling distribution of the sample mean is equal to the population mean:

$$\mu_{\bar{x}} = \mu$$

Where:

- $\mu_{\bar{x}}$= mean of the sampling distribution of the sample mean
- μ= population mean

5. **Standard error of the sample mean:** The standard deviation of the sampling distribution of the sample mean is called the standard error of the mean.

It is commonly expressed as:

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

Where:

- SE= standard error
- σ= population standard deviation
- n= sample size

When the population standard deviation is unknown, the sample standard deviation $s$may be used:

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

As the sample size increases, the standard error generally decreases.

Central Limit Theorem in Clinical Research

The Central Limit Theorem becomes highly relevant in clinical and pharmaceutical research, given that researchers generally sample and do not study the population as a whole.

This can include measures such as:

- Blood pressure
- Blood glucose level
- Body weight
- Drug concentration
- Cholesterol level
- Heart rate

The sample means can be used by the researchers to make inferences about population means. The Central Limit Theorem provides a theoretical foundation for the numerous statistical tests based on the normally distributed sample mean.

Simple Illustration of the Central Limit Theorem

For instance, consider a population consisting of patients whose blood glucose varies significantly. This is not necessarily normally distributed. Scientists randomly pick out patients and get their blood glucose levels and then compute the mean of the selected sample.

When the sample size is large enough:

Individual observations → Sample means → Approximately normal sampling distribution

Thus, even when the original observations are not normally distributed, the distribution of sample means can become approximately normal.

3.2.2. Confidence Intervals and Their Interpretation

Confidence Interval (CI) can be defined as an interval of numbers that is calculated using sample data and can be used for estimation of an unknown population parameter. A confidence interval does not offer just one number but offers a range that is indicative of the uncertainity in the value.

In a clinical trial, the average amount of reduction in blood pressure can be:

8 mmHg

with a 95% confidence interval of 6 to 10 mmHg.

The interval provides information about the precision of the estimated treatment effect.

Point Estimate

Point estimates are the individual values that can be derived from samples in order to estimate a population parameter.

They include the following examples:

- Sample mean used as an estimate of the population mean.
- Sample proportion used as an estimate of the population proportion.
- Difference in sample means used as an estimate of the difference in population means.

For instance, if the average decrease in blood pressure of the subjects in the study is 8mmHg, then 8mmHg is the point estimate.

A confidence interval is a range based on the point estimate.

Confidence Level

The confidence level indicates the degree of confidence associated with the method used to construct the interval.

Common confidence levels include:

- 90%
- 95%
- 99%

Use of the 95% confidence interval is quite common in clinical and drug-related research.

As the confidence level increases, the width of the interval becomes larger, keeping all other factors constant.

General Form of a Confidence Interval

A confidence interval can generally be expressed as:

$$\text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin of Error}$$

For example:

$$\text{CI} = \bar{x} \pm \text{Margin of Error}$$

The margin of error represents the amount of uncertainty around the point estimate.

Confidence Interval for a Population Mean

If all the required conditions are fulfilled and population standard deviation is known, then the confidence interval of a population mean can be stated as:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Where:

- $\bar{x}$= sample mean
- $z_{\alpha/2}$= appropriate critical value
- σ= population standard deviation
- n= sample size

When the population standard deviation is unknown, a t-distribution is commonly used:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Where:

- s= sample standard deviation
- $t_{\alpha/2}$= appropriate critical value from the t-distribution

Example of a Confidence Interval

Suppose a clinical study estimates that a new medicine reduces systolic blood pressure by an average of:

10 mmHg

The researchers calculate a 95% confidence interval of:

7 mmHg to 13 mmHg

The result can be reported as:

Estimated mean reduction = 10 mmHg (95% CI: 7–13 mmHg).

The interval gives a range of plausible values for the population mean reduction based on the statistical method used.

Interpretation of a 95% Confidence Interval

The proper definition for 95% confidence interval is one that is derived from the performance of the procedure in the long run, such that if the sampling was done repeatedly in the right way and the confidence interval of 95% was computed, then almost all of them will include the true value of the parameter of the population. For an individual computed confidence interval, the definition is one in which researchers say the confidence interval gives a plausible range of values of the population parameter.

1. **Width of Confidence Interval:** The width of confidence interval tells us about the accuracy of an estimate.
2. **Narrow Confidence Interval:** Narrow Confidence Interval usually means greater accuracy. For instance:

9.5 to 10.5

3. **Wide Confidence Interval:** A wide interval indicates greater uncertainty and lower precision. For example:

6 to 14

Thus, scientists usually favor estimates that are fairly precise with relatively narrower confidence intervals.

Factors Affecting Confidence Interval Width

There are certain variables that influence the confidence interval.

1. **Sample Size:** The larger the sample size, the smaller the standard error will be, which leads to a narrower confidence interval.
2. **Variability:** Higher variability in the sample leads to higher standard error and hence, a wider confidence interval.
3. **Confidence Level:** The confidence level is increased from 95% to 99%, then it results in a wider confidence interval.

Thus:

Higher confidence level → Wider interval

Larger sample size → Narrower interval

Confidence Intervals in Clinical Research

Confidence intervals are used extensively in clinical and pharmaceutical studies to describe uncertainty in such estimates as:

- Mean blood pressure
- Mean drug concentration
- Difference between treatment means
- Treatment response rate
- Risk ratio
- Odds ratio
- Relative risk
- Incidence rate
- Proportion of patients experiencing an adverse effect

For example, researchers may report:

Risk ratio = 0.72; 95% CI = 0.55–0.94

The confidence interval gives an indication of what values are consistent with the data and model.

3.3. HYPOTHESIS TESTING AND STATISTICAL SIGNIFICANCE

Hypothesis testing is one of the methods that can be used to determine whether or not there is enough evidence contained in a sample to support a specific claim about the population in question. In clinical and pharmaceutical research, hypothesis testing is used by the researcher to determine issues such as whether or not a new drug is better compared to an already existing medication, whether or not there has been any change in the blood pressure of someone undergoing some sort of treatment, or whether or not there is any difference between two groups in their treatment results.

3.3.1. Null and Alternative Hypotheses

Hypotheses on statistics can be defined as assumptions or statements made about population parameters or relationships between variables. In hypothesis testing, there are usually two conflicting statements; these include the null hypothesis and the alternative hypothesis.

Null Hypothesis

The null hypothesis, denoted as H_0, is a hypothesis stating that there is no effect, no difference, or no relationship in the population that is being studied. The null hypothesis is the one that researchers assume to be true at the outset before any statistical testing.

For instance, if a researcher aims to study whether a certain drug lowers blood pressure,

The null hypothesis could be:

$$H_0: \mu_1 = \mu_2$$

It also means that the mean reduction in blood pressure by the new drug is equal to that of the comparator drug.

Another example is:

$$H_0: \mu = 120$$

This implies that the mean blood pressure of the population is 120 mm Hg.

The null hypothesis acts as a benchmark for evaluating the sample results.

Alternative Hypothesis

H_1or H_a, known as the alternative hypothesis, states that there is an effect, difference, or relationship in the population. This states that the effect or difference that the researcher wishes to demonstrate may exist.

Example: In order for a researcher to establish if a new drug results in a different level of blood pressure reduction from an existing drug:

$$H_a: \mu_1 \neq \mu_2$$

This means that there is a difference between the mean blood pressure decreases of the two treatments.

In research to determine if a new treatment reduces blood pressure more than an existing treatment:

$$H_a: \mu_1 > \mu_2$$

The direction of the alternative hypothesis depends on the research question.

Types of Alternative Hypotheses

1. Two-Tailed Alternative Hypothesis: The two tailed test is done when the researcher wants to find a difference in any direction.

$$H_a: \mu_1 \neq \mu_2$$

The hypothesis tests whether one treatment can result in either a higher or lower effect compared to the other.

2. One-Tailed Alternative Hypothesis: One-tailed hypothesis is applied if the research problem shows a specific direction.

For example:

$$H_a: \mu_1 > \mu_2$$

or

$$H_a: \mu_1 < \mu_2$$

The decision to perform a one-tailed test is to be made depending on the research hypothesis and methodology employed rather than the convenience of a statistically better outcome.

Example of Null and Alternative Hypotheses

If a scientist wished to test whether a newly invented medication lowers the average systolic blood pressure more effectively than an existing treatment,

Null Hypothesis

$$H_0: \mu_{\text{new}} = \mu_{\text{standard}}$$

There is no significant difference between the means of systolic blood pressure reductions from the two treatments.

Alternative Hypothesis

$$H_a: \mu_{\text{new}} \neq \mu_{\text{standard}}$$

There exists a difference in the mean reduction of systolic blood pressure in the two groups under the study.

Then the researcher collects the data and does the relevant test on statistics to establish whether there is any enough evidence to reject the null hypothesis.

Rejecting and Failing to Reject the Null Hypothesis

Once the results of a hypothesis test have been analyzed, one of the following decisions is usually made:

Reject H_0: When there is enough statistical evidence to refute the null hypothesis, the null hypothesis is rejected. This means that there is enough statistical evidence against H_0.

Fail to Reject H_0: When there is not enough statistical evidence to refute the null hypothesis, the null hypothesis is not rejected. It doesn't mean that the null hypothesis is true; it just means that there is not enough evidence to refute it.

3.3.2. Type I and Type II Errors, p-Values and Statistical Significance

Statistical tests of hypotheses have uncertainty due to the fact that conclusions are drawn from samples and not the whole population. As such, a researcher may draw wrong conclusions. There are two main forms of errors; Type I error and Type II error.

Type I Error

Type I Error arises when the null hypothesis is rejected even though it is true. The likelihood of committing a Type I error is indicated by:

$$\alpha$$

where $\alpha$is called the significance level.

A commonly used significance level is:

$$\alpha = 0.05$$

This corresponds to a 5% significance level.

Example: Now imagine a new medication which does not really have any difference from the old one. But if the test shows that there is indeed a difference, then the type I error has been made.

Type II Error

A Type II error takes place in the situation where a false null hypothesis is not rejected. The Type II error probability is denoted as:

$$\beta$$

Example: Assuming that the new drug is better than the existing one, but there is no statistical evidence of it in the experiment, it is a Type II error.

Statistical Power

Statistical power is the probability that a statistical test will correctly reject a false null hypothesis.

It is expressed as:

$$\text{Power} = 1 - \beta$$

For example, if:

$$\beta = 0.20$$

then:

$$\text{Power} = 1 - 0.20 = 0.80$$

or 80% power.

Statistical power is influenced by factors such as sample size, variability, effect size, and the chosen significance level.

Table 3: Type I and Type II Errors

Decision	**H_0 is True**	**H_0 is False**
Reject H_0	Type I Error	Correct decision
Fail to reject H_0	Correct decision	Type II Error

Thus:

- **Type I error:** Rejecting a true H_0
- **Type II error:** Failing to reject a false H_0

p-Value

A p-value is a numerical value that is used in a hypothesis test to determine how well the obtained data fits the null hypothesis. In particular, the p-value is defined as the probability of obtaining a test statistic as or more extreme than the one that was obtained on the basis of the statistical test, given that the null hypothesis holds true.

Common Interpretation

For a significance level of:

$$\alpha = 0.05$$

researchers commonly use the following decision rule:

- If $p < 0.05 \rightarrow$ Reject H_0
- If $p \geq 0.05 \rightarrow$ Fail to reject H_0

The exact threshold should be specified before analyzing the data.

Example of a p-Value

Suppose a clinical trial compares a new medicine with a standard medicine and obtains:

$$p = 0.03$$

Using a significance level of:

$$\alpha = 0.05$$

we have:

$$0.03 < 0.05$$

Therefore, the result is considered statistically significant at the 5% level, and the researcher rejects the null hypothesis.

If the study instead produces:

$$p = 0.18$$

then:

$$0.18 > 0.05$$

The researcher would fail to reject the null hypothesis.

Statistical Significance

Statistical significance means that the observed outcome is unlikely to happen given the null hypothesis based on the selected statistical model and significance level.

When:

$$p < \alpha$$

the result is generally described as statistically significant.

When:

$$p \geq \alpha$$

the result is generally described as not statistically significant.

For example, if $\alpha = 0.05$:

Table 4: Interpretation of p-Values and Statistical Significance

p-value	Interpretation
$p < 0.05$	Statistically significant
$p \geq 0.05$	Not statistically significant

3.4. STATISTICAL INFERENCE USING PYTHON

Statistical inference refers to the process of making deductions and estimations from a sample about the population. Inferential statistics are widely used in pharmaceutical and clinical trials to estimate population parameters, construct confidence intervals, and conduct hypothesis testing on the effect of certain treatment or drug on patients. There are several Python libraries for statistical inference including NumPy, SciPy, Pandas, and stats models. This chapter focuses on the process of calculation of confidence interval and hypothesis testing in Python. Then we explain the process of interpretation of the results of inferential statistics using pharmaceutical examples.

3.4.1. Calculation of Confidence Intervals and Hypothesis Tests

Python may be applied for making statistical inferences from a sample. Two major uses of statistical inference with Python include calculation of confidence intervals and hypothesis testing. The first gives an estimated range of values of a population parameter while the second allows us to see if we have enough statistical proof for our claim.

Calculation of Confidence Intervals Using Python

A confidence interval gives a set of values within which an unknown population parameter is estimated based on sample data. For instance, scientists can compute a confidence interval to estimate the average change in blood pressure among patients under a certain medication.

Confidence intervals can be calculated using Python's SciPy package.

Example: Suppose the reduction in systolic blood pressure among 10 patients is:

```python
import numpy as np
from scipy import stats

data = np.array([8, 10, 7, 9, 11, 6, 10, 8, 9, 12])

mean = np.mean(data)
sem = stats.sem(data)

confidence_interval = stats.t.interval(
    0.95,
    df=len(data)-1,
    loc=mean,
    scale=sem
)

print("Mean:", mean)
print("95% Confidence Interval:", confidence_interval)
```

In this case:

- np.mean(data) gives us the sample mean.
- stats.sem(data) gives us the standard error of the mean.
- stats.t.interval() gives us the 95% confidence interval assuming a t-distribution.

This provides us with an interval that is a range of possible values for the mean reduction in blood pressure.

Hypothesis Tests Using Python

A hypothesis test is used to evaluate a claim about a population using sample data. The analysis generally involves:

1. Formulating the null hypothesis (H_0).
2. Formulating the alternative hypothesis (H_a).
3. Selecting a significance level, commonly $\alpha = 0.05$.
4. Performing an appropriate statistical test.
5. Calculating the test statistic and p-value.
6. Comparing the p-value with the significance level.
7. Drawing a statistical conclusion.

Python's SciPy library provides functions for many commonly used hypothesis tests.

Example: One-Sample t-Test

If the mean reduction in blood pressure was 8 mmHg, what the researcher wishes to test out is whether the mean reduction obtained is different from 8 mmHg.

```python
import numpy as np
from scipy import stats

data = np.array([8, 10, 7, 9, 11, 6, 10, 8, 9, 12])

t_statistic, p_value = stats.ttest_1samp(data, popmean=8)

print("t-statistic:", t_statistic)
print("p-value:", p_value)
```

The function ttest_1samp () performs a one-sample t-test.

The hypotheses can be written as:

$$H_0: \mu = 8$$

$$H_a: \mu \neq 8$$

The null hypothesis is rejected if the calculated p-value is less than 0.05 at a significance level of 5%, whereas the null hypothesis is accepted if the p-value is greater than or equal to 0.05.

Common Hypothesis Tests Available in Python

Python can perform many statistical tests used in clinical and pharmaceutical research.

Table 5: Common Hypothesis Tests and Their Applications in Python

Statistical Test	Common Application
One-sample t-test	Comparing a sample mean with a specified value
Independent t-test	Comparing means of two independent groups
Paired t-test	Comparing measurements taken from the same subjects
ANOVA	Comparing means across three or more groups
Chi-square test	Examining relationships between categorical variables
Mann–Whitney U test	Comparing two independent groups when a non-parametric approach is appropriate
Wilcoxon signed-rank test	Comparing paired observations using a non-parametric approach
Correlation tests	Assessing relationships between quantitative variables

It all depends on a variety of factors, such as the research question, research design, nature of the data, and others.

3.4.2. Interpretation of Inferential Statistical Results Using Pharmaceutical Data

Interpretation of inferential statistical findings entails the use of statistical measures such as estimates, confidence intervals, test statistics, and p-values in order to assess what the sample information indicates about the population. In pharmaceutical studies, such techniques can be utilized in the assessment of the efficacy of drugs, side effects, concentrations of drugs, patient reactions, and differences between groups that received treatments. The p-value must not be interpreted alone; other factors include confidence intervals, effect size, design of the study, and assumptions of the technique.

Interpretation of the Sample Mean

Sample mean is the mean that results from values seen in the study subjects. For instance, consider an average decrease in blood pressure in patients taking a new drug as follows:

$$\bar{x} = 10 \text{ mmHg}$$

This translates to mean reduction in blood pressure of 10 mmHg for the subjects in the sample. The mean of the sample can be used to estimate the population mean.

Interpretation of a Confidence Interval

Suppose a pharmaceutical study reports:

$$\text{Mean reduction} = 10 \text{ mmHg}$$

$$95\%\ CI = 8 \text{ to } 12 \text{ mmHg}$$

The confidence interval refers to the possible values that the population mean reduction can take based on the statistical process. If the confidence interval is narrower compared to a very wider one, then it implies precision.

Interpretation of a p-Value

Suppose a hypothesis test comparing a new drug with a standard treatment produces:

$$p = 0.03$$

If the predetermined significance level is:

$$\alpha = 0.05$$

then:

$$0.03 < 0.05$$

The test is thus significant at the 5% significance level, and the researcher fails to accept the null hypothesis. But the p-value does not show whether the treatment effect is large enough.

Example Using Pharmaceutical Data

Suppose researchers compare the reduction in blood pressure between two treatment groups.

New Drug: Mean reduction:

$$10 \text{ mmHg}$$

Standard Drug: Mean reduction:

$$7 \text{ mmHg}$$

The estimated difference is:

$$10 - 7 = 3 \text{ mmHg}$$

Suppose the analysis produces:

$$p = 0.02$$

and a 95% confidence interval for the difference of:

$$0.5 \text{ to } 5.5 \text{ mmHg}$$

The study would be that the study offers statistical proof that there exists a difference in the mean decrease in blood pressure among the two interventions at the 5% level of significance. The

confidence interval is used to give an indication of the range in which the true difference between the populations can take. The researchers will need to evaluate if the difference is clinically significant.

Interpretation of Statistical Significance

A result is commonly described as statistically significant when:

$$p < \alpha$$

For a significance level of 0.05:

$$p < 0.05$$

indicates statistical significance.

If:

$$p \geq 0.05$$

The finding is generally referred to as statistically insignificant.

It should be noted that "statistically insignificant" does not mean that the effect is not present. It simply means that the research did not provide enough statistical proof to refute the null hypothesis within the given procedure.

REFERENCES

1. Akram, W., & Garud, N. (2021). Design expert as a statistical tool for optimization of 5-ASA-loaded biopolymer-based nanoparticles using Box Behnken factorial design. *Future Journal of Pharmaceutical Sciences, 7*, 146.
2. Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N. J., Shkedy, Z., & Van der Elst, W. (2016). *Applied surrogate endpoint evaluation methods with SAS and R*. CRC Press.
3. Apweiler, R., Beissbarth, T., Berthold, M. R., Blüthgen, N., Burmeister, Y., Dammann, O., Deutsch, A., Feuerhake, F., Franke, A., Hasenauer, J., Hoffmann, S., Höfer, T., Jäger, D., Kaderali, L., Kestler, H. A., Mewes, H. W., Müller, O. J., Saez-Rodriguez, J., & Wolkenhauer, O. (2018). Whither systems medicine? *Experimental & Molecular Medicine, 50*, e453.
4. Banerjee, S. (2016). Spatial data analysis. *Annual Review of Public Health, 37*, 47–60.
5. Bordukova, M., Makarov, N., Rodriguez-Esteban, R., Schmich, F., & Menden, M. P. (2024). Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opinion on Drug Discovery, 19*(1), 33–42.
6. Bujak, R., Struck-Lewicka, W., Markuszewski, M. J., & Kaliszan, R. (2015). Metabolomics for laboratory diagnostics. *Journal of Pharmaceutical and Biomedical Analysis, 113*, 108–120.
7. Chow, S.-C., Shao, J., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculations in clinical research* (3rd ed.). Chapman and Hall/CRC.
8. Collins, D. C., Sundar, R., Lim, J. S. J., & Yap, T. A. (2017). Towards precision medicine in the clinic: From biomarker discovery to novel therapeutics. *Trends in Pharmacological Sciences, 38*(1), 25–40.
9. Commenges, D., & Jacqmin-Gadda, H. (2015). *Dynamical biostatistical models*. CRC Press.
10. DeJong, C., Aguilar, T., Tseng, C. W., Lin, G. A., Boscardin, W. J., & Dudley, R. A. (2016). Pharmaceutical industry-sponsored meals and physician prescribing patterns for Medicare beneficiaries. *JAMA Internal Medicine, 176*(8), 1114–1122.
11. Dermawan, D., & Alotaiq, N. (2025). From lab to clinic: How artificial intelligence (AI) is reshaping drug discovery timelines and industry outcomes. *Pharmaceuticals, 18*(7), 981.
12. Ewusie, J. E., Soobiah, C., Blondal, E., Beyene, J., Thabane, L., & Hamid, J. S. (2020). Methods, applications and challenges in the analysis of interrupted time series data: A scoping review. *Journal of Multidisciplinary Healthcare, 13*, 411–423.

13. Goldsack, J. C., Coravos, A., Bakker, J. P., Bent, B., Dowling, A. V., Fitzer-Attas, C., Godfrey, A., Godino, J. G., Gujar, N., Izmailova, E., Manta, C., Peterson, B., Vandendriessche, B., Wood, W. A., Wang, K. W., & Dunn, J. (2020). Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for biometric monitoring technologies (BioMeTs). *npj Digital Medicine, 3*, 55.

14. Gray, A. P., Chung, E., Hsu, R. L., Araki, D. T., & others. (2025). Global, regional, and national sepsis incidence and mortality, 1990–2021: A systematic analysis. *The Lancet Global Health*.

15. Gyawali, B., de Vries, E. G. E., Dafni, U., Amaral, T., Barriuso, J., Bogaerts, J., Cobleigh, M., de Azambuja, E., Haanen, J., Lassen, U., & others. (2021). Biases in study design, implementation, and data analysis that distort the appraisal of clinical benefit and ESMO-Magnitude of Clinical Benefit Scale (ESMO-MCBS) scoring. *ESMO Open, 6*(3), 100117.

16. Harris, E. K., & Albert, A. (2024). *Survivorship analysis for clinical studies*. CRC Press.

17. Hazra, A., & Gogtay, N. (2016). Biostatistics series module 1: Basics of biostatistics. *Indian Journal of Dermatology, 61*(1), 10–20.

18. Hazra, A., & Gogtay, N. (2016). Biostatistics series module 3: Comparing groups: Numerical variables. *Indian Journal of Dermatology, 61*(3), 251–260.

19. Hazra, A., & Gogtay, N. (2016). Biostatistics series module 4: Comparing groups—Categorical variables. *Indian Journal of Dermatology, 61*(4), 385–392.

20. Hazra, A., & Gogtay, N. (2016). Biostatistics series module 6: Correlation and linear regression. *Indian Journal of Dermatology, 61*(6), 593–601.

21. Johnson, S. C., Cunningham, M., Dippenaar, I. N., Sharara, F., Wool, E. E., Agesa, K. M., Han, C., Miller-Petrie, M. K., Wilson, S., Fuller, J. E., Balassyano, S., & others. (2021). Public health utility of cause of death data: Applying empirical algorithms to improve data quality. *BMC Medical Informatics and Decision Making, 21*, 175.

22. Kelly, W. K., & Halabi, S. (Eds.). (2018). *Oncology clinical trials: Successful design, conduct, and analysis*. Demos Medical Publishing.

23. Kromer, A. P. E., Sieber-Schäfer, F., & others. (2024). Design of experiments grants mechanistic insights into the synthesis of spermine-containing PBAE copolymers. *ACS Applied Materials & Interfaces, 16*(29), 37545–37557.

24. Lavielle, M., & Ribba, B. (2016). Enhanced method for diagnosing pharmacometric models: Random sampling from conditional distributions. *Pharmaceutical Research, 33*, 2979–2988.

25. Leider, J. P., Castrucci, B. C., Plepys, C. M., Blakely, C., Burke, E., & Sprague, J. B. (2015). On academics: Characterizing the growth of the undergraduate public health major: U.S., 1992–2012. *Public Health Reports, 130*(1), 104–113.

26. Lu, Y., Fang, J. Q., Tian, L., & Jin, H. (2015). *Advanced medical statistics*. World Scientific.

27. Ludwig, R. J., Anson, M., Zirpel, H., Thaci, D., & others. (2025). A comprehensive review of methodologies and application to use the real-world data and analytics platform TriNetX. *Frontiers in Pharmacology, 16*, 1516126.

28. Olsen, A. A., McLaughlin, J. E., & Harpe, S. E. (2020). Using multiple linear regression in pharmacy education scholarship. *Currents in Pharmacy Teaching and Learning, 12*(10), 1258–1268.

29. Panayides, A. S., Pattichis, M. S., Leandrou, S., Pitris, C., Constantinidou, A., & Pattichis, C. S. (2018). Radiogenomics for precision medicine with a big data analytics perspective. *IEEE Journal of Biomedical and Health Informatics, 23*(5), 2063–2079.

30. Proitsi, P., Kim, M., Whiley, L., Pritchard, M., Leung, R., Soininen, H., Kloszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Sham, P., Lovestone, S., Powell, J. F., & Legido-Quigley, C. (2015). Plasma lipidomics analysis finds long chain cholesteryl esters to be associated with Alzheimer's disease. *Translational Psychiatry, 5*, e494.

31. Qian, T., Zhu, S., & Hoshida, Y. (2019). Use of big data in drug development for precision medicine: An update. *Expert Review of Precision Medicine and Drug Development, 4*(3), 189–200.

32. Robaczewska, J., Vanhaverbeke, W., & Lorenz, A. (2019). Applying open innovation strategies in the context of a regional innovation ecosystem: The case of Janssen Pharmaceuticals. *Global Transitions, 1*, 120–131.

33. Sadiq, I. Z., Usman, A., Muhammad, A., & others. (2025). Sample size calculation in biomedical, clinical and biological sciences research. *Journal of Umm Al-Qura University for Applied Sciences*.

34. Saleem, F., Hassali, M. A., Shafie, A. A., Ul Haq, N., Farooqui, M., Aljadhay, H., & Ahmad, F. U. D. (2015). Pharmacist intervention in improving hypertension-related knowledge, treatment medication adherence and health-related quality of life: A non-clinical randomized controlled trial. *Health Expectations, 18*(5), 1270–1281.

35. Sallam, M., Salim, N. A., Barakat, M., & Al-Tammemi, A. B. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J, 3*(1), e103.

36. Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M. J., & Williams, Z. (2016). Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Medical Research Methodology, 16*, 77.

37. Turongkaravee, S., Jittikoon, J., Lukkunaprasit, T., Sangroongruangsri, S., Chaikledkaew, U., Thakkinstian, A., & others. (2021). A systematic review and meta-analysis of genotype-based and individualized data analysis of *SLCO1B1* gene and statin-induced myopathy. *The Pharmacogenomics Journal, 21*, 296–307.

38. Wali, N., & Hashimi, M. S. (2025). Advanced computational modeling and uncertainty analysis for COVID-19 projection: A data science approach to real-world epidemiological challenges. *Applied Data Science Research*.

39. Weisberg, H. I., & Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials, 12*(4), 357–364.

40. Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics, 20*(2), 273–286.

Unit 4...

CORRELATION AND REGRESSION IN PHARMACEUTICAL RESEARCH

This unit highlights some of the most important statistical techniques used to analyze and understand the relationship between variables in pharmaceutical, clinical and biological studies. The topics that will be covered under this unit include correlation, Pearson's correlation coefficient, scatter diagrams, dose-response relationship, simple linear regression, regression coefficients, odds ratio and risk analysis in clinical settings. These techniques allow the researcher to find out whether the relationship between two variables is positive or negative, what trend exists and how one can use this information to predict future results. Most importantly, this unit emphasizes the importance of interpreting the results of statistical analysis correctly.

4.1. FUNDAMENTALS OF CORRELATION

Correlation is a statistical method for studying the association of two variables. This method can tell whether changes in one variable are correlated with changes in another and whether the relationship is positive or negative as well as strong or weak. The correlation analysis technique is especially relevant for pharmaceutical, clinical and biological studies, as scientists have to investigate the correlation between some measurable traits quite often. Such a correlation might exist between drug dosage and plasma level of the drug, age and blood pressure, body weight and drug dose, or drug level and effectiveness.

Correlation does not indicate a causal link between the variables, but rather tells whether the variables are changing systematically in a particular way. Thus, this method is mostly used to find out the existence of an association and its magnitude, while causal relationships have to be investigated with the help of other methodologies.

4.1.1. Concept of Correlation and Pearson Correlation Coefficient

Correlation describes the extent and direction of relationship between two quantitative variables in terms of how they tend to increase or decrease together. The Pearson correlation coefficient (r) is a statistical tool that measures the strength and direction of linear relationship between two variables in the form of values ranging from −1 to +1. If the value of r is positive, then it represents a positive relationship between the two variables; if it is negative, then it represents a negative relationship; and if it is near to zero, it means that there is no significant linear relationship between them. For instance,

a positive correlation can be found between the dose of drug and the plasma concentration of the drug, wherein the higher dose will result in a higher concentration of the drug.

A. Concept of Correlation

Correlation refers to the degree and direction of association between two variables.

Suppose we have two variables:

- **X** = Drug dose
- **Y** = Plasma drug concentration

If an increase in one variable results in an increase in the other, there exists a positive correlation between the variables.

If an increase in one variable results in a decrease in the other, there exists a negative correlation between the variables.

For example:

Table 1: Relationship Between Drug Dose and Plasma Drug Concentration

Drug Dose (mg)	Plasma Concentration (µg/mL)
50	4
100	8
150	12
200	16
250	20

As the drug dose rises here, there is a rise in the plasma drug concentration as well. This shows that the two variables have a positive correlation.

There are three important aspects of correlation which may be explained as follows:

1. Direction: Direction indicates whether the variables move:

- In the same direction → positive correlation
- In opposite directions → negative correlation
- Without a systematic linear pattern → little or no linear correlation

2. Strength: Strength represents the degree of the relationship between the two variables. The closer the correlation is to +1 or −1, the stronger the relationship is; on the other hand, the closer the correlation is to 0, the weaker the relationship is.

3. Type of Relationship: Correlation, especially Pearson correlation, looks for a linear relationship. The linear relationship is the relationship where changes in one variable can be illustrated by the straight line.

B. Pearson Correlation Coefficient

Pearson's Correlation Coefficient is one of the most popularly used correlation coefficients in measuring correlation between two variables.

It is denoted by the symbol:

$$r$$

The value of $r$ranges from:

$$-1 \leq r \leq +1$$

Pearson's correlation coefficient measures the degree to which two numeric variables are related linearly.

The equation is:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

Where:

- r= Pearson correlation coefficient
- X= individual values of the first variable
- Y= individual values of the second variable
- $\bar{X}$= mean of the first variable
- $\bar{Y}$= mean of the second variable

The numerator indicates the covariance between the two variables, whereas the denominator normalizes this figure to ensure that the correlation coefficient remains within the range of −1 and +1.

Interpretation of Pearson Correlation Coefficient

The value of $r$provides information about both direction and strength.

Table 2: Interpretation of Pearson's Correlation Coefficient (r)

Pearson's r	**Interpretation**
+1.00	Perfect positive correlation
+0.70 to +0.99	Strong positive correlation

+0.40 to +0.69	Moderate positive correlation
+0.10 to +0.39	Weak positive correlation
0.00	No linear correlation
−0.10 to −0.39	Weak negative correlation
−0.40 to −0.69	Moderate negative correlation
−0.70 to −0.99	Strong negative correlation
−1.00	Perfect negative correlation

Such values are only indicative since what constitutes a "strong" or "weak" correlation may vary depending on the scientific discipline and circumstances of research.

Example 1: Strong Positive Correlation

Suppose a study obtains:

$$r = +0.90$$

This implies that there is a very positive correlation between the two variables.

For instance, the increased drug dosage may usually correspond with an increase in drug concentration in the blood plasma.

Example 2: Strong Negative Correlation

Suppose:

$$r = -0.85$$

This suggests a clear negative linear relationship.

For instance, where an increase in one of the specific variables concerning the treatment leads to decrease in symptoms, there could be a negative correlation between the two variables.

Example 3: Weak Correlation

Suppose:

$$r = +0.15$$

It means there is a weak positive linear correlation.

The variables may have a tendency to increase simultaneously, but the correlation is weak.

Example 4: Zero Correlation

Suppose:

$$r = 0.02$$

This shows that there is virtually no correlation between the variables along a straight line.

Nevertheless, it is crucial to note that although r=0, it doesn't necessarily imply that there isn't any relationship at all between the variables. The relationship could be nonlinear.

4.1.2. Interpretation of Positive and Negative Correlations

The sign of Pearson's correlation coefficient indicates the direction of the relationship.

There are three basic possibilities:

1. Positive correlation
2. Negative correlation
3. No or very weak linear correlation

A. Positive Correlation

A positive correlation occurs when two variables tend to move in the same direction.

In mathematical terms:

$$r > 0$$

This means:

- When Xincreases, Ytends to increase.
- When Xdecreases, Ytends to decrease.

Example

Consider the relationship between drug dose and plasma drug concentration:

Table 3: Drug Dose and Corresponding Plasma Concentration

Dose (mg)	Concentration (µg/mL)
50	5
100	9
150	14
200	18
250	23

The general pattern is that higher doses are associated with higher plasma concentrations.

Therefore, the relationship is positive.

If the calculated correlation is:

$$r = +0.92$$

the relationship is considered a strong positive correlation.

Important point

The plus sign does not mean that the correlation is necessarily strong.

For example:

$$r = +0.20$$

And

$$r = +0.90$$

are both positive, but $+0.90$represents a much stronger linear relationship.

Thus:

The sign tells us the direction, while the absolute value tells us the strength.

B. Negative Correlation

A negative correlation occurs when two variables tend to move in opposite directions.

In mathematical terms:

$$r < 0$$

This means:

- When Xincreases, Ytends to decrease.
- When Xdecreases, Ytends to increase.

Example

Suppose researchers investigate the relationship between a physical activity score and resting heart rate:

Table 4: Variation in Resting Heart Rate with Activity Score

Activity Score	Resting Heart Rate
20	84
30	80
40	76
50	72
60	68

As activity score increases, resting heart rate tends to decrease.

Therefore, the relationship is negative.

If:

$$r = -0.88$$

the variables have a strong negative linear correlation.

Again, the negative sign indicates the direction, not weakness.

For example:

$$r = -0.20$$

is a weak negative correlation, whereas:

$$r = -0.90$$

is a strong negative correlation.

C. Comparison of Positive and Negative Correlation

Table 1 is an illustration of correlation that shows the positive and negative types with respect to the sign, the direction of the relationship, X and Y changes, and healthcare/pharmaceutical applications. Positive correlation involves changes of the two variables in the same direction, while negative correlation involves changes of the two variables in different directions.

Table 5: Comparison of Positive and Negative Correlation

Feature	Positive Correlation	Negative Correlation
Sign of r	$r > 0$	$r < 0$
Direction	Same direction	Opposite directions
X increases	Y tends to increase	Y tends to decrease
X decreases	Y tends to decrease	Y tends to increase
Example	Dose ↑, concentration ↑	Activity ↑, heart rate ↓
Strong example	$r = +0.90$	$r = -0.90$

Notice that:

$$r = +0.90$$

And

$$r = -0.90$$

have approximately the same strength, but their directions are opposite.

4.2. VISUALIZATION OF RELATIONSHIPS IN PHARMACEUTICAL DATA

Pharmaceutical relationship visualization is the term that implies the use of visualization and analysis techniques for the purpose of discovering the nature of the interdependence of two or more variables. In pharmaceutical research, there are numerous variables which do not behave independently. For instance, the dose of medication could impact plasma concentrations, medication concentration might affect the treatment outcome, and treatment duration could impact the patient's results. Visualizations can help the researchers to discover patterns, trends, nature of the interdependence of variables, anomalies, and potential associations.

In some cases, a proper visualization will give a better representation of the relationship compared to numerical data. Some of the most effective visualization tools are scatter plots and dose-response relationship plots. They are especially valuable in such fields as pharmacology, pharmacokinetics, toxicology, and drug development.

4.2.1. Scatter Plots and Trend Visualization

The scatter plot is a tool that is used to plot the association between two quantitative variables such that each value in the dataset is represented by a dot using its values for both the explanatory and response variable. In data analysis in the pharmaceutical industry, scatter plots assist analysts to determine trends such as the direction and nature of relationships, outliers and linearity. The use of trend visualization assists analysts in understanding whether the response increases, decreases, does not change, or changes in a non-linear manner with respect to the change in the explanatory variable. For instance, a positive trend is noted when there is an increase in the drug dosage leading to an increase in the plasma drug concentration whereas a negative trend occurs when there is an increase in the drug dosage leading to a decrease in blood pressure.

❖ **Concept of a Scatter Plot**

Scatterplot refers to a graphical approach to depicting the relationship between two numerical variables. The observations are plotted according to their numerical values for two variables.

Generally:

- The X-axis is the independent variable
- The Y-axis is the dependent variable
- Each dot represents an individual observation or experimental subject

For example, in a pharmaceutical experiment, scientists may look into the connection between:

- Drug dose and plasma concentration of the drug
- Patient's age and drug metabolism rate

- Concentration of the drug and therapeutic effect
- Body mass and necessary drug dosage
- Treatment period and biomarker value
- Blood pressure and concentration of the drug.

Example

Suppose the following data are obtained from a pharmacokinetic study:

Table 6: Patient-Wise Drug Dose and Plasma Concentration Data

Patient	Drug Dose (mg)	Plasma Concentration (µg/mL)
1	50	2.1
2	100	4.0
3	150	5.8
4	200	8.2
5	250	9.7

Here:

- The drug dosage is the independent variable.
- The plasma concentration is the dependent variable.
- Each individual gives one set of data points.

If larger dosages are linked to increased plasma concentration, then the plotted points will indicate an increasing trend, showing a positive relationship.

Purpose of Scatter Plots

Scatter plots are useful for several purposes in pharmaceutical data analysis.

1. Identifying Relationships

Scatter diagram assists in checking if there is any relationship between the two variables. In this case, it could assist in checking whether an increase in dosage causes an increase in plasma concentration.

2. Determining Direction

The overall pattern of points can indicate whether the relationship is:

- **Positive** – both variables tend to increase together.
- **Negative** – one variable tends to increase as the other decreases.
- **No clear relationship** – changes in one variable do not show a consistent pattern with the other.

3. Assessing Strength

Observations clustered around an apparent trend form a visual clue regarding the degree of the relationship's strength. Observations clustered around an underlying trend suggest the existence of a strong relationship, while observations dispersed over a wider area reflect a weaker relationship.

4. Detecting Outliers

A scatter diagram is able to show the outliers in terms of data values. For instance, when most of the subjects have dose-concentration relationship, one subject having a very high concentration value is likely to need further examination.

Possible reasons include:

- Measurement error
- Incorrect sample handling
- Non-adherence
- Individual metabolic differences
- Drug interaction
- Genetic variation
- Unusual pharmacokinetic behavior

An outlier should not automatically be removed. Its scientific or clinical significance should first be evaluated.

5. Identifying Non-linear Relationships

All pharmaceutical associations are not always linear. The scatter diagram can exhibit an association that is non-linear in nature, like curved. For instance, the drug response may rise at an increasing rate with the dose but level off to attain its highest response rate.

❖ **Trend Visualization**

Trend visualization involves using graphic shapes or curves to depict the general trend of the relationship between variables.

A trend is useful for the researcher in finding out whether the response variable generally:

- increases,
- decreases,
- remains relatively constant, or
- changes in a non-linear manner

as the explanatory variable changes.

As another example, when the drug dosage is increased from 50 mg to 250 mg, and the plasma concentration increases as well, there is a trend in the dataset that is increasing.

In addition, the trend line can be used to show the general association of two variables. Nevertheless, the trend line has to be considered with care because it shows the general tendency but not all observations.

Types of Trends in Pharmaceutical Data

1. **Positive Trend:** A positive trend is said to exist when increased values of one variable are accompanied by increased values of another variable.

Example: Increasing drug dose → increasing plasma drug concentration.

2. **Negative Trend:** It is said that there is a negative trend if high values of one variable correspond to low values of the other variable.

Example: Increasing drug dose → decreasing blood pressure.

3. **No Clear Trend:** When the points are spread out without any pattern, then there is likely to be no correlation between the two variables.

Example: Patient identification number and drug response would normally have no meaningful relationship.

4. **Non-linear Trend:** In non-linear relationships, the relationship takes a change in direction or speed as the explanatory variable changes.

Example: The drug concentration can cause an increasing therapeutic response at low concentrations and then level off at high concentrations.

❖ **Relationship Between Scatter Plots and Correlation**

Scatter Plots & Correlation are related concepts. The correlation coefficient (r) indicates numerically the nature and extent of a linear relationship, while the scatter plot gives us a graphical representation of that relationship.

For instance:

- A positive r-value indicates a positive linear relationship.
- A negative r-value indicates a negative linear relationship.
- A small r-value indicates an absence of linear relationship.

But we must not rely on correlation alone without looking at the data graphically. This is because some outliers or non-linear relationships can have significant effects on the correlation coefficient.

4.2.2. Dose-Response Relationships and Their Interpretation

Dose-response relationship is a study of the relation between the dose of a drug and the intensity of the effect it creates on the biological system. The information gained through dose-response relationship is useful in studying the effect that different doses of the drug can have on the body, and is used in the fields of drug discovery, pharmacodynamics, toxicology, clinical trials, dosing, and safety testing. In general, as the dose increases, there can be more intensity of effect until the effect reaches its maximum level known as Emax, and then as the dose further increases, there would be no much increase in the effect, giving a plateau. Dose-response relationship is also useful in determining potency, which is how much drug is needed to produce a particular effect, and efficacy, which is the maximum effect that the drug can create.

❖ **Concept of Dose-Response Relationship**

Dose response relationship refers to the correlation between the dosage or concentration of a particular drug and the effect of the drug on the body. It is one of the most vital relationships in pharmacology since it helps the scientists understand how alterations in the drug dosage will affect the body response.

Generally:

Dose → Drug exposure → Biological response

For instance, if the dosage of an antihypertensive medication is increased, the reduction in blood pressure will be higher, up to some limit.

Dose-response analysis is used in:

- Drug discovery
- Preclinical studies
- Clinical trials
- Pharmacodynamics
- Toxicology
- Dose selection
- Safety evaluation
- Determination of therapeutic ranges

Basic Components of a Dose-Response Relationship

1. Dose: The dose is the quantity of drug administered to a patient or experimental subject.

It may be expressed as:

- mg

- mg/kg
- µg/kg
- mg/day
- concentration units

For example, a drug may be administered at doses of 10 mg, 20 mg, 40 mg, and 80 mg.

2. Response: The response is the measurable biological, physiological, or clinical effect produced by the drug.

Examples include:

- Reduction in blood pressure
- Reduction in blood glucose
- Increase in heart rate
- Reduction in pain score
- Inhibition of bacterial growth
- Change in enzyme activity
- Tumor-cell inhibition

The response may be expressed as a numerical measurement, percentage change, or other suitable clinical or biological measure.

Interpretation of Dose-Response Relationships

A dose-response relationship is generally interpreted by examining how the response changes as the dose increases.

1. Increasing Response

At lower doses, increasing the dose may produce an increasing response.

For example:

Table 7: Drug Dose and Reduction in Blood Pressure

Dose (mg)	Reduction in Blood Pressure (mmHg)
10	4
20	8
40	14
80	18

The response increases as the dose increases.

This suggests that the drug is producing a stronger pharmacological effect over this dose range.

2. Maximum Response

If high enough doses are administered, there is no doubt that there comes a time when the maximum response level is reached. At this time, administering more doses will add nothing to the therapeutic effectiveness of the drug. The maximum possible response is normally denoted by Emax, which stands for maximum drug effect. The reason for this is that going past this level of dosage may increase the chances of negative consequences.

3. Plateau:

A plateau occurs when increasing the dose produces little or no further increase in response.

For example:

Table 8: Drug Dose and Corresponding Response

Dose (mg)	Response (%)
10	15
20	35
40	60
80	85
160	87

The response increases substantially at first but changes very little between 80 mg and 160 mg.

This indicates that the drug is approaching its maximum effect.

4. Potency

Potency can be defined as the quantity or concentration of a drug necessary to elicit a certain response. The more potent a drug is, the less quantity or concentration is needed to elicit a certain response compared to another drug which is less potent. Thus, in an experiment where Drug A elicits a 50% response at 10 mg while Drug B elicits the same 50% response at 50 mg, Drug A is more potent.

The concept of potency is often related to EC50 or ED50.

- **EC50 =** concentration that causes 50% of the maximal effect.
- **ED50 =** dose that causes 50% of a stated effect.

5. Efficacy

Efficacy is defined as the maximal response that the drug can elicit at experimental conditions. The drug may be very potent, yet still elicit less maximal efficacy than another drug.

Consequently, potency and efficacy are two distinct notions:

Table 9: Comparison of Drug Potency and Efficacy

Feature	Potency	Efficacy
Meaning	Amount required to produce a specified effect	Maximum effect that can be produced
Common measure	EC50 or ED50	Emax
Main question	"How much drug is needed?"	"How much effect can the drug produce?"

6. Therapeutic and Toxic Responses

Dose-response relationships are essential in knowing the correlation between therapeutic action and toxicity.

With increasing dosage:

- The therapeutic effect is increased.
- At the same time, adverse effects are increased.
- At some high level of exposure, toxicity is observed.

Hence, dose-response data can help in finding therapeutic doses without unacceptable levels of toxicity.

7. Graded and Quantal Dose-Response Relationships

Dose-response relationships can broadly be studied in two ways.

Graded Dose-Response Relationship

A graded response measures the magnitude of response in an individual biological system or subject.

Examples:

- Reduction in blood pressure
- Change in heart rate
- Degree of enzyme inhibition
- Change in blood glucose

The response can take different numerical values.

Quantal Dose-Response Relationship

A quantal response records whether a defined event occurs or does not occur.

Examples:

- Patient responds / does not respond
- Animal survives / does not survive
- Seizure occurs / does not occur

The response is therefore categorized into distinct outcomes.

Factors Affecting Dose-Response Relationships

Dose-response relationship differs from individual to individual as well as from experiment to experiment. The important parameters are:

1. **Patient characteristics:** Age, body weight, gender, genetics, physiological state of the patient can influence the response to the drug.
2. **Pharmacokinetics:** Factors related to absorption, distribution, metabolism and excretion affect the drug concentration at the target site.
3. **Pharmacodynamics:** Receptor sensitivity and the number of receptors as well as the biological effects can influence the dose-response curve.
4. **Disease state:** Same dose of the drug will have different effect on healthy and diseased patients.
5. **Drug interactions:** Other drugs may enhance or diminish the effect of certain drug.
6. **Tolerance:** Dose-response curve should be interpreted considering the tolerance.

4.3. SIMPLE LINEAR REGRESSION

The simple linear regression analysis is a statistical tool that helps to explain and model the association between an independent or explanatory variable (X) and dependent or response variable (Y). This tool becomes very important for researchers who are interested in exploring how one pharmaceutical or clinical variable is related to another one and predicting the response variable using values of the explanatory variable. Simple linear regression analysis may help researchers to explore the association between drug dose (X) and plasma concentration of the drug (Y) or the association between treatment duration (X) and the biomarker (Y).

4.3.1. Concept and Structure of Simple Linear Regression

Simple linear regression is a statistical technique used to investigate the linear relationship between two variables such that the value of one of the variables (X) explains or predicts the value of the other

variable (Y). This relationship can be expressed using the model Yˆ= b_0 + b_1 X, where Yˆis the estimated value of Y, b_0 is the intercept, and b_1 is the slope or regression coefficient. In this case, drug dose (X) can be used to predict the plasma drug concentration (Y). The intercept can be interpreted as the estimated value of Y when the value of X is zero while the slope is the estimated value of change in Y resulting from a unit change in X. Regression also takes into account the distinction between the observed and the estimated value, whose difference is referred to as the residual.

❖ **Concept of Simple Linear Regression**

Linear regression tests whether there is any linear relationship between two quantitative variables. The test takes into consideration the available information to find an equation of a straight line that describes the overall relationship between the two variables.

The basic regression equation is:

$$\hat{Y} = b_0 + b_1 X$$

Where:

- $\hat{Y}$= predicted value of the dependent variable
- X= value of the independent or explanatory variable
- b_0= intercept
- b_1= regression coefficient or slope

The equation allows researchers to estimate the expected value of Y for a given value of X.

For example, if:

- **X = drug dose**
- **Y = plasma drug concentration**

a regression equation can be used to estimate the expected plasma concentration for a particular drug dose.

Independent and Dependent Variables

Simple linear regression involves two main variables.

1. Independent or Explanatory Variable (X)

The independent variable is the variable used to explain or predict changes in the response variable.

Examples include:

- Drug dose
- Treatment duration

- Patient age
- Body weight
- Drug concentration

2. Dependent or Response Variable (Y)

The dependent variable is the outcome that researchers want to explain or predict.

Examples include:

- Plasma drug concentration
- Blood pressure
- Blood glucose level
- Biomarker concentration
- Therapeutic response

For example:

Drug dose (X) → Plasma drug concentration (Y)

Here, drug dose is the explanatory variable and plasma drug concentration is the response variable.

Structure of the Regression Equation

The simple linear regression equation is:

$$\hat{Y} = b_0 + b_1 X$$

Intercept (b_0): The intercept represents the predicted value of Y when X is equal to zero.

For example, if a regression model relates drug dose to a response:

$$\hat{Y} = 5 + 0.8X$$

the intercept is 5.

This means that the predicted value of Y is 5 when X = 0.

However, the intercept may not always have a meaningful biological interpretation, particularly when X = 0 is outside the range of the observed data.

Slope (b_1): The slope represents the expected change in Y for a one-unit increase in X.

In:

$$\hat{Y} = 5 + 0.8X$$

the slope is 0.8.

This means that for every one-unit increase in X, the predicted value of Y increases by 0.8 units, on average.

If the slope is positive:

$$b_1 > 0$$

Y tends to increase as X increases.

If the slope is negative:

$$b_1 < 0$$

Y tends to decrease as X increases.

If:

$$b_1 = 0$$

there is no linear change in the predicted value of Y with X.

Example of Simple Linear Regression

Suppose a pharmaceutical researcher wants to examine the relationship between drug dose and plasma drug concentration.

The estimated regression equation is:

$$\hat{Y} = 1.2 + 0.04X$$

where:

- X = drug dose in mg
- Y = predicted plasma concentration in µg/mL
- 1.2 = intercept
- 0.04 = slope

For a dose of 100 mg:

$$\hat{Y} = 1.2 + (0.04 \times 100)$$
$$\hat{Y} = 5.2$$

Therefore, the predicted plasma concentration is 5.2 µg/mL according to this regression model.

Observed and Predicted Values

Regression distinguishes between the observed value and the predicted value.

- **Observed value (Y):** The actual value measured in the study.
- **Predicted value ($\hat{Y}$):** The value estimated by the regression equation.

The difference between the observed and predicted values is called the residual.

$$e = Y - \hat{Y}$$

A small residual indicates that the predicted value is close to the observed value, whereas a large residual indicates a greater difference between the model prediction and the actual observation.

4.3.2. Interpretation of Regression Coefficients

Regression coefficients provide important information about the relationship between X and Y.

The two main coefficients in simple linear regression are:

- **Intercept (b_0)**
- **Slope (b_1)**

1. Interpretation of the Intercept (b_0)

The intercept represents the expected or predicted value of Y when X = 0.

For example:

$$\hat{Y} = 10 + 0.5X$$

Here:

$$b_0 = 10$$

Therefore, when X is zero, the predicted value of Y is 10 units.

Pharmaceutical Example

When X denotes time in treatment and Y denotes biomarker level, the intercept is the biomarker level predicted when the time in treatment is zero. Nonetheless, this would be true only if the value of X=0 is scientifically sensible and within an acceptable range of values in the data set.

2. Interpretation of the Slope (b_1)

Slope is always considered the most critical regression coefficient for understanding the relationship. It refers to the average increase in Y corresponding to one unit of X.

For example:

$$\hat{Y} = 2 + 0.6X$$

Here:

$$b_1 = 0.6$$

This means that for every one-unit increase in X, the predicted value of Y increases by 0.6 units on average.

Positive Regression Coefficient

If:

$$b_1 > 0$$

the regression relationship is positive.

For example:

$$\hat{Y} = 3 + 0.5X$$

The positive slope indicates that increasing X is associated with an increase in the predicted value of Y.

Pharmaceutical example:

Drug dose ↑ → plasma drug concentration ↑

Negative Regression Coefficient

If:

$$b_1 < 0$$

the regression relationship is negative.

For example:

$$\hat{Y} = 100 - 0.7X$$

The negative slope indicates that increasing X is associated with a decrease in the predicted value of Y.

Pharmaceutical example:

Drug dose ↑ → blood pressure ↓

when blood pressure is used as the response variable.

Zero Regression Coefficient

If:

$$b_1 = 0$$

the model predicts no linear change in Y as X changes.

This suggests that X does not provide a linear predictive relationship for Y within the model.

Example of Interpreting Regression Coefficients

Suppose a study produces the regression equation:

$$\hat{Y} = 20 - 0.4X$$

where:

- X = drug dose in mg
- Y = systolic blood pressure in mmHg

Intercept

$$b_0 = 20$$

The intercept shows the estimated value of Y where X equals zero. Interpretation of the intercept makes sense only if a zero-dosage level holds any significance for the problem under study.

Slope

$$b_1 = -0.4$$

As the slope of the graph is negative, it implies that an increase in the dosage of the drug by one unit leads to 0.4 units fall in systolic blood pressure.

Regression Coefficients and Units

An important feature of regression coefficients is that their interpretation depends on the units of measurement.

For example, if:

- X = dose in mg
- Y = concentration in µg/mL

and:

$$b_1 = 0.05$$

The coefficient tells us that a 1 mg change in the dose causes an expected 0.05 µg/mL increase in the concentration.

Therefore, it is always essential to consider units when interpreting regression coefficients.

Regression and Correlation

Regression and correlation are closely related but serve different purposes.

Table 10: Comparison of Correlation and Simple Linear Regression

Feature	Correlation	Simple Linear Regression
Main purpose	Measures association	Models and predicts relationship
Variables	Two variables	X and Y
Direction	Positive or negative	Determined by slope
Main measure	Pearson's r	Intercept and slope
Prediction	Not its primary purpose	Important purpose
Units	Correlation is unitless	Coefficients have units
Causation	Does not establish causation	Does not establish causation by itself

Correlation helps determine the linear relationship between two variables, while regression gives us the equation that can be used to predict the response variable based on the predictor variable.

4.4. ODDS RATIO AND CLINICAL RISK ANALYSIS

Odds ratio (OR) and risk analysis in clinical studies are two statistical methods used in medical research and pharmacology to assess the relationship between a risk factor and an outcome. The application of these methods is especially convenient for situations where the outcome is categorical, that is, either the presence or the absence of a disease, the occurrence or lack of adverse reaction to medication, and the success or failure of a treatment. Using OR, researchers can estimate how likely an outcome is in each group compared to the other one.

4.4.1. Introduction to Odds Ratio

Odds ratio (OR) is a statistical tool that allows one to estimate the odds of the occurrence of an event in comparison between two groups, for example, exposed vs. unexposed groups, treatment vs. control groups, etc. The term "odds" refers to the proportion of the probability of the event to occur against the probability of the event not to occur; the odds ratio allows one to compare odds between groups and is usually calculated according to a 2 × 2 contingency table using $OR = \frac{ad}{bc}$. OR = 1 means no correlation, OR > 1 suggests higher odds of the event occurring in the exposed group, while OR < 1 – lower odds. Odds ratio is widely applied in medicine and pharmacology studies and is particularly useful in the estimation of adverse drug reactions, risk factors for the development of diseases, effectiveness of treatments, etc. It is usually provided with a 95% confidence interval, which contains 1 meaning lack of statistical significance at the 5% level.

- **Concept of Odds**

It is crucial for one to understand the concept of odds before comprehending the odds ratio. This is because odds refers to the ratio of the likelihood of occurrence of an event and its non-occurrence.

$$\text{Odds} = \frac{P(\text{Event})}{1 - P(\text{Event})}$$

For example, suppose 20 out of 100 patients experience an adverse drug reaction.

Probability of adverse reaction:

$$P = \frac{20}{100} = 0.20$$

Probability of no adverse reaction:

$$1 - 0.20 = 0.80$$

Therefore:

$$\text{Odds} = \frac{0.20}{0.80} = 0.25$$

Thus, the odds of experiencing the adverse reaction are 0.25.

❖ **Concept of Odds Ratio**

The odds ratio is a statistical term which involves comparing the probability of an event happening in one group with that of happening in another group.

It is commonly used to compare:

- Exposed vs. unexposed groups
- Treatment vs. control groups
- Patients receiving one drug vs. another drug
- Patients with a risk factor vs. those without it

The basic interpretation is:

$$OR = \frac{\text{Odds of outcome in exposed group}}{\text{Odds of outcome in unexposed group}}$$

An OR therefore tells us how the odds of an outcome differ between two groups.

❖ **2 × 2 Contingency Table**

Odds ratio is commonly calculated using a 2 × 2 contingency table.

Table 11: Two-by-Two Contingency Table for Exposure and Outcome

	Outcome Present	**Outcome Absent**	**Total**
Exposed	a	b	a+b
Unexposed	c	d	c+d

Here:

- **a** = exposed individuals with the outcome
- **b** = exposed individuals without the outcome
- **c** = unexposed individuals with the outcome
- **d** = unexposed individuals without the outcome

The odds ratio is calculated as:

$$OR = \frac{a/b}{c/d}$$

which can also be written as:

$$\boxed{OR = \frac{ad}{bc}}$$

Example of Odds Ratio

Suppose a study investigates whether exposure to a particular drug is associated with an adverse drug reaction.

Table 12: Distribution of Drug Exposure by Adverse Reaction Status

	Adverse Reaction	**No Adverse Reaction**
Drug exposed	30	70
Not exposed	10	90

Here:

$$a = 30,\ b = 70,\ c = 10,\ d = 90$$

Therefore:

$$OR = \frac{30 \times 90}{70 \times 10}$$

$$OR = \frac{2700}{700} = 3.86$$

Thus, the odds of experiencing the adverse reaction are approximately 3.86 times higher in the exposed group than in the unexposed group.

Interpretation of Odds Ratio

The value of OR provides information about the association between exposure and outcome.

Table 13: Interpretation of Odds Ratio Values

Odds Ratio	**General Interpretation**
OR = 1	No association between exposure and outcome
OR > 1	Higher odds of the outcome in the exposed group
OR < 1	Lower odds of the outcome in the exposed group

OR = 1

If:

$$OR = 1$$

the odds of the outcome are the same in both groups.

There is no evidence of an association based on the estimated OR.

OR > 1

If:

$$OR > 1$$

the exposure is associated with higher odds of the outcome.

For example:

$$OR = 2.5$$

means the exposed group has 2.5 times the odds of the outcome compared with the reference group.

OR < 1

If:

$$OR < 1$$

the exposure is associated with lower odds of the outcome.

For example:

$$OR = 0.40$$

indicates that the exposed group has 0.40 times the risk of developing the disease as the reference group.

An OR less than 1 could suggest a protective effect, based on the research methodology and outcome variable.

Odds Ratio and Relative Risk

Odds ratio should not be confused with relative risk (RR).

Relative risk compares probabilities or risks:

$$RR = \frac{\text{Risk in exposed group}}{\text{Risk in unexposed group}}$$

Odds ratio estimates odds not risk.

The difference is critical since in certain instances OR and RR may vary significantly in numeric terms especially when the event is prevalent.

Confidence Interval for Odds Ratio

An odds ratio is often reported together with a 95% confidence interval (CI).

For example:

$$OR = 2.50,\ 95\%\ CI = 1.40 - 4.46$$

This gives an idea about the uncertainty associated with the OR estimate. For interpretation of the common 95% CI:

- If the CI contains 1, then there is no statistical significance for the association at 5%.
- If the CI lies above 1, it indicates increased odds.
- If the CI lies below 1, it indicates decreased odds.

It is important to consider CI along with the study design and sample size.

4.4.2. Application of Odds Ratio in Clinical Risk Analysis

The odds ratio (OR) is used in clinical risk assessment for assessing the association between the exposure, intervention, risk factor, or patient variable and the clinical outcome, especially when the outcome is dichotomous. The OR is utilized in risk-factor analysis, drug safety surveillance, clinical intervention studies, case-control studies, and epidemiology for comparison of the odds of an outcome between different groups. An OR greater than one means there are higher odds of the outcome, whereas an OR less than one implies that there are lower odds of the outcome as compared with the reference group. In case the OR of Drug A for an adverse effect is 3.2, then Drug A exposure is linked with higher odds of the adverse effect. Nonetheless, for clinical interpretation, other aspects need to be considered, including the confidence interval, p-value, baseline risk, type of study, sample size, potential confounders, and clinical relevance, since the OR simply shows association but not causation.

❖ **Concept**

Odds ratio is applied in clinical risk assessment to determine the relationship between an exposure, intervention, or risk factor and a specific clinical outcome.

This outcome will typically have two levels only, and this means that it could be a case of:

- Disease present / disease absent
- Adverse reaction / no adverse reaction
- Treatment success / treatment failure
- Complication / no complication
- Hospitalization / no hospitalization
- Death / survival

1. Analysis of Risk Factors

Odds ratios can be used in examining whether a certain risk factor leads to an increased chance of developing a clinical condition.

For instance, one can examine whether a certain patient attribute leads to an increased frequency of adverse reactions due to a certain drug.

If:

$$OR > 1$$

the factor is associated with higher odds of the outcome.

If:

$$OR < 1$$

the factor is associated with lower odds.

2. Evaluation of Adverse Drug Reactions

Pharmacovigilance studies can use odds ratios to investigate associations between drug exposure and adverse events.

For example:

Drug exposure → Adverse drug reaction

When the observed OR is much greater than 1, the information indicates that the odds of the adverse event for the exposed group is higher than that for the comparison group.

Nevertheless, the interpretation needs to be done in conjunction with confidence intervals, research design, confounding variables, and other information.

3. Clinical Treatment Studies

Odds ratios can also be used to compare treatment outcomes.

For example, suppose a clinical trial compares:

- **Drug A**
- **Drug B**

and the outcome is treatment success.

If:

$$OR = 1.8$$

When comparing Drug A to Drug B, the odds ratio of success is 1.8 times higher than for Drug B, assuming that the odds ratio is calculated with Drug B being the reference group.

The interpretation always depends on how the outcome and reference groups are defined.

4. Case-Control Studies

Odds ratio is especially significant in case-control research.

In a case-control study:

- Cases experience the outcome/disease.
- Controls do not experience the outcome/disease.
- Comparison of the previous exposure of both groups is made.

Odds ratio represents the relation between the exposure and the outcome.

As an illustration, researchers could compare previous exposure to the drug in cases and controls that did not have an adverse event.

5. Epidemiological Research

Odds ratios are also widely used in epidemiological studies to examine associations between exposures and diseases.

Examples include studying relationships between:

- Smoking and disease
- Drug exposure and adverse events
- Environmental exposure and illness
- Lifestyle factors and clinical outcomes
- Genetic factors and disease occurrence

The OR describes an association, but it does not automatically demonstrate causation.

6. Interpretation in Clinical Decision-Making

Odds ratio can provide useful evidence about whether a treatment or exposure is associated with a clinical outcome.

However, clinical interpretation should not rely on OR alone.

Researchers should consider:

- Magnitude of the OR
- Confidence interval
- Statistical significance
- Absolute risk
- Baseline risk
- Study design
- Sample size
- Potential confounding
- Clinical importance

A statistically significant association may not necessarily represent a clinically important effect.

Example of Clinical Risk Analysis

Suppose researchers' study whether exposure to Drug A is associated with a particular adverse event. The results are:

$$OR = 3.2$$

with:

$$95\%\ CI = 1.8 - 5.7$$

Since the calculated OR value is greater than 1, exposure to Drug A increases the odds of the event occurring compared to the reference group. Since the confidence interval is not inclusive of 1, there is a statistically significant relationship between the variable of interest and Drug A exposure at the 5% level of significance. Nevertheless, caution is required when interpreting this result because of the nature of the association.

REFERENCES

1. Awaisu, A., Bakdach, D., Elajez, R. H., & Zaidan, M. (2015). Hospital pharmacists' self-evaluation of their competence and confidence in conducting pharmacy practice research. *Saudi Pharmaceutical Journal, 23*(3), 257–265.
2. Beckett, L. A., Donohue, M. C., Wang, C., Aisen, P., Harvey, D. J., & Saito, N. (2015). The Alzheimer's Disease Neuroimaging Initiative phase 2: Increasing the length, breadth, and depth of our understanding. *Alzheimer's & Dementia, 11*(7), 823–831.
3. Bespalov, A., Wicke, K., & Castagné, V. (2019). Blinding and randomization. In A. Bespalov, M. C. Michel, & T. Steckler (Eds.), *Good research practice in non-clinical pharmacology and biomedicine* (pp. 81–100). Springer.
4. Blum, A., Hopcroft, J., & Kannan, R. (2017). *Foundations of data science*. Cambridge University Press.
5. Bors, D. (2018). *Data analysis for the social sciences: Integrating theory and practice*. SAGE Publications.
6. Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hassan, M., Laska, E., Lewis, G., Newhouse, P. A., & others. (2024). Illusory generalizability of clinical prediction models. *Science, 383*(6679), 164–167.
7. Conklin, L. S., Damsker, J. M., Hoffman, E. P., Jusko, W. J., Mavroudis, P. D., Schwartz, B. D., Mengle-Gaw, L. J., Smith, E. C., Mah, J. K., Guglieri, M., Nevo, Y., Kinnett, K., Tulinius, M., Wrogemann, K., Sweeney, H. L., & others. (2018). Phase IIa trial in Duchenne muscular dystrophy shows vamorolone is a first-in-class dissociative steroidal anti-inflammatory drug. *Pharmacological Research, 136*, 140–150.
8. D'Agostino, R. B., & Stephens, M. A. (Eds.). (2017). *Goodness-of-fit techniques*. CRC Press.
9. Dag, O., Dolgun, A., & Konar, N. M. (2018). Onewaytests: An R package for one-way tests in independent groups designs. *The R Journal, 10*(1), 175–199.
10. Dobrota, R., Maurer, B., Graf, N., Jordan, S., Mihai, C., Kowal-Bielecka, O., Allanore, Y., Distler, O., & others. (2016). Prediction of improvement in skin fibrosis in diffuse cutaneous systemic sclerosis: A EUSTAR analysis. *Annals of the Rheumatic Diseases, 75*(10), 1743–1748.
11. Fang, Y. (2026). *Every coin has two sides: An introduction to causal inference in pharmaceutical statistics*. CRC Press.

12. Garcia, T. P., & Marder, K. (2017). Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current Neurology and Neuroscience Reports, 17*, 14.

13. Hammouri, H. M., Sabo, R. T., Alsaadawi, R., & Kheirallah, K. A. (2020). Handling skewed data: A comparison of two popular methods. *Applied Sciences, 10*(18), 6247.

14. Hao, L., Wang, J., Page, D., Asthana, S., Zetterberg, H., Carlsson, C., Okonkwo, O. C., Li, L., & others. (2018). Comparative evaluation of MS-based metabolomics software and its application to preclinical Alzheimer's disease. *Scientific Reports, 8*, 9291.

15. Hoffman, J. I. E. (2015). *Biostatistics for medical and biomedical practitioners: An interpretative guide for medicine and biology*. Academic Press.

16. Kristoffersson, A. N., David-Pierson, P., Parrott, N. J., Kuhlmann, O., Lave, T., Friberg, L. E., & Nielsen, E. I. (2016). Simulation-based evaluation of PK/PD indices for meropenem across patient groups and experimental designs. *Pharmaceutical Research, 33*, 1115–1125.

17. Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, M., Abebe, Z., Abil, O. Z., Aboyans, V., Abrham, A. R., Abu-Raddad, L. J., & others. (2018). Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1859–1922.

18. Le, P. D. U. M., Tran, P. D. T. T., Vinall, P. D. R., Phan, P. D. V., & others. (2024). Impact of high-stakes exams, incentives, and preparation on student performance: Insights from the pharmacy curriculum outcomes assessment. *Currents in Pharmacy Teaching and Learning*.

19. Leyens, L., Reumann, M., Malats, N., & Brand, A. (2017). Use of big data for drug development and for public and personal health and care. *Genetic Epidemiology, 41*(1), 51–60.

20. Li, Z., Jiang, X., Wang, Y., & Kim, Y. (2021). Applied machine learning in Alzheimer's disease research: Omics, imaging, and clinical data. *Emerging Topics in Life Sciences, 5*(6), 765–777.

21. Lim, J., Walley, R., Yuan, J., Liu, J., Dabral, A., Best, N., Grieve, A., Hampson, L. V., Wolfram, J., Woodward, P., & others. (2018). Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Therapeutic Innovation & Regulatory Science, 52*(5), 546–559.

22. Liu, J., Lahousse, L., Nivard, M. G., Bot, M., Chen, L., van Klinken, J. B., Thesing, C. S., Beekman, M., van der Kallen, C. J. H., Schalkwijk, C. G., Arts, I. C. W., van der Spek, A., Verhoeven, A., Li-Gao, R., Tigchelaar, E. F., Swertz, M. A., Uitterlinden, A. G., Vojinovic, D.,

Amin, N., … van Duijn, C. M. (2020). Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas. *Nature Medicine, 26*, 110–117.

23. Liu, Q. F., Kim, H. M., Lim, S., Chung, M. J., Lim, C. Y., Koo, B. S., Kang, S. S., & others. (2020). Effect of probiotic administration on gut microbiota and depressive behaviors in mice. *DARU Journal of Pharmaceutical Sciences, 28*, 181–189.

24. Mallinckrodt, C., & Lipkovich, I. (2016). *Analyzing longitudinal clinical trial data: A practical guide*. Chapman and Hall/CRC.

25. Matabuena, M., Petersen, A., Vidal, J. C., & Gude, F. (2021). Glucodensities: A new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research, 30*(6), 1445–1464.

26. Mircioiu, C., & Atkinson, J. (2017). A comparison of parametric and non-parametric methods applied to a Likert scale. *Pharmacy, 5*(2), 26.

27. Naghavi, M., Duncan, B. B., Schmidt, M. I., & others. (2025). Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990–2023: A systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*.

28. National Center for Health Statistics. (2017). *Health, United States, 2016: With chartbook on long-term trends in health*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

29. Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice* (2nd ed.). Wiley.

30. Schumacher, A. E., Duncan, B. B., Schmidt, M. I., & others. (2024). Global age-sex-specific all-cause mortality, life expectancy, and population estimates in 204 countries and territories and 811 subnational locations, 1950–2021, and the impact of the COVID-19 pandemic: A comprehensive demographic analysis for the Global Burden of Disease Study 2021. *The Lancet, 403*(10440), 1989–2056.

31. Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., & Ahmad, H. F. (2020). Big data analytics enhanced healthcare systems: A review. *The Journal of Supercomputing, 76*, 1754–1799.

32. Shen, X., Shen, S., Li, J., Hu, Q., Nie, L., Tu, C., Wang, X., Poulsen, D. J., Orsburn, B. C., Wang, J., Qu, J., & others. (2018). IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. *Proceedings of the National Academy of Sciences of the United States of America, 115*(21), E4767–E4776.

33. Turongkaravee, S., Jittikoon, J., Lukkunaprasit, T., Sangroongruangsri, S., Chaikledkaew, U., Thakkinstian, A., & others. (2021). A systematic review and meta-analysis of genotype-based and individualized data analysis of *SLCO1B1* gene and statin-induced myopathy. *The Pharmacogenomics Journal, 21*, 296–307.

34. Uddin, M. J., Groenwold, R. H. H., Ali, M. S., de Boer, A., Roes, K. C. B., Chowdhury, M. A. B., & Klungel, O. H. (2016). Methods to control for unmeasured confounding in pharmacoepidemiology: An overview. *International Journal of Clinical Pharmacy, 38*, 714–723.

35. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery, 18*(6), 463–477.

36. Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., Patro, R., & Robinson, M. D. (2019). RNA sequencing data: Hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science, 2*, 139–173.

37. Wang, J. L., Chiou, J. M., & Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application, 3*, 257–295.

38. Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., Freitag, D. F., Benoit, J., Hughes, M. C., Khan, F., Slater, P., & others. (2021). The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials, 22*, 537.

39. Xu, J., Lin, Y., Yang, M., & Zhang, L. (2020). Statistics and pitfalls of trend analysis in cancer research: A review focused on statistical packages. *Journal of Cancer, 11*(10), 2957–2961.

40. Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., Goyal, H., & others. (2018). Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine, 6*(11), 216.

Unit 5...

STATISTICAL ANALYSIS USING PYTHON THROUGH CASE-BASED LEARNING

This course covers the utilization of statistical tools implemented in Python to conduct pharmaceutical and clinical research by using a problem-based approach. Specifically, the unit covers processes related to the organized generation, examination, description, analysis, and interpretation of pharmaceutical data gathered from domains including clinical trials, pharmacovigilance, pharmacokinetics, pharmacodynamics, quality control, manufacturing, and hospital records. The unit builds expertise in the use of Python packages such as SciPy, Stats models, Pandas, and Scikit-learn to conduct descriptive statistics, correlation, simple linear regression, statistical inference, and report writing in pharmaceutical research. This includes the selection of appropriate statistical tests, interpretation of coefficients and p-values, dealing with uncertainties, and accurate and scientific reporting of pharmaceutical research results.

5.1. PHARMACEUTICAL DATA ANALYSIS WITH PYTHON

Python-based analysis of pharmaceutical data refers to the employment of programming software for the management, cleaning, exploratory analysis, analysis, and interpretation of data collected from pharmaceutical and clinical research. Examples of sources of pharmaceutical data include clinical trials, pharmacovigilance reports, quality control laboratories, pharmacokinetic and pharmacodynamic tests, hospitals, and manufacturing facilities. The use of Python is helpful in automating calculations, working with massive amounts of data, identifying outliers, and making reproducible statistical analysis possible. However, the choice of appropriate analysis techniques is still determined by the characteristics of data.

5.1.1. Introduction to SciPy, Stats models and Scikit-learn

SciPy, Stats models, and Scikit-learn are some crucial Python modules that have different functionalities when it comes to analysing pharmaceutical data. SciPy offers scientific and statistical functions, which can be utilized in testing hypotheses, working with probability distributions, correlating variables, optimizing problems and performing various numerical operations. Stats models library is mostly used for carrying out statistical modelling and inference. This includes regression analysis, test of hypotheses, confidence interval and summary of model among other

aspects. Scikit-learn module is used in machine learning tasks, such as pre-processing of data, classification, regression, clustering and making predictions.

A. SciPy

SciPy is a scientific computing library developed around the Python numerical environment. It offers methods that can be utilized in statistics and science computations. Some applications of SciPy in pharmaceutical research include conducting statistical tests, computations of probability-related values, investigation of dependencies between variables, and numerical analysis.

Such applications include:

- t-tests and other statistical tests
- correlation analysis
- probability distributions
- statistical hypothesis testing
- optimization
- numerical calculations
- analysis of experimental data

A simple example of importing the statistical module is:

```
Python

from scipy import stats
```

For instance, a scientist may conduct a statistical test for the comparison of data collected from two different groups of pharmaceutical treatments.

Use of SciPy becomes handy for scientists who want statistical functions without having to develop calculations from scratch.

B. Stats models

Stats models is a package in the Python programming language that focuses more on statistical modelling and statistical inference. This is especially helpful in instances where scientists require a statistical analysis of the model.

Stats models can be used for:

- Linear regression
- Multiple regression
- Logistic regression
- Hypothesis testing

- Confidence intervals
- Regression coefficients
- Statistical significance
- Model summaries
- Time-series analysis

For example, a researcher may investigate whether drug dose predicts plasma drug concentration.

A basic linear regression structure may be represented as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y= response variable
- X= explanatory variable
- β_0= intercept
- β_1= regression coefficient
- ϵ= error term

Statistical models allow estimation of the coefficient along with the standard error, confidence intervals, test statistics, and p-value, making it especially useful in pharmaceutical studies since inference is very crucial in such studies.

C. Scikit-learn

Scikit-learn is a well-known library in the field of machine learning in the Python programming language. Scikit-learn provides algorithms that can be used to develop models capable of making predictions.

Scikit-learn supports:

- Classification
- Regression
- Clustering
- Data preprocessing
- Feature selection
- Model evaluation
- Cross-validation
- Predictive modelling

For instance, scientists working in pharmacology can employ classification to ascertain if a certain patient will have any negative reaction to the drugs administered to him/her based on certain patient and medication attributes.

Alternatively, a regression model can also be used to predict the continuous outcome like the amount of the drug.

Machine learning process involves:

Data → Preprocessing → Model Training → Prediction → Model Evaluation

Scikit-learn is therefore particularly useful when pharmaceutical research involves prediction and pattern recognition rather than only traditional statistical inference.

Table 1: Comparison of SciPy, Stats models and Scikit-learn

Library	Main Purpose	Pharmaceutical Application
SciPy	Scientific and statistical calculations	Hypothesis tests, correlation, probability analysis
Stats models	Statistical modelling and inference	Regression, coefficients, p-values, confidence intervals
Scikit-learn	Machine learning and prediction	Classification, prediction, clustering, model evaluation

All these libraries can be used simultaneously. In this way, one may use SciPy for statistical tests, Stats models for regressions and inferences, and Scikit-learn for predictions. Such combination gives a lot of flexibility to analyze pharmaceutical data.

5.1.2. Preparation and Exploration of Pharmaceutical Datasets

Data preparation and data exploration are essential early steps in analyzing pharmaceutical data. Before conducting any statistical tests or machine learning analysis, it is necessary to know the characteristics of the data set and whether it is appropriate for analysis. The data sets can come from sources such as clinical trials, pharmacovigilance, quality control labs, pharmacokinetics research, manufacturing procedures, and hospital documentation. Data preparation includes arranging the variables, verifying their types, finding missing or invalid data, detecting anomalies in the data, and checking whether the data set is ready for analysis. Data exploration includes analyzing the distribution and patterns in the data.

A. Importing Pharmaceutical Data

Pharmaceutical datasets may be stored in formats such as CSV, Excel, or other structured files. Pandas can be used to import tabular data into a Data Frame.

For example:

```python
import pandas as pd

data = pd.read_csv("pharmaceutical_data.csv")
```

Data Frame gives structure to the data set in the form of observations being in rows and variables being in columns. Pandas works on structured data.

Some variables that can be part of a pharmaceutical data set include:

Table 2: Example of a Pharmaceutical Dataset

Patient ID	Age	Dose (mg)	Concentration (µg/mL)	Response
P01	35	50	2.4	1
P02	42	100	4.8	1
P03	51	150	6.7	0
P04	29	200	8.9	1

Here, age, dose, concentration, and response represent different types of pharmaceutical or clinical information.

B. Examining the Dataset

After importing the data, researchers should examine its basic structure.

Useful Pandas commands include:

```python
data.head()
```

This displays the first few observations.

```python
data.info()
```

This provides information about variables, data types, and non-missing observations.

```python
data.shape
```

This identifies the number of rows and columns.

```python
data.describe()
```

This provides descriptive statistics for appropriate numerical variables.

These initial checks help researchers understand the size and structure of the dataset before conducting further analysis.

C. Checking Data Types

Correct data types are important for statistical analysis. Pharmaceutical datasets may contain:

- Numerical variables
- Categorical variables
- Binary variables
- Dates and times
- Text-based identifiers

For example:

- **Age** → numerical
- **Drug dose** → numerical
- **Drug concentration** → continuous numerical
- **Treatment group** → categorical
- **Adverse reaction** → binary
- **Patient ID** → identifier

Mishandling of data types can result in issues in statistical modelling. Thus, the researcher must ensure that each variable is recorded properly.

D. Identifying Missing Values

Data sets often contain missing values in the field of pharmaceuticals. The patient's lab value might be missing, the clinical event data might not be captured, and manufacturing measurement might not be available.

Missing data can be analyzed with:

```python
data.isnull().sum()
```

This indicates the number of missing observations in each variable.

The researchers should then come up with a suitable way forward, like:

- Dropping the observations where necessary
- Filling in with the use of a proper imputation technique
- Reasoning behind the missing values
- The magnitude of the missing values

Missing values should not just be filled in without regard to the study design and context.

E. Detecting Duplicate Records

Duplicate observations can occur during data entry or when information from multiple sources is combined.

They can be examined using:

```python
data.duplicated().sum()
```

If duplicate records are identified, researchers should determine whether they represent genuine repeated observations or accidental duplication before removing them.

F. Checking Descriptive Statistics

Descriptive statistics provide an initial understanding of pharmaceutical variables. Python can calculate measures such as:

- Mean
- Median
- Minimum
- Maximum
- Standard deviation
- Quartiles

For example:

```python
data["Concentration"].describe()
```

This allows researchers to learn about the central tendency and variation in drug concentration data.

Descriptive analysis is especially helpful in this regard since pharmaceutical data can have hundreds or even thousands of data points.

G. Exploring Data Distribution

Pharmaceutical measurements need to be studied regarding their distribution. It is vital since biological and pharmaceutical measurements can either be normally distributed, positively skewed, negatively skewed, or any other type of distribution. The measurements can vary greatly.

Useful graphical methods include:

- Histograms
- Box plots
- Scatter plots
- Bar charts
- Normal probability plots

For example, a histogram of plasma drug concentration can show whether most observations are concentrated around a central value or whether a long tail is present.

H. Detecting Outliers

Outliers are observations that vary greatly from the rest of the observations.

For instance, where all plasma drug levels range from 5 to 15 μg/mL but one observation shows 50 μg/mL, the researcher must consider why the observation is like that.

Reasons could be:

- Measurement error
- Data-entry error
- Incorrect sample handling
- Individual biological variation
- Drug interaction
- Unusual pharmacokinetic behaviour

An outlier should not automatically be removed. Its scientific validity should be investigated first.

I. Exploring Relationships Between Variables

Researchers can examine relationships between pharmaceutical variables using correlation analysis and graphical methods.

For example:

Drug dose → Plasma drug concentration

For instance, a scatterplot can demonstrate whether greater doses correspond to increased concentrations. In addition, researchers may examine correlations between:

- Drug concentrations and treatment effectiveness
- Patient age and clearance rate for the drug
- Drug dose and the likelihood of experiencing an adverse event
- Length of therapy and biomarkers

Visual representation is helpful because it helps researchers observe patterns, anomalies, and potential non-linear associations.

5.2. CASE-BASED DESCRIPTIVE AND CORRELATION ANALYSIS

Case-based descriptive and correlation analysis is an important method of analyzing and interpreting the information gathered in pharmaceutical research. Descriptive analysis involves the summary of essential features of pharmaceutical and clinical data sets through the use of tools like the mean, median, standard deviation, min, max, frequencies, and percentages. Correlation analysis is done to find out if there is any relationship between two variables and the strength and direction of that relationship. There are many useful libraries in Python to carry out these calculations, such as Pandas, NumPy, and SciPy. This method can be employed in pharmaceutical research in cases involving patient data, drug dosages, plasma concentration levels of drugs, laboratory tests, and therapeutic outcomes among others.

5.2.1. Descriptive Statistical Analysis Using Python

Descriptive statistics is always the first step involved in quantitative data analysis. It involves summarizing all the information provided in a dataset without making any attempts to predict the future behavior of a population. Descriptive statistics may also be used in the field of pharmacology to provide a summary of various patient characteristics such as age, body weight, blood pressure, dosage, plasma levels, drug efficacy, dissolution, and assaying among others.

For example, consider a dataset containing information about five patients:

Table 3: Patient-Level Pharmaceutical Dataset

Patient	Age	Dose (mg)	Drug Concentration (µg/mL)	Response Score
P1	25	50	2.1	65
P2	32	75	2.8	72

P3	45	100	3.6	81
P4	51	100	3.9	85
P5	39	75	3.0	76

A researcher may want to determine the average patient age, typical drug dose, average drug concentration, variability in concentration, and range of response scores.

The mean is the arithmetic average of observations and is calculated as:

$$\bar{x} = \frac{\sum x_i}{n}$$

Mean is normally calculated for the situation where the distribution is almost normally distributed.

Median refers to the value that lies in the middle of the set of observations that have been arranged in ascending or descending order. It is mostly applied to situations where there is skewness and presence of extreme observations in pharmaceutical and clinical data.

Standard deviation shows how spread the observations are from the mean. When the standard deviation is low, it implies that observations are closely clustered around the mean, and when it is high, it means that observations are widely spread.

Minimum and Maximum show the lowest and highest observation values in a dataset.

Quartiles refer to divisions that occur in ordered data in four parts. First quartile (Q1) refers to 25th percentile, and third quartile (Q3) refers to 75th percentile. The difference between them is referred to as IQR:

$$IQR = Q_3 - Q_1$$

The IQR represents the spread of the middle 50% of observations and can be useful when identifying unusually high or low observations.

Descriptive Analysis Using Python

Pharmaceutical datasets can be imported and analyzed using Pandas.

```python
import pandas as pd

data = pd.read_csv("pharmaceutical_data.csv")

print(data.head())
print(data.describe())
```

The describe () function allows for the retrieval of many valuable statistics related to the numeric variable; they include observations, mean, standard deviation, minimum, quartiles, and maximum.

Single statistical values can also be computed:

```python
print("Mean dose:", data["Dose"].mean())
print("Median dose:", data["Dose"].median())
print("Standard deviation:", data["Dose"].std())
print("Minimum dose:", data["Dose"].min())
print("Maximum dose:", data["Dose"].max())
```

These calculations allow researchers to obtain a quick statistical summary of pharmaceutical measurements.

For categorical variables, frequencies and percentages can be calculated using Pandas:

```python
print(data["Treatment"].value_counts())
print(data["Treatment"].value_counts(normalize=True) * 100)
```

For instance, if 60 out of 100 patients get Treatment A, then the frequency will be 60 and the percentage will be 60%. Data quality must also be considered when describing the data.

Researchers must ensure that there are no missing values, duplicates, incorrect data types, or outliers in the data.

The missing values can be detected by:

```python
print(data.isnull().sum())
```

Duplicate observations can be checked using:

```python
print(data.duplicated().sum())
```

Outliers should not immediately be deleted. In the case of clinical trials, an outlier might very well indicate a real biological effect or a unique feature of that particular patient. It is thus important to first study the scientific importance of outliers.

Case Example

For instance, let us consider a study where the systolic blood pressure of 100 patients is analyzed after taking antihypertensive medication. This produces a systolic blood pressure mean of 128 mmHg and a standard deviation of 10 mmHg.

The result can be reported as:

$$128 \pm 10 \text{ mmHg}$$

This means that the mean systolic blood pressure is 128 mmHg, and standard deviation shows the degree to which an individual observation varies from the mean.

Thus, descriptive statistics help the researchers determine the central tendency, dispersion, range, and distribution of data in pharmaceutics before proceeding with further analysis.

5.2.2. Correlation Analysis and Interpretation of Results

This statistical method is applied to determine if there is an association between two quantitative variables. Correlation analysis plays an important role in the field of pharmacology since scientists use this technique to study associations between different variables, including the drug dose and plasma level, plasma level and response, age and clearance of the drug, etc.

The main indicator of correlation analysis is Pearson's coefficient, which is denoted as r. The values of the coefficient vary between -1 and +1.

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

The general interpretation is:

Table 4: Interpretation of Correlation Coefficient Values

Correlation coefficient	Interpretation
$r = +1$	Perfect positive correlation
$0 < r < 1$	Positive correlation
$r = 0$	No linear correlation
$-1 < r < 0$	Negative correlation
$r = -1$	Perfect negative correlation

Positive correlation is an indication that the greater the values of one variable, the greater are the values of the other variable. For instance, an increase in drug dosage can be related to an increase in plasma concentration of the drug.

Negative correlation is an indication that the higher the values of one variable, the lower are the values of the other variable. For instance, an increase in the period of treatment can be related to a decrease in a certain clinical parameter.

The closer the value of correlation coefficient to zero, the weaker the relationship between the variables. It does not imply, however, that there is no relationship between the variables because it may be nonlinear.

Correlation Analysis Using Python

SciPy can be used to calculate the Pearson correlation coefficient and its associated p-value.

```python
from scipy.stats import pearsonr

r, p_value = pearsonr(
    data["Dose"],
    data["Drug_Concentration"]
)

print("Correlation coefficient:", r)
print("P-value:", p_value)
```

Pandas can also be used to calculate correlations among several numerical variables:

```python
correlation_matrix = data[
    ["Age", "Dose", "Drug_Concentration", "Response_Score"]
].corr()

print(correlation_matrix)
```

Correlation matrices present the correlation coefficient of each combination of selected variables, thus making them useful when analyzing pharmaceutical data with more than one quantitative variable.

Interpretation of Results

Suppose the analysis produces:

```
Correlation coefficient (r) = 0.86
P-value = 0.001
```

The value of 0.86 of r shows a positive linear relationship between the two variables. In case the variables are the drug dose and plasma concentration, the result implies that the patients with high doses have high plasma drug concentration levels.

The p-value gives an idea on how to test the null hypothesis that the population linear correlation coefficient is equal to zero. The small value of the p-value of 0.001 rejects the null hypothesis.

Another example is:

$$r = -0.72$$

This suggests a strong negative linear relationship, implying that high levels of one variable correspond to low levels of another variable.

It is essential to bear in mind that correlation does not imply causation. Thus, even if a strong correlation was found between dose and effectiveness of a drug, this would not necessarily mean that variations in the dose have caused changes in effectiveness. There could be other variables affecting the relationship under study, such as patient's age, disease progression, adherence, presence of other medications, metabolism, and physiology.

This is why correlation must be considered in combination with the research design, sample size, p-value, confidence interval, biological significance of results, and possible confounders.

Case-Based Example

Consider a pharmaceutical study investigating whether drug dose is associated with plasma drug concentration. Data from patients are analyzed using Python, producing:

$$r = 0.86,\ p = 0.001$$

This shows a high level of positive linear correlation between dose and plasma concentration. The small value of the p-value statistically proves that the positive linear correlation is not a product of chance and cannot be attributed to sampling variability alone if the hypothesis of zero correlation in the population is true.

Therefore, this result may stimulate further study through regression analysis or pharmacokinetics to examine how change in dose is related to change in the concentration of the drug. Hence, descriptive analysis describes the main features of pharmaceutical data, whereas correlation analysis measures the degree and direction of correlation between two variables.

5.3. REGRESSION ANALYSIS AND STATISTICAL INFERENCE

Regression analysis is an essential statistical method applied in pharmaceutical research to examine the association between the dependent and independent variables. This statistical technique allows researchers to know how change in an independent variable relates to change in the outcome variable

and even predict future outcomes. Regression analysis is applied in pharmaceutical research to determine the correlation between drug dose and plasma drug level, drug level and drug effect, or formulation factors and drug release properties. The statistical inference that comes with regression includes the estimated coefficients and their levels of uncertainty, statistical significance, and model performance. In Python, some of the modules useful for regression analysis include Stats models, SciPy, Pandas, and NumPy.

5.3.1. Simple Linear Regression Using Pharmaceutical Data

Simple linear regression is a technique that is applied in determining the relationship between an independent variable X and a dependent variable Y. In simple linear regression, the aim is to determine the amount of change in the dependent variable corresponding to a one-unit change in the independent variable.

The general simple linear regression equation is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y= dependent or outcome variable
- X= independent or predictor variable
- β_0= intercept
- β_1= regression coefficient or slope
- ϵ= random error

The fitted regression equation is commonly written as:

$$\hat{Y} = b_0 + b_1 X$$

Here, $\hat{Y}$represents the predicted value of the dependent variable.

In a study in pharmaceutical research, where the researcher is interested in finding out whether drug dose predicts plasma drug concentration, dose becomes the independent variable while plasma drug concentration becomes the dependent variable.

For example:

Table 5: Patient-Wise Drug Dose and Plasma Concentration

Patient	Dose (mg)	Plasma Concentration (µg/mL)
P1	25	1.2
P2	50	2.0

P3	75	2.8
P4	100	3.5
P5	125	4.1

The regression model can be expressed as:

$$\text{Concentration} = \beta_0 + \beta_1(\text{Dose}) + \epsilon$$

If the estimated model is:

$$\hat{Y} = 0.35 + 0.031X$$

the intercept is 0.35 and the slope is 0.031.

The y-intercept is the predicted value of the dependent variable when X = 0. For instance, in some cases of drug research, the y-intercept might lack biological interpretation because an independent variable cannot be equal to zero.

The slope represents the rate of change of the dependent variable for each unit increase in the independent variable. For instance, in the illustration above, the slope of 0.031 implies that each unit increase in the level of dose corresponds to an increase of 0.031 µg/mL of plasma concentration.

Performing Simple Linear Regression in Python

Stats models can be used to perform regression analysis and obtain a detailed statistical summary.

```python
import statsmodels.api as sm

X = data["Dose"]
Y = data["Drug_Concentration"]

X = sm.add_constant(X)

model = sm.OLS(Y, X).fit()

print(model.summary())
```

The sm.OLS() function performs Ordinary Least Squares (OLS) regression. The fit() method estimates the regression coefficients from the data.

The regression coefficients can also be obtained directly:

```python
print(model.params)
```

Predicted values can be calculated using:

```python
predicted = model.predict(X)

print(predicted)
```

Regression analysis is valuable in that it does more than just determine if there is any association between two variables. In regression analysis, one gets an equation showing the association, which can be used to make an estimate of what the predicted value of the response will be at a specified value of the predictor.

For instance, if one has dose as the predictor and concentration of drug as the response variable, then the researcher will be able to get the expected concentration of the drug at a certain dose. These estimates, however, must not exceed what is scientifically sensible.

5.3.2. Interpretation of Output Summaries and p-Values

Some of the statistical measures used when performing regression analysis using stats models include several. It is very crucial to know what these measures mean in order to interpret the results obtained from the study.

The regression summary includes R-squared, Adjusted R-squared, coefficient, standard error, t-statistic, p-value, and confidence interval.

R-squared (R^2) is the measure showing how much variability of the dependent variable is explained by the independent variable.

For example:

$$R^2 = 0.81$$

means that about 81% of the variation in the dependent variable can be attributed to the linear relationship with the predictor variable. However, it does not imply that 81% of the variation in the outcome is caused by the predictor.

The adjusted R^2 is simply the R^2 adjusted for the number of predictors and sample size. The adjusted R^2 is especially important in the comparison of models with different predictors, even though in simple linear regression there is just one predictor.

The coefficient is the estimate of the effect of the predictor on the outcome. For instance:

Table 6: Regression Coefficients and Statistical Significance

Variable	Coefficient	Standard Error	t-value	p-value
Intercept	0.35	0.18	1.94	0.124
Dose	0.031	0.004	7.75	0.001

A dose coefficient of 0.031 means that for every unit increase in dose, there will be an average increase of 0.031 units in the concentration of plasma.

The standard error is used to measure the precision of a regression coefficient estimate. The lower the standard error, the more precise the coefficient estimate.

The t-statistic tests whether an individual regression coefficient is different from a particular value, usually zero.

The p-value helps us test the null hypothesis. In case of a slope coefficient, the null hypothesis is:

$$H_0: \beta_1 = 0$$

while the alternative hypothesis is:

$$H_1: \beta_1 \neq 0$$

If the p-value is less than the selected significance level, commonly 0.05, the null hypothesis is rejected in the conventional hypothesis-testing framework.

For example:

$$p = 0.001$$

Because $0.001 < 0.05$, it can be concluded that there is sufficient statistical evidence that the slope is different from zero. In case of an experiment in pharmacology on dose and concentration of plasma, the above conclusion will indicate statistical evidence of a linear association between the dose and concentration.

It must be emphasized that statistical significance does not mean pharmaceutical or clinical significance. Even though a small effect might be statistically significant, an important effect might not be significant statistically.

The 95% confidence interval offers a range of plausible values for the population regression coefficient based on the assumptions of the model. For instance, the 95% confidence interval of the dose coefficient is:

$$0.022 \text{ to } 0.040$$

The slope estimate obtained is consistent with values lying in this range when the model and the sampling distribution have been taken into consideration. If the confidence interval does not contain zero, then this indicates that the value is statistically significant at approximately 5% two-sided.

The interpretation of a regression analysis must therefore incorporate all these elements such as the coefficient, nature of correlation, size, p-value, confidence interval, and R squared.

Case-Based Example

Let's assume that a drug researcher is studying the effect of drug dose and its concentration in blood and finds the results shown below:

```
Intercept       = 0.35
Dose coefficient = 0.031
R-squared       = 0.81
p-value         = 0.001
95% CI          = 0.022 to 0.040
```

The regression equation is:

$$\hat{Y} = 0.35 + 0.031X$$

The positive sign of the coefficient means that greater doses lead to greater levels of drug concentration in the plasma. The coefficient 0.031 means that when the dose increases by one unit, the concentration is expected to increase by 0.031 units. The R2 value of 0.81 shows that 81% of variability of the concentration can be explained by the linear model. The p-value of 0.001 provides strong evidence that the slope is significantly different from zero. The confidence interval from 0.022 to 0.040 shows the range of uncertainty of the estimation of the slope.

The result cannot be considered as evidence of the causality between the dose and the observed concentration since other variables such as absorption, distribution, metabolism, and elimination could influence the plasma concentration.

5.4. STATISTICAL REPORTING AND RESEARCH INTERPRETATION

Statistics reporting is a vital component of pharmaceutical research due to the fact that statistics have to be reported in a clear, precise and scientific way. Statistics alone cannot give us a final conclusion in the research; therefore, the statistics need to be well arranged, interpreted and presented so that the reader will get an understanding of what was discovered in the experiment. The report on statistics will include dataset details, data analysis methods, descriptive statistics, statistical test used, effect estimates, p-values, confidence interval, and results. In a pharmaceutical study, proper statistical reporting will assist the researcher to convey scientific results without making any misleading

conclusions. Python generated statistics can be used in the reports of research after proper interpretation.

5.4.1. Preparation of Statistical Reports

Statistical Report is an organized description of the statistical methods used and results found through research. It is supposed to provide enough information to enable the reader to understand what data has been analyzed, what statistical methods have been used, what the results are, and their interpretations.

First, it is necessary to describe the purpose of the research and the dataset involved. In particular, the report needs to show the population and experimental materials being studied, the size of the sample, the key variables included, treatment groups if any, and the kind of data obtained. In a pharmaceutical clinical trial, for instance, this will include patient's age, treatment group, dosage, plasma concentrations, effectiveness, and side effects.

Next, there should be a clear description of the statistical methods. The researcher needs to show which statistical method was used such as descriptive statistics, correlation analysis, regression analysis, hypothesis testing, and others. The choice of the method used needs to be consistent with the data involved.

For quantitative variables, results can be given in terms of mean and standard deviation. If data is skewed, the median and the interquartile range may be more appropriate. For categorical variables, results can be given in terms of frequencies and percentages.

For example:

The average age of the subjects in the study was 42.6 ± 8.4 years. The average plasma drug concentration was 3.25 ± 0.72 μg/mL.

A statistical report should also include a clear description of inferential statistics. For example, in presenting the results of regression analysis, the analyst can give the coefficient estimate, confidence intervals, p-value, and R^2 value.

An example is:

$$\beta = 0.031,\ 95\%\ CI = 0.022 - 0.040,\ p = 0.001$$

This provides much greater detail as compared to merely stating that the result was “significant.”

The p-value should be stated properly without simply stating any small p-values as “$p < 0.05$.” For instance, “when feasible it is preferable to report $p=0.003$ instead of $p<0.05$.”

Confidence intervals also should be provided wherever possible because it provides extra information about the precision of the estimated effect.

Python can assist in the production of statistical results:

```python
import pandas as pd
import statsmodels.api as sm

data = pd.read_csv("pharmaceutical_data.csv")

X = sm.add_constant(data["Dose"])
Y = data["Drug_Concentration"]

model = sm.OLS(Y, X).fit()

print(model.summary())
```

The output of the results produced through Python programming must not be directly included in the research paper without interpretation. The researcher needs to check whether the variables used, sample size, statistical test applied, the assumptions made and the output obtained are correct.

An ideal statistical paper should have the following features: accuracy, conciseness, transparency, reproducibility, and scientific relevance. It should clearly differentiate between the results of statistics and scientific interpretations.

5.4.2. Interpretation and Presentation of Pharmaceutical Research Findings

Explanation refers to how the data collected has been interpreted in relation to the problem under investigation. This is not merely recitation of numerical figures. An effective interpretation relates the statistical findings to the pharma/clinical world while taking into account the limitations of the analysis.

For instance, if a study is designed to look at the relationship between drug dose and plasma concentration and gets:

$$r = 0.86,\ p = 0.001$$

This may be regarded as suggesting a strong positive linear relationship between the dose of the drug and the plasma concentration level. The low p-value is statistically significant in rejecting the null hypothesis of no population linear correlation at all.

If a regression analysis gives:

$$\hat{Y} = 0.35 + 0.031X$$

the positive slope suggests a direct relationship between dose and predicted plasma concentration. The coefficient needs to be interpreted by considering its units of measurement. For instance, a

coefficient of 0.031 implies that a one unit increase in dose results in a corresponding increase in predicted plasma concentration of 0.031 units.

Statistical significance versus practical significance should be considered when reporting the research findings in the pharmaceutical field. Statistical significance does not imply that the observed difference or relationship is large enough to be practically significant. On the other hand, the observed relationship may be practically significant but lack statistical significance because of limited sample size or variability in the data set.

In addition to measures of central tendency and spread, measures of uncertainty should be reported. For example,

Treatment A resulted in a mean reduction of systolic blood pressure of 12.4 mm Hg, compared to a reduction of 7.8 mm Hg for Treatment B, resulting in a mean difference of 4.6 mm Hg (95% CI: 1.8–7.4; $p = 0.002$)

such presentation gives more meaningful information than reporting the p-value alone.

Findings in the pharmaceutical field should be interpreted with respect to their biological and clinical significance. For example, the statistically significant relationship between drug dose and predicted plasma concentration can be expected based on pharmacokinetic principles but it is up to the researcher to consider such factors as metabolism, absorption, patient compliance, among others.

Results should be presented in properly constructed tables and well-structured text as appropriate. The table should have a descriptive title, appropriately labelled variables, units of measurements and adequate content for interpretation.

For example:

Table 7: Descriptive Statistics of Pharmaceutical Variables

Variable	Mean ± SD	Minimum	Maximum
Age (years)	42.6 ± 8.4	25	61
Dose (mg)	75.0 ± 18.5	50	100
Drug concentration (μg/mL)	3.25 ± 0.72	2.1	4.4

The results section needs to highlight the findings, whereas the discussion needs to interpret the findings within the framework of pharmacological research. It is advised that researchers not make any unsubstantiated interpretations or try to establish any causal relationship from association.

Limitations of the study must also be mentioned. Some of them can be a small sample size, incomplete data, variations in measurements, possibility of confounders, lack of generalizability, and assumptions related to the test used statistically.

A good way of reporting a statistical finding in pharmaceutical research is:

Finding → Statistical evidence → Magnitude → Scientific interpretation → Limitation

For example:

There is a positive relationship between the dosage of drugs and plasma concentration ($r=0.86$, $p=0.001$). The high value of the correlation coefficient implies that there is a relationship between higher dosages and higher plasma concentration in the group of participants. The information can be useful for the assessment of dose-concentration; nevertheless, there are other factors which may affect plasma concentration as well.

Hence, effective reporting of statistics is not only about presenting numbers calculated with the help of Python software. It consists in organizing the data correctly, employing correct statistical tests, giving precise estimates and uncertainty measurements, interpretation, and presenting pharmaceutical relevance. Thus, statistical results will be comprehensible and applicable in pharmaceuticals, clinical trials, drug formulations, pharmacokinetics, and other disciplines.

REFERENCES

1. Allen, R., & Moore, H. (2019). Perspectives on the Role of Mathematics in Drug Discovery and Development. *Bulletin of Mathematical Biology*, *81*(9), 3425.
2. Al-Worafi, Y. M. (2024). Medical and health sciences in developing countries: Research fields. In *Handbook of medical and health sciences in developing countries*. Springer.
3. Ameri, A., Khajouei, R., Ameri, A., & Jahani, Y. (2020). Acceptance of a mobile-based educational application (LabSafety) by pharmacy students: An application of the UTAUT2 model. *Education and Information Technologies, 25*, 419–435.
4. Arbet, J., Brokamp, C., Meinzen-Derr, J., Trinkley, K. E., Spratt, H. M., & others. (2021). Lessons and tips for designing a machine learning study using EHR data. *Journal of Clinical and Translational Science, 5*(1), e21.
5. Austin, C. P. (2021). Opportunities and challenges in translational science. *Clinical and Translational Science, 14*(5), 1629–1647.
6. Basu, C., & Carlin, B. P. (2020). Bayesian hierarchical models for data extrapolation and analysis in pediatric disease clinical trials. In *Bayesian methods in pharmaceutical research*. Chapman and Hall/CRC.
7. Bhangdia, K., May, M. L., Kocarnik, J. M., Pritchett, N., & others. (2026). Global, regional, and national burden of breast cancer among females, 1990–2023, with forecasts to 2050: A systematic analysis for the Global Burden of Disease Study 2023. *The Lancet Oncology*.
8. Cook, A. J., Delong, E., Murray, D. M., Vollmer, W. M., Heagerty, P. J., & others. (2016). Statistical lessons learned for designing cluster randomized pragmatic clinical trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core. *Clinical Trials, 13*(5), 504–512.
9. Hus, A. K., Xu, Y. Y., Steinmetz, J. D., Aalipour, M. A., & others. (2025). Global, regional, and national burden of headache disorders, 1990–2023: A systematic analysis for the Global Burden of Disease Study 2023. *The Lancet Neurology*.
10. James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … Murray, C. J. L. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A

systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1789–1858.

11. Kapadia, A. S., Chan, W., & Moyé, L. A. (2017). *Mathematical statistics with applications*. CRC Press.

12. Karp, N. A., & Reavey, N. (2019). Sex bias in preclinical research and an exploration of how to change the status quo. *British Journal of Pharmacology, 176*(21), 4107–4118.

13. Kocherlakota, S., & Kocherlakota, K. (2017). *Bivariate discrete distributions*. CRC Press.

14. Koenig, F., Slattery, J., Groves, T., Lang, T., Benjamini, Y., Day, S., ... & Posch, M. (2015). Sharing clinical trial data on patient level: opportunities and challenges. *Biometrical Journal*, *57*(1), 8-26.

15. Konrad, R., Zhang, W., Bjarndóttir, M., & Proaño, R. (2020). Key considerations when using health insurance claims data in advanced data analyses: An experience report. *Health Systems, 9*(4), 317–325.

16. Kumar, A., Terakosolphan, W., Hassoun, M., Vandera, K. K., Novicky, A., Harvey, R., Royall, P. G., Bicer, E. M., Eriksson, J., Edwards, K., Valkenborg, D., Nelissen, I., Hassall, D., Mudway, I. S., Forbes, B., & others. (2017). A biocompatible synthetic lung fluid based on human respiratory tract lining fluid composition. *Pharmaceutical Research, 34*, 2454–2465.

17. Liu, J., Barrett, J. S., Leonardi, E. T., Lee, L., Roychoudhury, S., Chen, Y., Trifillis, P., & others. (2022). Natural history and real-world data in rare diseases: Applications, limitations, and future perspectives. *Journal of Clinical Pharmacology, 62*(S2), S38–S55.

18. Macip, G., Garcia-Segura, P., Mestre-Ferrandiz, J., & others. (2022). Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Medicinal Research Reviews, 42*(2), 744–769.

19. Mageed, I. A. (2025). *The hidden mathematics to treat cancer*. Preprints.

20. Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2018). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics, 19*(2), 286–302.

21. Mendonça, S. A. M., Melo, A. C., Pereira, G. C. C., Santos, D. M. S. S., Grossi, E. B., Sousa, M. C. V. B., Ramalho-de-Oliveira, D., & others. (2016). Clinical outcomes of medication therapy management services in primary health care. *Brazilian Journal of Pharmaceutical Sciences, 52*(3), 365–373.

22. Meyers, L. S., Gamst, G., & Guarino, A. J. (2017). *Applied multivariate research: Design and interpretation*. Sage Publications, Inc.

23. Mulder, N., Adebamowo, C. A., Adebamowo, S. N., Adebayo, O., Adeleye, O., Alibi, M., ... & Souiai, O. (2017). Genomic research data generation, analysis and sharing–challenges in the African setting. *Data Science Journal*, *16*, 49-49.

24. Murray, C. J. L., Callender, C. S. K. H., Kulikoff, X. R., Srinivasan, V., Abate, D., Abate, K. H., Abay, S. M., Abbasi, N., Abbastabar, H., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abebe, D. G., Abay, K. A., Abebe, H. T., Abebe, M., Abebe, N., … GBD 2017 Population and Fertility Collaborators. (2018). Population and fertility by age and sex for 195 countries and territories, 1950–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet, 392*(10159), 1995–2051.

25. Nahum-Shani, I., Ertefaie, A., Lu, X., Lynch, K. G., McKay, J. R., Oslin, D. W., & Almirall, D. (2017). A SMART data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction, 112*(5), 901–909.

26. Neuenschwander, B., Roychoudhury, S., & Schmidli, H. (2016). On the use of co-data in clinical trials. *Statistics in Biopharmaceutical Research*, *8*(3), 345-354.

27. Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics, 17*(1), 29–39.

28. Ozgur, C., Kleckner, M., & Li, Y. (2015). Selection of statistical software for solving big data problems: A guide for businesses, students, and universities. *SAGE Open, 5*(2), 1–12.

29. Pan, Y., Kadash-Edmondson, K. E., Wang, R., Phillips, J., Liu, S., & Ribas, A. (2021). RNA dysregulation: An expanding source of cancer immunotherapy targets. *Trends in Pharmacological Sciences, 42*(4), 268–282.

30. Pang, T. Y., Lee, T. K., & Murshed, M. (2023). Towards a new paradigm for digital health training and education in Australia: Exploring the implication of the fifth industrial revolution. *Applied Sciences, 13*(11), 6854.

31. Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment. *Medicina, 56*(9), 455.

32. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. *Metabolomics, 11*, 1492–1513.

33. Roychoudhury, S., Anderson, K. M., Ye, J., Mukhopadhyay, P., Kundu, S., & others. (2023). Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. *Statistics in Biopharmaceutical Research, 15*(2), 280–294.

34. Ruberg, S. J., Beckers, F., Hemmings, R., Honig, P., Irony, I., LaVange, L., LeBlond, D., Loewy, J., Rantell, K., Shun, Z., & others. (2023). Application of Bayesian approaches in drug development: Starting a virtuous cycle. *Nature Reviews Drug Discovery, 22*, 235–250.

35. Sepanlou, S. G., Safiri, S., Bisignano, C., Ikuta, K. S., Merat, S., Saberifiroozi, M., Poustchi, H., Tsoi, D., Colombara, D. V., Abdoli, A., Adedoyin, R. A., Afarideh, M., Agrawal, S., Ahmad, S., Ahmadian, E., Ahmadpour, E., Akinyemiju, T., Alipour, V., Almasi-Hashiani, A., … Malekzadeh, R. (2020). The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Gastroenterology & Hepatology, 5*(3), 245–266.

36. Tolley, E. E., Ulin, P. R., Mack, N., Robinson, E. T., & Succop, S. M. (2016). *Qualitative methods in public health: A field guide for applied research* (2nd ed.). Jossey-Bass.

37. Tong, C. (2019). Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician, 73*(S1), 246–261.

38. Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine, 34*(6), 984–998.

39. Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA, 323*(9), 844–853.

40. Xu, J., Zhang, P., Huang, Y., Zhou, Y., Hou, Y., Bekris, L. M., Lathia, J. D., & others. (2021). Multimodal single-cell/nucleus RNA sequencing data analysis uncovers molecular networks between disease-associated microglia and astrocytes with implications for drug repurposing in Alzheimer's disease. *Genome Research, 31*(10), 1900–1912.

ABOUT THE AUTHORS



Prof. (Dr.) Amit Kumar currently working as a professor inFaculty of Pharmaceutical Sciences, Motherhood University, Roorkee, Uttarakhand, India. Prof. (Dr.) Amit Kumar is a pharmaceutical academician and researcher with over 17 years of experience in teaching, research, and academic administration. He holds a Ph.D. in Pharmaceutical Sciences, M.Pharm in Clinical Pharmacy, and B.Pharm. He has served as Principal and subsequently as Deputy Registrar and Registrar at Smt. Manjira Devi University, Uttarakhand. His areas of research include Drug Design and Development, Molecular Docking, Ethnopharmacology, and Phytochemical Screening. He has published 60 National and International research papers, contributed to 06 national and international patents, and authored 06 pharmacy books. He has guided 08 Ph.D. research scholars., 10 Postgraduate, and 15 Undergraduate research projects and has actively contributed to seminars, workshops, conferences, curriculum development, and scientific peer review publication. He is associated with IIP Press as a reviewer/editorial contributor and is a member of the Society of Higher Education (SHER).



Mrs. Bhagyashri V. Gawade is an Assistant Professor in Faculty of Health Sciences at JSPM's University, Wagholi, Pune, India and is pursuing her Ph.D. in Pharmaceutical Sciences at JSPM University. She holds an M. Pharm. in Pharmaceutics from SVPM's College of Pharmacy, Malegaon, Baramati and has more than 15 years of academic and teaching experience in pharmaceutical sciences. Her research interests include novel drug delivery systems, formulation development, Transdermal Drug Delivery System, Self Emulsifying Drug Delivery System and pharmaceutical research. She has authored and co-authored several research publications, published 1 Design Patent and a book chapter in the field of pharmaceutics. She has also received research funding under the ASPIRE scheme of Savitribai Phule Pune University and has guided number of M. Pharmacy and B. Pharmacy students at research level. She has contributed to academic administration, examination, accreditation, and student-development activities.



Dr. Charu Sinha is a healthcare management, quality, auditing, training and academic professional with a BDS and an MBA in Hospital & Healthcare Management, with more than 10 years of experience across Research, Academics, Training and Quality. Complemented by advanced credentials in ISO Quality Management Systems, Lean Six Sigma, Clinical Research and AI-enabled quality improvement. She has professional experience across hospital administration, clinical research, academic coordination, training, healthcare quality and process improvement, with associations including Centre for Sight, Apollo Hospitals, Indian Spinal Injuries Centre, Ikure, and many more As an ISO Lead Auditor, she has conducted and participated in management system audits across healthcare and diverse service/industrial sectors, covering ISO standards. She is also a Lean Six Sigma Master Black Belt with expertise in QMS, NABH, healthcare KPIs, risk-based thinking, process improvement and regulatory compliance. As Founder of Dental Plus Direct, she leads a multidisciplinary team and develops professional education, quality, accreditation and training initiatives in dentistry, collaborating with professional organizations for webinars and capacity-building programs. She has served as a guest speaker and trainer on healthcare quality, benchmarking and NABH, contributed to multidisciplinary research, publications, and patents. She is also skilled in AI-assisted research and productivity tools, SPSS, Advanced Excel, statistical analysis, data visualization and digital learning platforms, integrating quality management, healthcare expertise and emerging technologies in her professional and academic work.



Dr. Richa Srivastava is a Ph.D.-qualified Pharmaceutical Scientist and Professor at Ambekeshwar Institute of Pharmaceutical Sciences , with expertise in Pharmaceutics, pharmaceutical research, and novel drug delivery systems. Her academic and research experience includes polyherbal formulations, targeted nasal drug delivery, thermoreversible gels, liposomal drug delivery, and data analytics in polyherbal formulations. She has contributed to peer-reviewed research publications and presented her work at national and international scientific forums.Through Applied Biostatistics and Data Analytics for Pharmaceutical Sciences (Theory),she aims to provide pharmacy students with a clear and practical foundation in biostatistics and data analytics, helping them understand the principles needed to interpret pharmaceutical research and make informed scientific decisions.



Mr, Selvamuthukumar R, M.Pharm., is an academician and researcher in the field of Pharmaceutical Sciences with expertise in Pharmaceutics, novel drug delivery systems, and pharmaceutical formulation development. He is currently serving as an Associate Professor in the Department of Pharmaceutics and has over five years of teaching B. Pharmacy, D. Pharmacy, and Pharm. D and research experience. His research interests include nanosponges, microspheres, transdermal drug delivery systems, emulgels, and the application of artificial intelligence in pharmaceutical formulation development. He has qualified M. Pharmacy in Industrial Pharmacy branch from, Annamalai University, and B. Pharmacy from Annamalai University, Chidambaram, Tamilnadu, India. He has guided many students of B. Pharmacy at research level. He has more than 9 publications. He has actively contributed to scientific research through publications in peer-reviewed journals and participation in national and international conferences. His commitment to academic excellence and innovative research continues to support advancements in pharmaceutical education and drug delivery technology.



Mrs. Manisha Chandrakar is an Assistant Professor at Chhattisgarh Swami Vivekanand Technical University (CSVTU). She has academic and teaching experience in the field of Pharmaceutical Sciences. Her area of expertise is Pharmacology and related pharmaceutical sciences. She has a keen interest in research, innovation, and the development of scientific knowledge. She has published research and review papers in SCI, Scopus, and UGC-indexed journals. She has contributed to the field of academics through her teaching and research activities. She is the author/co-author of two books in the field of Pharmaceutical Sciences. She is also a patent holder reflecting her interest in research and innovation. Her research work includes pharmacological evaluation and exploration of medicinal plants. She is dedicated to quality education, scientific research, and the overall development of Pharmaceutical Sciences.

NEXUS
KNOWLEDGE PUBLICATION
Imprint of AKT Multitask Consultancy

editor@nknpub.com
https://nknpub.com
Sarkanda Bilaspur, Chhattisgarh, India

ISBN:- 978-81-180323-0-4
9 788118 032304
MRP ₹499/-